A Glimpse Under the Surface: Language Understanding May Need Deep Syntactic Structure

Eva Hajičová(✉)

Faculty of Mathematics and Physics Institute of Formal and Applied Linguistics,
Charles University in Prague,
Malostranské Nám. 25, 11800 Praha 1, Czech Republic
hajicova@ufal.mff.cuni.cz

Abstract. Language understanding is one of the crucial issues both for the theoretical study of language as well as for applications developed in the domain of natural language processing. As Katz (1969, p. 100) puts it "to understand the ability of natural languages to serve as instrument to the communication of thoughts and ideas we must understand what it is that permits those who speak them consistently to connect the right sounds with the right meanings." The proper task of linguistics consists then in the description (and) explanation of the relation between the set of the semantic representations and that of the phonetic forms of utterances; at the same time, among the principal difficulties there belongs "a specification of the set of semantic representations" (Sgall and Hajičová 1970, p. 5). In our contribution, we present arguments for the approach that follows the tradition of European structuralism which attempted at an account of linguistic meaning the elements of which are understood as "points of intersection" of conceptual contents (as a reflection of reality) and the organizing principle of the grammar of the individual language (Dokulil and Daneš 1958). In other words, we examine how "deep" the sematic representations have to be in order (i) to give an appropriate account of synonymy, and (ii) to help to distinguish semantic differences in cases of ambiguity (homonymy).

Synonymy can be understood as a relation between two sentences differing in a given opposition but having the same truth conditions, i.e. there does not exist a situation when one sentence would be true while the other sentence would not be true. A proof of non-existence, of course, is not possible, so that the statement of synonymy has always a nature of a hypothesis; this criterion helps us to decide for two suspicious sentences whether they are synonymous or not. Thus e.g. the sentences *Pavel sold Jirka a car. – Jirka bought a car from Pavel.* are not synonymous because if they are used in the context ... *with enthusiasm*, their meanings differ (in the first of them, the enthusiasm is on the side of Pavel, the seller, in the other on the side of Jirka, the buyer). Similar considerations hold for such pairs as Cz. *Jan si vzal Marii* (Jan married Mary) and *Jana si vzala Marie* (E. Jan-Acc married Mary-Nom.), if inserted into the context ... *for money*, or

K. Ekštein and V. Matoušek (Eds.): TSD 2017, LNAI 10415, pp. 3–7, 2017.
DOI: 10.1007/978-3-319-64206-2_1

for the non-synonymous sentences *I have read a letter about a quarrel of parents* and *I have read a letter about quarrelling parents.* because the continuation "... *but about their quarrel there was no mention in the letter*" is possible only with the second sentence. On the other hand, the pairs such as *He promised to do it in time* and *He promised he would do it in time* are considered to be synonymous: a context in which they would have different truth conditions has not yet been found (Panevová 1980).

Ambiguity is a notorious problem for both theoreticians and NLP researchers. The sources for ambiguity may be either lexical or they may lie in morphemics (e.g. the ambiguity between Nominative and Accusative in Cz: *Slepice honí kuřata* 'Hens chase chicken' where *slepice* and *kuřata* may be either Nom. or Acc. and the order is not decisive (who runs after whom?)), or in syntax (the well-known example *the criticism of the Polish delegate*: who criticized whom, or *the warning of the driver* – who warns?). The sources of amibiguities may accumulate in a single sentence – Cz. *Loví tlouště na višni* 'Catch(es) fish on morello/morello-tree': *loví* "catch" he-she-it-they, *tloušť* "(kind of) fish/" Acc sg/pl., *višni*: Loc. of *višeň* "morello" (fruit) or "morello-tree"; from the syntactic point of view there are several possible structural interpretations: Subj – Verb – Object – Loc: Subj is in the Location? Object is in the Location? (multiplied by the lexical ambiguity), or: Subj – Verb – Object – Instrument (?Manner) (= Morello as a bait put on a hook to catch fish).

A special attention in the paper will be paid to the case of synonymy/ambiguity/semantic differences related to **the information structure of sentences**. Among the examples discussed there are pairs of sentences such as *Everybody in this room knows at least two languages* vs. *At least two languages are known by everybody in this room*, or *Russian is spoken in Siberia* vs. *In Siberia one speaks Russian*, or *Tom only introduced Mary to Jane* vs. *Tom introduced Mary only to Jane*, or *Dogs must be CARRIED* (with the normal placement of intonation center at the end of the sentence, as denoted by the capitals) vs. *DOGS must be carried* (with the intonation center on DOGS) vs. *Carry dogs (CARRY dogs* vs. *Carry DOGS)*. Examples such as those document that if one wants to account in a consistent way for the semantic differences between sentences that on the surface look the same, it is necessary to postulate some kind of underlying structure (for a more detailed discussion of a formal account of information structure, see e.g. Hajičová et al. 1998).

Another support for this claim is the phenomenon of **surface deletions** (see Hajič et al. 2015; Hajičová et al. 2015). There belongs e.g. the phenomenon known recently in theoretical linguistics as a pro-drop parameter (called sometimes zero subject or null-subject). Czech belongs to the pro-drop type of language: the subject is often deducible from the morphology of the verb (*Přišel*-Masc. *domů* 'He came home' vs. *Přišla*-Fem. *domů* 'She came home') but due to the ambiguity of some verb endings this is not always the case (see above the sentence *Loví tlouště na višni* 'He-she-it-they catch(es) fish on morello-tree'). Other examples of surface deletions are infinitival constructions of the type: *John decided to leave Prague* (synonymous with *John decided that he would*

leave Prague) vs. *John recommended his friend to move to a better flat* (synonymous with *John recommended his friend that (he/she) moves to a better flat*), structures with comparison (*Paul knows a better lawyer than John:* meaning either ... *a better lawyer than John (is a lawyer)*, or ... *a better lawyer than John knows (a lawyer)*, or structures with the word '*kromě*' (besides): *Kromě Jany pozveme celou rodinu* (Besides Jane we will invite the whole family) which may mean either an addition (Jane will be invited (too)), or an exclusion (Jane will not be invited). Special problems are connected with deletions in structures with coordination (see Popel et al. 2013): it is not always clear which sentence elements are coordinated/deleted (cf. examples *red and white wine* vs. *Polish flag is white and red*, or *Romulus and Remus founded Rome* vs. *Michelangelo and Dante celebrated Rome*, or the ambiguity of the structure *sick and old people*: sick people [need not be old] and old people [need not be sick] vs. (both: sick and old) people.

The inclusion of an underlying (deep) level into the theoretical description of a language has led the research team of Prague theoretical and computational linguists to the postulation of a multilevel scheme in the theory of **Functional Generative Description** as proposed by Petr Sgall in the late sixties and developed since then by him and his pupils (for a most comprehensive treatment, cf. Sgall et al. 1986). This approach is also reflected in the proposal and build-up of the so-called **Prague Dependency Treebank** (PDT) for Czech, and the same scenario for the parallel annotation of the Prague Czech-English Dependency Treebank (PCEDT, with a two-level annotation of Czech and English; the original English texts are taken from the Penn Treebank, translated to Czech, see Hajič et al. 2011). The work on PDT started as soon as in the mid-nineties and the overall scheme was published already in 1998 (see e.g. Hajič 1998; for a detailed study on the treatment of some particular linguistic issues in PDT see Hajič et al. 2016). The basic idea was to build a corpus annotated not only with respect to the part-of-speech tags and some kind of (surface) sentence structure but capturing also the syntactico-semantic, underlying structure of sentences. The annotation is manual, and the "deep" syntactic dependency structure (with several semantically-oriented features, called "tectogrammatical" level of annotation) has been conceptually and physically separated from the surface dependency structure and its annotation, with full alignment between the elements (tree nodes) of both annotation levels being kept. The Prague Dependency Treebank consists of continuous Czech texts mostly of the journalistic style analyzed on three levels of annotation (morphological, surface syntactic and deep syntactic structure, including the annotation of the information structure of sentences, see Hajičová 2012). At present, the total number of documents annotated on all the three levels is 3,168, amounting to 49,442 sentences and 833,357 (occurrences of) nodes. The PDT version 1.0 (with the annotation of only morphology and the surface dependencies) is available from the Linguistic Data Consortium, as is the PDT version. Pronominal coreference is also annotated. Other additions (such as discourse annotation) appeared in PDT 2.5 and in PDT 3.0, which are both available from the LINDAT/CLARIN repository (Bejček et al. 2013).

The annotated corpus has a multifold exploitation. It is an indispensable resource for the study of particular linguistic phenomena in the given language, and, when a parallel corpus is available, also in comparison with other languages. The annotated material may serve as a basis for the compilation of lexicons (e.g. the VALLEX lexicon of Czech verbs with added information on valency of verbs, Lopatková et al. 2016 and the PDT-based lexicon of Czech verbs, Urešová 2011) and for the build-up of grammars (cf. Panevová et al. 2014). The annotation on the underlying, tectogrammatical level has also served as invaluable inspiration and data support for the build-up of some NLP applications (e.g. the Tecto-MT system, or the project Companions).

In our contribution, we will document that one of the basic features of this resource is its importance not only for the representation of the surface shape of the sentence but even more for the underlying sentence structure: it elucidates phenomena hidden on the surface but unavoidable for the representation of the meaning and functioning of the sentence.

Acknowledgement. This work has been supported by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071) and by the project No. GA17-07313S of the Grant Agency of the Czech Republic.

References

Bejček, E., et al.: Prague Dependency Treebank 3.0. Data/Software (2013)

Dokulil, M., Daneš, F.: K tzv. významové a mluvnickě stavbě věty [On the so-called semantic, grammatical structure of the sentence]. In: O vědeckém poznání soudobých jazyků, Praha: Nakladatelství Československé akademie věd, pp. 231–246 (1958)

Hajič, J.: Building a syntactically annotated corpus: the Prague dependency treebank. In: Hajičová, E. (ed.) Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová, pp. 106–132. Karolinum, Prague (1998)

Hajič, J., Hajičová, E., Mírovský, J., Panevová, J.: Linguistically annotated corpus as an invaluable resource for advancements in linguistic research: a case study. Prague Bull. Math. Linguist. **106**, 69–124 (2016)

Hajič, J., Hajičová, E., Mikulová, M., Mírovský, J., Panevová, J., Zeman, D.: Deletions and node reconstructions in a dependency-based multilevel annotation scheme. In: Gelbukh, A. (ed.) CICLing 2015. LNCS, vol. 9041, pp. 17–31. Springer, Cham (2015). doi:10.1007/978-3-319-18111-0_2

Hajič, J., Hajičová, E., Panevová, J., et al.: Prague Czech-English Dependency Treebank 2.0 (2011)

Hajičová, E.: What we have learned from complex annotation of topic-focus articulation in a large Czech corpus. Echo des Etudes Romanes **8**(1), 51–64 (2012)

Hajičová, E., Mikulová, M., Panevová, J.: Reconstructions of deletions in a dependency-based description of Czech: selected issues. In: Proceedings of the Third International Conference on Dependency Linguistics, Uppsala, pp. 131–140 (2015)

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

Katz, J.J.: The Philosophy of Language. Harper, New York (1969)

Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., Žabokrtský, Z.: Valenční slovník českých sloves VALLEX. [Valency Dictionary of Czech Verbs]. Karolinum, Praha (2016)

Panevová, J.: Formy a funkce ve stavbě české věty [Forms and Functions in the Structure of Czech Sentence]. Academia, Praha (1980)

Panevová, J., et al.: Mluvnice současné češtiny 2. Syntax češtiny na základě anotovaného korpusu. [Syntax of Czech on the Basis of an Annotated Corpus]. Karolinum, Prague (2014)

Popel, M., Mareček, D., Štěpánek, D., Zeman, D., Žabokrtský, Z.: Coordination Structures in dependency treebanks. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 517–527. Association for Computational Linguistics, Sofija (2013)

Sgall, P., Hajičová, E.: A "functional" generative description (background and framework). Prague Bull. Math. Linguist. **14**, 3–38 (1970). Revue roumaine de linguistique **16**, 9–37 (1971)

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

Urešová, Z.: Valence sloves v Pražském závislostním korpusu [Valency of verbs in the Prague Dependency Treebank]. UFAL, Prague (2011)