Information Structure in an Annotated Parallel English-Czech Corpus: A Comparative Probe

Eva Hajičová, Jiří Mírovský, Kateřina Rysová and Magdaléna Rysová
(Charles University, Institute of Formal and Applied Linguistics
Prague)

Keywords: topic–focus articulation, information structure, word order, contrastive analysis, Prague Czech-English Dependency Treebank

Schedule: Thu 11.30 Room 11

Information structure (IS) is one of the aspects of language structure that is reflected in some way or another in most linguistic theories since the pioneering studies of Mathesius (1929) and Halliday (1967); it is more or less explicitly assumed that from the semantic and/or pragmatic points of view the function of IS is well comparable across languages (Prince 1981, Partee 1991, Steedman 2000, Krifka 2006). It would then naturally follow that in translations the IS of the source and of the target language sentences should be preserved.

Based on the Praguian theory of Topic/Focus Articulation (TFA; Sgall 1967, Sgall et al. 1973, Hajičová et al. 1998), repeatedly tested on corpus material (Mírovský et al. 2013, Rysová et al. 2015, Hajičová and Mírovský 2018), and using a parallel English(source)–Czech(target) corpus annotated both for deep syntactic structure and for TFA (PCEDT, Hajič et al. 2012), we have followed two research questions:

(i) How far does the assignment of Focus proper (= the last element of global Focus) agree in English and in Czech?

(ii) If the assignment of Focus proper differs, is the Focus-proper element in English at least a member of the (global) Focus of the Czech sentence?

Results:

The total number of automatically aligned sentences without coordination of the main predicates was 3857; there were 2514 cases (65,3%) with the same syntactic value at the last position of both source and target text and 1287 cases where there was a difference in this value.

(i) After a manual inspection of the randomly selected 120 sentences with a different syntactic label in Focus proper position, we have filtered out cases where the syntactic label differed but the target lexical item corresponded to the source one. We had then at our disposal 24 examples of an actual difference in Focus proper. Most frequently, the difference concerned the mutual position of the main predicate and its modification of time or place (in English, this modification frequently was in the Focus proper) or the position of the predicate itself (in Czech, the predicate was in the Focus proper).

(ii) As for the second question, a manual filtering of the whole set of 171 sentences ended up with a set of 30 sentences in which the element assigned Focus proper in English does not appear even in the global Focus part of the target Czech sentences. Again, most differences concerned temporal or local modifications which in the Czech sentence appeared in the Topic rather than in the global Focus. A second group of examples concerned the mutual position of the main predicate and its argument Patient: Patient was “topicalized” in Czech.

Conclusion:

Our analysis confirms that the differences in IS between the examined languages are rather rare, though certain tendencies can be observed, namely in case of the temporal and local modifications of the main predicate. We have also carefully analyzed the context in which the sentences occur and it came out that the context actually offers both interpretations of the IS structure.

Acknowledgment:
This work has been supported and using language resources and tools distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2015071 and OP VVV VI CZ.02.1.01/0.0/0.0/16 013/0001781).

References

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., Žabokrtský, Z. (2012), Announcing Prague Czech-English Dependency Treebank 2.0. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, European Language Resources Association, İstanbul, Turkey, ISBN 978-2-9517408-7-7, pp. 3153–3160.

Hajičová, E., Mírovský, J. (2018), Discourse Coherence Through the Lens of an Annotated Text Corpus: A Case Study. In: *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association, Paris, France, ISBN 979-10-95546-00-9, pp. 1637–1642.

Hajičová, E., Partee, B. H. and P. Sgall (1998), *Topic-Focus Articulation, Tripartite Structures, and Semantic Content*, Dordrecht, Kluwer Academic Publishers.

Halliday, M. A. K. (1967), Notes on transitivity and theme in English. Part 2, *Journal of Linguistics 3*, 199–244.

Krifka, M. (2008), Basic Notions of Information Structure. *Acta linguistica Academica*, Vol. 55, Issue 3-4, 243–276.

Mathesius, V. (1929), Zur Satzperspektive im modernen Englisch. *Archiv für das Studium der neueren Sprachen und Literaturen*, 155(29), 202–210.

Mírovský, J., Rysová, K., Rysová, M., Hajičová, E. (2013), (Pre-)Annotation of Topic-Focus Articulation in Prague Czech-English Dependency Treebank. In: *Proceedings of the 6th International Joint Conference on Natural Language Processing, Asian Federation of Natural Language Processing*, Nagoya, Japan, ISBN 978-4-9907348-0-0, pp. 55–63.

Partee, B. H. (1991), Topic, Focus and Quantification. In: Moore, S., Wyner, A. (ed.): *Proceedings from SALT I*, Ithaka, N.Y.: Cornell University, pp. 257–280.

Prince, E. (1981), Toward a taxonomy of given/new information. In Cole, ed. *Radical Pragmatics (1981)*, 223–254.

Rysová K., Mírovský J., Hajičová E. (2015), On an apparent freedom of Czech word order. A case study. In: *14th International Workshop on Treebanks and Linguistic Theories (TLT 2015)*, Warszawa, Poland, ISBN 978-83-63159-18-4, pp. 93–105.

Sgall P. (1967), Functional Sentence Perspective in a Generative Description. In *Prague Studies in Mathematical Linguistics 2*, Prague, Academia, 203–225.

Sgall, P., Hajičová, E., Benešová, E. (1973), *Topic, Focus and Generative Semantics.* Kronberg/Taunus: Scriptor.

Steedman, M. (2000), Information structure and the syntax-phonology interface. *Linguistic Inquiry, 31*, pp. 649–689.