Selected bibliography

Bethin, Christina Y. (1998) *Slavic prosody*. Cambridge: Cambridge University Press.

Garde, Paul (1985) Le mythe de l' allongement comoensatoire en ukrainien. In Hursky, Jacob P. (ed.) *Studies in Ukrainian Linguistics in Honor of George Y. Shevelov*, 69-81. (*The Annals of the Ukrainian Academy of Arts and Scoiences in the U.S.*, 39/40.)

Gvozdanović, Jadranka (2009) *Celtic and Slavic and the Great Migrations*. Heidelberg: Universitätsverlag Winter (AATSEEL prize 2010).

Shevelov, George Y. (1979) *A historical phonology of the Ukrainian language*. Heidelberg: Universitätsverlag Winter.

Timberlake, Alan (1983a) Compensatory lengthening in Slavic, 1: Conditions and dialect geography. In: Markov, Vladimir and Dean S. Worth (eds.) *From Los Angeles to Kiev: Papers on the Occasion of the Ninth International Congress of Slavists*, Kiev, September 1983, 207-235. Columbus, OH: Slavica.

--- (1983b) Compensatory lengthening in Slavic, 2: Phonetic Reconstruction. In. Flier, Michael S. (ed.) *American Contributions to the Ninth International Congress of Slavists*, Kiev, September 1983, vol. 1: Linguistics, 293-319. Columbus OH: Slavica.

A contrastive look at information structure: A corpus probe

Eva Hajičová, Jiří Mírovský, Katya Brankatschk
Institute for Formal and Applied Linguistics
Charles University in Prague
Praha, Czech Republic
hajicova@ufal.mff.cuni.cz

Key words: information structure, dependency grammar, Czech, English, corpus annotation

1. Aim of the Contribution

Studies on the linguistic phenomenon subsumed at present under a common label of information structure (but investigated for many decades under most different terms such as topic-focus articulation, functional sentence perspective, theme-rheme structure, presupposition and focus etc.) have a long history and take as their starting point different theoretical considerations. The aim of our contribution is to present a corpus-based comparison of Czech and English and to consider which aspects of this phenomenon are of an universal nature and which are language specific.

2. Theoretical Background

Our study is based on the approach to information structure of the sentence (topic-focus articulation, TFA in the sequel) as developed within the functional generative description of language (FGD in the sequel; for a general account of this formal theory, see Sgall et al 1986). The basic idea is that the dichotomy of topic and focus corresponds to the relation of 'aboutness': the focus of the sentence says something ABOUT its topic. The formal

representation of a sentence on its syntactico-semantic (tectogrammatical) layer corresponds to a dependency tree structure in which each node of the tree is assigned – side by side with a label identifying its underlying syntactic function (such as Actor, Patient, Addressee, Effect, Origin, different kinds of place, time, direction and the like) – a specification of contextual boundness or non-boundness. Based on these primary features, the sentence representation can be divided into the topic part and the focus part of the whole sentence. The TFA structure of the sentence is assumed to be semantically relevant, since e.g. *On Saturdays I work on my dissertation* differs from *On my dissertation I work on Saturdays*. For a description of the TFA theory, see Sgall et al 1986 and Hajičová, Partee and Sgall 1998; for its application to Czech, see esp. Sgall, Hajičová and Buráňová, 1980).

3. TFA annotation of Czech corpus

The FGD approach has served as a theoretical background for the design of the multilayered annotation scheme of the Prague Dependency Treebank (PDT). PDT is an annotated collection of Czech texts, randomly chosen from the Czech National Corpus, with a mark-up on three layers: (a) morphemic, (b) surface shape "analytical", and (c) underlying syntactic (tectogrammatical). The current version (the description of which is publicly available on http://ufal.mff.cuni.cz/pdt2.0, with the data themselves available at LDC under the catalog No. LDC2006T01), contains 3,165 documents (text segments mainly of a journalistic genre) comprising of 49,431 sentences and 833,195 occurrences of tokens (word forms and punctuation marks) annotated on all the three layers.

On the tectogrammatical layer, which is our main concern from the theoretical point of view, every node of the tectogrammatical representation (TGTS, a dependency tree) is assigned a label consisting of: the lexical value of the word, its '(morphological) grammatemes' (i.e. the values of morphological categories such as Feminine, Plural, Preterite etc.), its 'functors' (with a more subtle differentiation of syntactic relations by means of subfunctors, e.g. 'in', 'at', 'on', 'under'), and the topic-focus articulation (TFA) attribute containing the values for contextual boundness. In addition, some basic coreferential links (including intersentential ones) are also added. It should be noted that TGTSs may contain nodes not present in the morphemic form of the sentence in case of surface deletions.

A similarly based annotation, though not covering all the features captured by the PDT, exists for English in the so-called Prague English Dependency Treebank (PEDT) comprising tectogrammatical (syntactico-semantic) annotation of texts from the Wall Street Journal (12,440 annotated and checked trees), see Cinková et al. (2009), and first of all the Prague Czech English Dependency Treebank (PCEDT; Čmejrek et al. 2004) comprising an annotation of Czech and English parallel texts (21,600 sentences) in the lines of PDT. PCEDT 1.0 also comprises a parallel Czech-English corpus of plain texts from Reader's Digest 1993-1996 consisting of 53,000 parallel sentences.

4. Methodology

The material described in Sect. 3 above has allowed for a more detailed contrastive analysis of tectogrammatical (underlying syntactic) sentence structures as for the topic focus structure of Czech and English sentences. As there exists a detailed manual for annotators instructing

them how to assign syntactic and TFA annotations for Czech sentences in PDT (Mikulová et al. 2005), we could formulate the following research inquiry:

The basic hypothesis behind our approach to TFA (supported, as a matter of fact, by an extensive linguistic literature on most different languages) claims that TFA is a universal phenomenon of language that is semantically relevant but can be expressed by different means in different languages. To what extent and in which points is then possible to apply the instructions of the manual for Czech to the analysis of English?

5. Czech and English: Commonalities and Differences

5.1 Basically, in both languages a common strategy in communication is to proceed from a retrievable, identifiable information to an unretrievable one. This strategy can be documented for Czech by the fact that in the PDT, there is only a small portion of cases in which a contextually bound item in the topic of the sentence does not provide a coreferential link (i.e. it does not serve as an anaphor); for a detailed quantitative as well as qualitative analysis, see Hajičová and Mírovský, in prep.). As for English, a good indicator is the appearance of a indefinite article in the subject position of sentences, if one assumes an unmarked position of the intonation center at the end of the sentence. Such cases are rather rare and can be explained by an interaction of other factors (as the marked meaning of the indefinite article "one of the"); these cases are analyzed and documented on the material from the Czech-English corpus in Mírovský and Hajičová (in prep.).

5.2 One of the basic instructions in the annotation manual for Czech, also supported by an empirical analysis of Czech (see Sgall, Hajičová and Buráňová 1980), is to consider the part of the sentence preceding the verb as the topic of the sentence, and the part of the sentence following the verb the focus of the sentence, with the verb belonging either to the topic or to the focus, according to the preceding context. This basically holds for Czech, though even here, there is a trend to place the verb into the second position (see Zikánová and Týnovský 2009). In English, with its basically grammatically fixed word order, the position after the verb is to be examined more carefully; here again the use of the definite article is a good indicator (see Mírovský and Hajičová in prep. for some quantitative results). A marked position of the intonation center is more frequent here than in Czech.

5.3 Another difference concerns the position of the so-called focusing particles (focalizers) in Czech (*jenom, také, dokonce*, …) and their English counterparts (*only, also, even*, …) and their semantic scope. While in Czech a typical position of a focalizer in the surface shape of the sentence is immediately before the sentence element the focalizer is "associated with", in English this need not be the case, as illustrated by the example *John only introduced Bill to SUE.* and its interpretations illustrated by the continuations (a) … and not to MARY., (b) … and not Nick to MARY., (c) … and did not say hello to the HOSTESS/and he LEFT.
In Czech, we have to distinguish these readings by placing the focalizer immediately before the focused element (or group of elements, i.e. before the focus of the sentence) even in the surface shape). It is interesting to notice that contrary to the general characteristics of Czech as a language with a relatively ‚free' word order (i.e. without grammatical word-order restrictions), in the placement of the focalizer *only* English is more flexible than Czech is:

this particle can be placed either immediately before the element it is ‚associated with' or between the subject and the verb.

5.4 There is also a difference between English and Czech concerning the fact that a focalizer may have a "backward" scope more frequently in English than in Czech. For example, the intonation center in the PEDT sentence *Scenario 1, known as the "Constant Dollar Freeze", reimburses the Pentagon for INFLATION only.*, if pronounced, would be placed on the word *inflation* (as indicated here by capitalization); the postposited focalizer *only* having its scope to the left. In the Czech translation of this sentence, the focalizer *jenom* has to be placed in front of the focused element. *Scénář 1, známý jako „konstantní zmrazení dolaru", nahrazuje Pentagonu výdaje jen kvůli INFLACI*. Our analysis has shown that in Czech, backward scope of focalizers is not that frequent as in English, but it is also possible.

5.5 As the previous paragraphs indicate, one of the main means of expression of topic-focus articulation in English is the placement of the intonation center. For Czech, it may be basically assumed that the unmarked placement is on the last element of the sentence and a pitch shift is realized only in marked situation. In English, due to the gramatically fixed character of the word order, the pitch may be more easily shifted to some non-final sentence element. As a matter of fact, there may be no other way how to mark the focus. Those who work with a written text then must always consider the analyzed sentence within the whole context, with an appropriate sentence prosody.

6. Summary

In our contribution, we want to demonstrate how a systematically annotated parallel corpus can be used to carry out a contrastive study of a linguistic phenomenon, in our case the information structure of the sentence. In the present abstract we could only briefly describe the material used and sketch the methodology, which in the full version of the paper will be documented by Czech and English authentic examples from the respective treebanks and by quatitative evidence.

References:

Cinková S. et al. (2009) Tectogrammatical Annotation of the Wall Street Journal. Prague Bulletin of Mathematical Linguistics, 92.

Čmejrek M., Cuřín J., Havelka J., Hajič J., Kuboň J. (2004). Prague Czech-English Dependecy Treebank: Syntactically Annotated Resources for Machine Translation. In 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal.

Hajič, J. et al. (2006). Prague Dependency Treebank 2.0. Linguistic Data Consortium, Philadelphia.

Hajičová E. (2010), Rhematizers Revisited. Linguistica Pragensia, Vol. 20 Issue: 2 57-70 Hajičová E , Partee B. and P. Sgall (1998), Topic-Focus Articulation, Tripartite Structures and Semantic Content, Dordrecht: Kluwer.

Mikulová, M. et al. (2005). Annotation on the Tectogrammatical Layer in the Prague Dependency Treebank: Annotation Manual. Universitas Carolina Pragensis, Prague.

Hajičová E. and J. Mírovský (in prep.), Contextual Boundness and Coreference – What the Prague Dependency Treebank Tells Us. Prepared for The Prague Bulletin of Mathematical Linguistics.

Mírovský J. and E. Hajičová (in prep.), Indefinite Subjects in English – A Corpus Probe. Prepared for Linguistica Pragensia.

Sgall, P., E. Hajičová and E. Buráňová, (1980). Aktuální členění věty v češtině [Topic-Focus Articulation in Czech], Prague:Academia

Sgall, P., E. Hajičová and J. Panevová (1986). The Meaning of the Sentence in Its Semantic and Pragmatic Aspect. Dordrecht: Reidel Publishing Company, and Prague: Academia.

Zikánová, Š. and M. Týnovský (2009) Identification of Topic and Focus in Czech: Comparative Evaluation on Prague Dependency Treebank. In: Zybatow, Gerhild et al. (eds.), Studies in Formal Slavic Phonology, Morphology, Syntax, Semantics and Information Structure. Formal Description of Slavic Languages 7, Frankfurt am Main: Peter Lang, pp. 343-353.

We' as a means of national identity construction in political discourse (a case study of the public political talk show `Shuster live` in Ukraine)

Olena Hlazkova
University of Alberta,
Edmonton, Canada
hlazkova@ualberta.ca

Key words: national identity, political discourse, media

Abstract

The modern Ukraine is undergoing serious social and political changes. Being a young post soviet state it is still in the process of self-formation and identification. Some Ukrainian historians and sociologists point out that it is difficult to understand the national identity type which is being formed in Ukraine now (Sumniv 2008: 484). The attempts to come up with a clear definition of national identity were made by Ukrainian and Russian researchers (Onyshkevych & Rewacowich 2009, Wolchik & Zviglianich 2000, Molchanov 2002, Palii 2005). The approach to study this question was limited either to historical aspect (Molchanov 2002) or to the cultural sphere and studies on modern literature (Gnatiuk 2005), based on the theory that national identity is formed by the cultural elite of the nation first. Both approaches are valid and important as one can not leave behind the historical or cultural aspect in the process of nation's self-identification. But these two theories do not give a full picture of such a complex and multilevel notion as national identity. Since national identity is "politically shaped" (Molchanov 2002: 10) in my research I will focus on the notion of national identity understood and declared by politicians who possess more power to influence and form social opinion.

For this study I have chosen the mediated political discourse in form of a political talk show, which enables analysis of the present-day situation in the country. The chosen programme release was dedicated to the topic of national security and national independence. While performing their monological speeches the politicians would use WE form to index