A Plea for Information Structure as a Part of Meaning Representation

Eva Hajičová

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague

hajicova@ufal.mff.cuni.cz

Abstract

The view that the representation of information structure (IS) should be a part of (any type of) representation of meaning is based on the fact that IS is a semantically relevant phenomenon (Sect. 2.1). In the contribution, three arguments supporting this view are briefly summarized, namely, the relation of IS to the interpretation of negation and presupposition (Sect. 2.2), the relevance of IS to the understanding of discourse connectivity and for the establishment and interpretation of coreference relations (Sect. 2.3). A possible integration of the description of the main ingredients of IS into a meaning representation is illustrated in Section 3.

1 Introduction

After the more or less isolated (though well substantiated) inquiries into the issues concerning one of the bridges between sentence form and its function in discourse (starting with the pioneering studies by Czech scholars in the first half of the last century followed by such prominent linguists and semanticists as M. A. K. Halliday, B. H. Partee, M. Rooth, E. Prince, K. Lambrecht, M. Steedman, E. Vallduví & E. Engdahl, to name just a few),[1] the last two decades of the last century witnessed an increasing interest of linguists in the study of information structure (IS). These approaches used different terms (theme-rheme, topic-focus, functional sentence perspective, presupposition and focus, background and focus, and a general term information structure (being the most frequent) and claimed to be based on different fundamental oppositions and scales (given - new, aboutness relation, activation or topicality scale) but all were more or less in agreement that this phenomenon, in addition to the syntactic structure of the sentence, is to be taken into account in an integrated description of the sentence and/or discourse, and that it significantly contributes to the study of the functioning of language.

The theory of information structure we subscribe to (cf. e.g. Sgall 1967; 1979; Sgall, Hajičová and Panevová 1986) called Topic-Focus Articulation (TFA) is based on the "aboutness" relation: the Focus of the sentence says something ABOUT its Topic. This dichotomy is based on the primary notion of contextual boundness (see below, Section 3) and its representation is a part of the representation of the sentence on its underlying (deep, tectogrammatical) syntactic level, which is assumed to be a linguistically structured level of meaning. In addition to the basic dichotomy the TFA theory works with a hierarchy of the so-called communicative dynamism, ie. an ordering of the meaningful lexical items (ie. items other than function words) of the sentence from the least communicatively important elements of the sentence to the elements with the highest degree of communicative importance. The TFA is considered to be a recursive phenomenon, which makes it possible to recognize – aside with the global Topic and the global Focus and based on the features of contextual boundness – also local topics and local foci. In this way, the TFA framework offers a possibility, if needed, to recognize more distinctions in addition to the basic dichotomy (as done, e.g. by the focus – background approach of Vallduví and Engdahl (1996), or as needed, according to e.g. Büring (1997) or Steedman (2000), for a proper account of prosody).

[1] For the bibliographical references, see the Section References at the end of the paper.

Proceedings of the First International Workshop on Designing Meaning Representations, pages 66–72
Florence, Italy, August 1st, 2019

2 Information Structure as a Semantically Relevant Phenomenon

2.1 Basic argument

The crucial argument in support of an inclusion of the representation of information structure into a representation of meaning relates to the fact that IS is *semantically relevant*, as can be documented by examples (1) to (3), taken from early literature on these issues (the capitals denote the intonation center).

(1) (a) Dogs must be CARRIED.

(a') CARRY dogs.

(b) DOGS must be carried.

(b') Carry DOGS. (Halliday 1967)

(2) (a) English is spoken in the SHETLANDS.

(b) In the Shetlands, ENGLISH is spoken. (Sgall 1967)

(3) (a) Mary always takes John to the MOVIES.

(b) Mary always takes JOHN to the movies. (Rooth 1985)

For the sake of simplicity, let us reduce here the more differentiated approach of TFA into an articulation of the sentence into its Topic (what is the sentence about) and Focus (what the sentence says about its Topic). Then it can be easily seen that the (a) and (b) sentences in the above sets (capitals indicating the intonation center) differ in this articulation and, correspondingly, differ in their meaning: (1)(b) is non-sensical (one can use the underground elevator also without a dog), (2)(a) even false (English is spoken in other countries as well) and (3)(a) and (b) reflect different situations in the real world (It is always the movies where John is taken vs. It is always John who is taken to the movies). In the surface shape of the sentences, the different interpretations of the (a) and (b) sentences in each set are rendered by different surface means, such as word order or the position of the intonation center, but have to be accounted for in the representation of their meaning if the sentences have to receive the appropriate corresponding reading. For an example from a typologically different language with a rather flexible word order, cf. the Czech equivalents of the sentences (1) through (3), with the assumed prototypical placement of the intonation center at the end of the sentence (indicated again by capitals).

(1') (a) Psy neste v NÁRUČÍ.

(b) V náručí neste PSY.

(2') (a) Anglicky se mluví na Shetlandských OSTROVECH.

(b) Na Shetlandských ostrovech se mluví ANGLICKY.

(3') (a) Marie bere Honzu vždy do KINA.

(b) Marie bere do kina vždy HONZU.

2.2 Negation and presuppostion

Semantic relevance of IS is attested also by the analysis of the semantics of *negation* and of the specification of the notion of *presupposition*. If IS of a sentence is understood in terms of an aboutness relation between the Topic of the sentence, then in the prototypical case of negative sentences, the Focus does not hold about the Topic; in a secondary case, the negative sentence is about a negated topic and something is said about this topic.[2] Thus, prototypically, the sentence (4) is about John (Topic) and it holds *about* John that he didn't come to watch TV (negated Focus).

(4) John didn't come to watch TV.

However, there may be a secondary interpretation of the negative sentence, e.g. in the context of (5).

(5) John didn't come, because he suddenly fell ill.

One of the interpretations of (5) is that the sentence is *about* John's not-coming (Topic) and it says about this negated event that is happened because he suddenly fell ill (Focus).

As Hajičová (e.g.1973; 1984) documented, there is a close relation between IS, negation and presupposition (see the original analysis of presupposition as a specific kind of the entailment

[2] An objection that one cannot speak about a non-existent topic does not arise: one can speak about an absence as well as about not-coming, not-visiting (cf. Strawson's example below), etc. See also Heim's treatment of the definite-indefinite noun phrases and her notion of file change semantics in which meanings are analyzed as context-change potentials (Heim 1982; 1983). See also the pioneering study of the relation between theme-rheme and negation by Zemb (1968).

relation by Strawson (1952) and Strawson's (1964) notion of referential availability in his analysis of the sentence *The exhibition was visited by the King of France*. and its negation):

(6) (a) John caused our VICTORY.

(b) John didn't cause our VICTORY.

(c) Though he played well as usual, the rest of the team was very weak (and nothing could have prevented our defeat).

(7) (a) Our victory was caused by JOHN.

(b) Our victory was not caused by JOHN.

(8) We won.

Both (6)(a) and (7)(a) imply (8). However, it is only the negative counterpart of (7)(a), namely (7)(b), that implies (8), while (6)(b) may appear also in a context suggesting that we were defeated, see (6)(c). In terms of presuppositions, the statement (8) belongs to the presuppositions of (7)(a) since it is entailed both by the positive as well as by the negative sentence, but not to the presuppositions of (6)(a) as it is not entailed by the negative sentence.[3]

2.3 Discourse connectivity

Another phenomenon, though going beyond the interpretation of a single sentence but important for the interpretation of a text (discourse), is *discourse connectivity*. There have been several proposals in literature how to account for these relations, the *centering* theory being one of the most deeply elaborated (cf. Grosz, Joshi and Weinstein, 1983 and its corpus-based evaluation in Poesio et al. 2004). It is based on the model of the local attentional states of speakers and hearers as proposed by Grosz and Sidner (1986). Each utterance in discourse is considered to contain a backward looking center, which links it with the preceding utterance, and a set of entities called forward looking centers; these entities are ranked according to language-specific ranking principles stated in terms of syntactic functions of the referring expressions. Related treatment rooted in the Praguian traditional account of IS is the idea of so-called thematic progressions (Daneš 1970), explicitly referring to the relation between the theme (Topic) and the rheme (Focus) of a sentence and the theme (Topic) or the rheme (Focus) of the next following sentence (a simple linear thematic progression and a thematic progression with a continuous theme), or to a 'global' theme (derived themes) of the (segment of the) discourse. As demonstrated in Hajičová and Mírovský (2018a), an annotation of a text (corpus) in terms of Topic and Focus makes it possible to find these links between sentences and in this way to account for the structure of discourse. In a similar vein, it has been demonstrated that a meaning representation including some basic attributes of IS serves well for an establishment and interpretation of coreference relations (Hajičová and Mírovský 2018b).

3 Information Structure in an Annotated Corpus

The observations documenting the semantic relevance of the information structure (Sect. 2.1 and 2.2 above) indicate that the information structure (Topic-Focus articulation) of the sentence belongs to the domain of the (syntactico-) semantic structure of the sentence rather than exclusively to the domain of discourse (or, in more general terms, to the domain of pragmatics), as sometimes claimed. However, this is not to deny the interrelation or interaction between the two domains and, as illustrated in Section 2.3, the inclusion of the basic features of IS into the representation of meaning may serve well also for the description of the structure of discourse.

In this final section of our paper we present an example of the annotation scenario illustrating how IS is represented in the Praguian dependency-based sentence representations. For a simplified example of such a representation for sentences in (1), see the Appendix.

The overall annotation scenario includes three levels: (a) morphemic (with detailed part-of-speech tags and rich information on morphological categories), (b) surface shape ("analytical", in the form of dependency-based tree structures with the verb as the root of the tree and with relations labeled by superficial syntactic

[3] The specific kind of entailment illustrated here by the above examples was introduced in Hajičová (1972) and called *allegation*: an allegation is an assertion A entailed by an assertion carried by a sentence S, with which the negative counterpart of S entails neither A nor its negation (see also the discussion by Partee 1996).

functions such as Subject, Object, Adverbial, Attribute, etc.), and (c) underlying dependency-based syntactic level (so-called tectogrammatical) with dependency tree structures labeled by functions such as Actor, Patient, Addressee, etc. and including also information on the IS (Topic-Focus articulation) of sentences.[4] For this purpose, a special TFA attribute is established in the scenario for the representation of a sentence on the tectogrammatical level, with three possible values, one of which is assigned to every node of the tree; these values specify, whether the node is contextually bound non-contrastive, contextually bound contrastive, or contextually non-bound. A contextually bound (*cb*) node represents an item presented by the speaker as referring to an entity assumed to be easily accessible by the hearer(s), i.e. more or less predictable, readily available to the hearers in their memory, while a contextually non-bound (*nb*) node represents an item presented as not directly available in the given context, cognitively 'new'. While the characteristics 'given' and 'new' refer only to the cognitive background of the distinction of contextual boundness, the distinction itself is an opposition understood as a grammatically patterned feature, rather than in the literal sense of the term. This point is illustrated e.g. by (9).

(9) (Tom entered together with his friends.) My mother recognized only HIM, but no one from his COMPANY.

Both Tom and his friends are 'given' by the preceding context (indicated here by the preceding sentence in the brackets), but in the given sentence they are structured as non-bound (which is reflected in the surface shape of the sentence by the position of the intonation center).

The appurtenance of an item to the Topic or Focus of the sentence is then derived on the basis of the features *cb* or *nb* assigned to individual nodes of the tree (see Sgall 1979):

(a) the main verb (V) and any of its direct dependents belong to F iff they carry index *nb*;

(b) every item that does not depend directly on V and is subordinated to an element of F different from V, belongs to F (where "subordinated to" is defined as the irreflexive transitive closure of "depend on");

(c) iff V and all items directly depending on V are *cb*, then it is necessary to specify the rightmost *k' node* of the *cb* nodes dependent on V and ask whether some of nodes *l* dependent on *k'* are *nb*; if so, this *nb* node and all its dependents belong to F; if not so, then specify the immediately adjacent (i.e. preceding) sister node of *k'* and ask whether some of its dependents is *cb*; these steps are repeated until an *nb* node depending (immediately or not) on a *cb* node directly dependent on V is found. This node and all its dependent nodes are then specified as F.

(d) every item not belonging to F according to (a) - (c) belongs to T.

This algorithm has been implemented and is applied in all experiments connected with research questions related to IS.

As described in Zikánová et al. (2009), the SH algorithm was applied to a part of the PDT data (about 11 thousand sentences). The results indicate that a clear division of the sentence into Topic and Focus according to the hypothesized rules has been achieved in 94.28% of sentences to which the procedure has been applied; 4.41% of sentences contained the type of focus referring to a node (or nodes) that belong(s) to the communicatively most dynamic part of the sentence though they depend on a contextually bound node. The real problem of the algorithm then rests with the case of ambiguous partition (1.14%) and cases where no focus was recognized (0.11%). In Rysová et al. (2015) some of the shortcomings of the previous implementation described in Zikánová et al. (2009) were removed and the algorithm was evaluated in a slightly different way: as the gold data we used data annotated by a linguist assuming that the results would better reflect the adequacy of the algorithm for transforming values of contextual boundness into the division of the sentence into the Topic and the Focus. Our gold data consisted of 319 sentences from twelve PDT documents annotated by a single linguistic expert. Without taking into account (already annotated but now hidden) values of contextual boundness, the annotator

[4] In addition, two kinds of information are being added in the latest version of PDT, namely annotation of discourse relations based on the analysis of discourse connectors (inspired by the Pennsylvania Discourse Treebank) and information on grammatical and on textual intra- and inter-sentential coreference relations.

marked each node as belonging either to the Topic or to the Focus. On these gold data, the new implementation of the algorithm was evaluated, see Table 1.[5]

Measure	SH Algorithm
F1-measure in topic	0.89
F1-measure in focus	0.95
overall accuracy on tectogrammatical nodes	0.93
overall accuracy on whole sentences	0.75

Table 1: Evaluation of the SH algorithm.

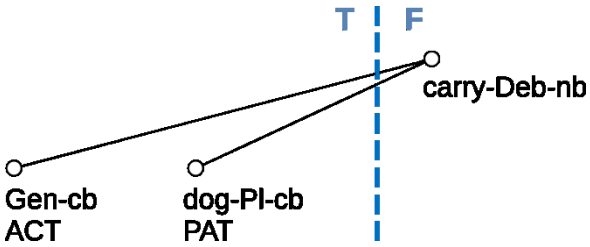
4 Summary

In our contribution we argue that a meaning representation of any type should include information on basic features of information structure. Our argument stems from the fact that information structure (at least the articulation of a sentence into its Topic and Focus) is semantically relevant which is demonstrated on several examples, taking into account also the representation of negation and presupposition. An inclusion of the representation of information structure into an overall representation of meaning also helps to account for some basic features of discourse connectivity and coreference relations. In the Appendix, we have briefly characterized one possible way of representation of the basic features of information structure.

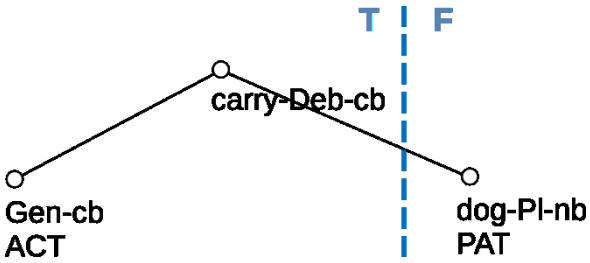
5 Appendix

To attest the plausibility of a representation of IS in an annotated corpus, we present here rather simplified representations of the sentences given above in (1). The symbols ACT, PAT and Gen stand for the deep syntactic functions Actor, Patient and General Actor, respectively, Deb(itive) and Imper stand for deontic and sentential modality, and cb and nb stand for the contextually bound and contextually non-bound values of the TFA attribute. The vertical dotted line denotes the boundary between Topic and Focus.

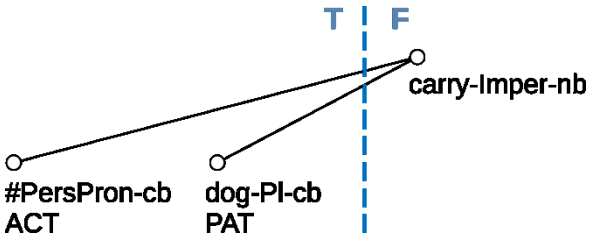
(1) (a) Dogs must be CARRIED.



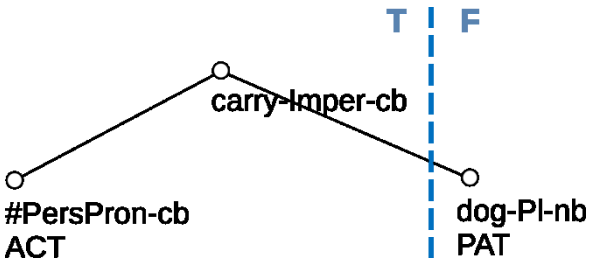
(1) (b) DOGS must be carried.



(1) (a') CARRY dogs.



(1) (b') Carry DOGS.



Acknowledgments

The author gratefully acknowledges support from the Ministry of Education of the Czech Republic (project LM2018101 – LINDAT/CLARIAH-CZ). I am also very much grateful to Jiří Mírovský for his help with the formatting of the text.

[5] It significantly outperformed the baseline, which was defined as follows: in the linear (surface) form of the sentence, each word before the autosemantic part of the predicate verb belongs to Topic, the rest of the sentence belongs to Focus.

References

Bäuerle, R., Schwarze C. & A. von Stechow, Eds. (1983), *Meaning, Use and Interpretation of Language.* Berlin: Mouton de Gruyter.

Bŭring, D. (1997). *The Meaning of Topic and Focus – The 59th Street Bridge Accent*. Londoin: Routledge.

Cole, P., Ed.. (1981*). Radical Pragmatics*. New York: Academic Press.

Daneš, F. (1970). Zur linguistischen Analyse der Textstruktur. *Folia linguistica* 4, 1/2, 72-78.

Grosz, B. & Sidner, C. L. (1986). Attention, Intentions and the structure of discourse. *Computational Linguistics, 12,* 175–204.

Grosz, B. J., Joshi, A. K. & S. Weinstein (1995). Centering: A Framework for modeling the local coherence of discourse. *Computational Linguistics, 21*, 203-225.

Hajič, J., Bejček, E., Bémová, A., Buráňová, E., Hajičová, E., Havelka, J., Homola, P., Kárník, J., Kettnerová, V., Klyueva, N., Kolářová, V., Kučová, L., Lopatková, M., Mikulová, M., Mírovský, J., Nedoluzhko, A., Pajas, P., Panevová, J., Poláková, L., Rysová, M., Sgall, P., Spoustová, J., Straňák, P., Synková, P., Ševčíková, M., Štěpánek, J., Urešová, Z., Vidová Hladká, B., Zeman, D., Zikánová, Š. & Žabokrtský, Z. (2018). Prague Dependency Treebank 3.5. Institute of Formal and Applied Linguistics, LINDAT/CLARIN, Charles University, LINDAT/CLARIN PID: http://hdl.handle.net/11234/1-2621

Hajičová E. (1973), Negation and topic vs. comment. *Philologica Pragensia* 17, 18-25. Reprinted in Hajičová (2017), 50-62.

Hajičová E. (1984), Presupposition and allegation revisited. *Journal of Pragmatics* 8:155-167. Reprinted in Hajičová (2017), 63-77.

Hajičová, E. (2017). *Syntax-Semantics Interface*. Prague: Karolinum

Hajičová E. & J. Mírovský (2018a), Discourse Coherence Through the Lens of an Annotated Text Corpus: A Case Study. In: Proceedings of LREC 2018, Miyazaki, Japan.

Hajičová E. & J. Mírovský (2018b), Topic/Focus vs. Given/New: Information Structure and Coreference Relations in an Annotated Corpus. Presented at the 2018 Annual Conference of Societas Linguistica Europaea, Tallin, Latvia.

Halliday, M. A. K. (1967a). *Intonation and Grammar in British English*. The Hague: Mouton.

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

Heim, I. (1982). *The semantics of definite and indefinite noun phrases*. PhD Thesis, Univ. of Massachusetts, Amherst.

Heim, I. (1983). File chnge semantics and the familiarity theory of definiteness. In: Bäuerle, Schwarze & von Stechow, Eds. (1983), 164-189.

Lambrecht, K. (1994). *Information Structure and Sentence Form. Topic, Focus and the Mental Representations of Discourse Referents*. Cambridge: Cambridge University Press.

Partee, B. H. (1996), Allegation and local accommodation. In: Partee & Sgall, Eds. (1996), 65-86.

Partee, B. H. & Sgall, P., Eds. (1996). *Discourse and meaning.* Amsterdam/Philadelphia: John Benjamins.

Prince, E. (1981). Toward a taxonomy of given/new information. In: Cole, Ed.. (1981), 223-254.

Rooth, M. (1985). *Association with focus*. PhD Thesis, Univ. of Massachusetts, Amherst.

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

Sgall P. (1967), Functional Sentence Perspective in a generative description of language. *Prague Studies in Mathematical Linguistics* 2, Prague, Academia, 203-225.

Sgall P. (1979), Towards a Definition of Focus and Topic. *Prague Bulletin of Mathematical Linguistics* 31, 3-25; 32, 1980, 24-32; printed in *Prague Studies in Mathematical Linguistics* 78, 1981, 173-198.

Sgall, P., Hajičová, E. & and J. Panevová (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* Dordrecht: D. Reidel.

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

Steinberg, D. D. & Jakobovits, L. A. Eds. (1971). *Semantics. - An interdisciplinary reader*. Cambridge, Mass.: Cambridge University Press.

Strawson, P. (1952). *Introduction to Logical Theory*. London: Methuen

Strawson, P. (1964). Identifying reference and truth values. *Theoria 30*, 96-118. Reprinted in Steinberg & Jakobovits, Eds. (1971), 86-99.

Vallduví, E. & Engdahl, E. (1996). The linguistic realization of information packaging. *Linguistics* 34, 459-519.

Zemb, J-M. (1968). *Les structures logiques de la proposition allemande. Contribution à l'étude des rapports entre la langue et la pensée*. Paris: O.C.D.L.

Zikánová, Š. & M. Týnovský (2009). Identification of Topic and Focus in Czech: Comparative Evaluation on Prague Dependency Treebank. In: Zybatow, G. 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., 343–353.