The fundamental significance of information structure

Eva Hajičová and Petr Sgall

1. Remarks on semantics and pragmatics

Jacob Mey has contributed most effectively to recognizing the importance of pragmatics. Even a specification of truth conditions (as Carnap's propositions, or in a situational framework) requires the reference of certain items in the sentence to be determined. Thus, pragmatic factors have to be reflected as fundamental and included into the procedure of interpretation. This also concerns tense and modalities, as well as information structure or **Topic-Focus Articulation** (TFA).

Many linguistic approaches are crucially mistaken underestimating the fundamental position of TFA. However, already the Aristotelian notions may be understood as referring to Topic (T) and Focus (F): τὸ ὑποκείμενον, 'the given circumstances', and τὸ κατηγορούμενον, 'the enounced'. Natural language differs from the calculi of logic or from programming languages due to its interactivity, to the contextual anchoring of utterances (sentence occurrences). The speaker does not just tell someone something, but rather s/he tells the addressee(s) something **about** something. If the content of an utterance is seen as an operation on the hearer's memory, then the interpretation should reflect that a declarative sentence asserts that its F holds about its T. TFA reflects the 'given-new' strategy of communication, but differs from it belonging to individual languages, rather than to the domain of cognition. This follows from the semantic relevance of TFA (cf. Section 3) and from the differences in the means expressing TFA: although the placement of the intonation center always is relevant, means such as word order, morphemic items, or syntactic constructions differ from one language to the other. Thus, the interactivity of communication is reflected not only in the patterning of discourse, but even in the structure of a sentence as a type, as a unit of language (*langue*, linguistic competence, I-language).

The position of TFA in language remains undetermined if a specific level of information structure is postulated or if two relevant dichotomies are distinguished without systematically analysing the relevance of such a division for the scope of negation and of other focusing operators (focalizers). A comparison of different approaches to information structure would deserve a specific study. Let us just recall that in Prague, the interest in TFA has been strong for decades, especially thanks to V. Mathesius and J. Firbas. Research in the descriptive framework of the Functional Generative Description (**FGD**) and discussions on its comparison with other views have convinced us that TFA is more basic for sentence structure than its predicate-argument pattern is (be the latter understood as based on constituents or on dependency), although in grammar TFA has been disguised into the opposition of subject and predicate as patterned by morphemics and/or by word order. We argue in Sect. 2 that the main requirements on the description of TFA can be met in the theoretical framework of FGD (see Hajičová, Partee and Sgall 1998; Petkevič 1995). FGD applies the following principles:

(A) it stresses the opposition of unmarked (primary, prototypical) items of all levels as opposed to their **marked** (secondary, peripheral) counterparts, which occur under specific contextual or other conditions;

(B) it works with **dependency**, i.e. a set of relations between a head word and its dependents; the underlying or **tectogrammatical** representation (TR) of a sentence is based on the verb (V) with its valency (obligatory and optional arguments and adjuncts).

Prototypical TRs are dependency **trees**, i.e. rooted trees the edges of which render dependency relations, their nodes being linearly ordered. Only other than dependency relations (coordination, apposition) require networks of more than two dimensions. A TR thus comprises three orderings, two of which are defined on the set of all its nodes: (i) a partial ordering, interpreted as the dependency relations, (ii) a linear ordering interpreted as the underlying word order (communicative dynamism, see below), i.e. as primarily corresponding to the temporal progression of the utterance, (iii) an optional partial ordering defined on individual subparts of the TR, interpreted as coordination or apposition.

The **orientation** of a dependency relation may be defined so that in the 'endocentric' constructions, e.g. *go slowly* or *old man,* the head cannot be deleted without loosing well-formedness, and 'exocentric' constructions are similarly specified on the level of word classes: e.g. V occurs both in constructions with an object and without it.

2. The position of information structure in language

Every node of a TR is labelled by a symbol composed of a lexical part, a morphological index (indicating the values of number, tense, modalities, etc.), and a symbol indicating a dependency relation (arguments, i.e. Actor, Objective, Addressee, Origin and Effect, and adjuncts, such as Locative, temporal and directional relations, Cause, Condition, Means, etc.), and possibly by an index for a CB root and another for a contrastive node. The set of TRs can be specified by a restricted set of very general rules, either

(i) in a declarative way, using unification and checking the conditions specified in the valency frames of the head words – the presence of each obligatory dependent, the order of dependents in F (see below on systemic ordering), the saturation of every argument (present just once as depending on one head),

(ii) or as a generative procedure, not stronger than a pushdown-store generator, deriving the tree from top to bottom and from left to right.

Since dependency based syntax describes embedded clauses as dependent on an item within their governing clauses, the TFA of a **complex** sentence may be rendered without any substantial additions to the framework. Each of the coordinated clauses in a **compound** sentence is handled as having its own TFA. The left-to-right order of nodes in TRs (communicative **dynamism**, CD) starts with T proper and proceeds to F proper (most dynamic), bearer of the (typically falling) intonation centre. Differences between CD and the surface order are due e.g. to shallow rules such as "adjective before noun", or to placing F in a marked (not clause-final) position.

The TRs are prototypically disambiguated, although peripheral cases such as marked patterns of quantifier scopes (branching, group reading) are left underspecified. Similarly as context based accommodation or bridging coreference, they belong to the domain of natural language inferencing, i.e. to individual abilities, rather than to linguistic competence. The TRs can be understood as 'the **meaning** of the sentence' (or as determining the meaning) in that their set is appropriate as the input for semantico-pragmatic interpretation. If the reference of the referring expressions is specified and different cases of figurative meaning are taken into account, a Carnapian proposition (sentence intension) is determined.

A prototypical TR is a **projective** tree: for every triple of nodes *a, b, c,* if *a* depends on *b*, and *c* is placed to the left of *a* (*b*) and to the right of *b* (*a*), then *c* is subordinated to *b*. Thanks to the projectivity and to similar conditions holding for combinations of dependency with coordination and apposition, a TR can be **linearized**, with every dependent enclosed in a pair of parentheses (an index of which at the parenthesis oriented towards the head indicates the kind of the dependency or other relation), see exx. (1)(b) and (21). The linearization can capture not only binary relations, but also cases of coordination of any number of items, possibly having any number of dependents.

This view of sentence structure is being tested and enriched in the program of the Prague Dependency Treebank (PDT), see Böhmová et al. (2003), which comprises three layers of annotation:

(i) the **morphemic** layer with about 2000 tags (1100 are actually used) for the highly inflectional Czech language, assigned stochastically, with a success rate higher than 95%;

(ii) an intermediate auxiliary layer of '**analytic**' ("surface") syntax, in which all surface wordforms and punctuation marks are represented by nodes of dependency trees: approximately 100 000 Czech sentences have been annotated by a semi-automatic procedure;

(iii) underlying **syntax**: tectogrammatical tree structures (TGTSs) with function words (preposition groups, periphrastic verb forms, etc.) as parts of complex word forms, and with a more subtle classification of objects, adverbials and attributes, including a TFA annotation.

3. Topic-Focus articulation and semantics

The semantic relevance of TFA is based on the relation of **aboutness**, i.e., the interpretation of a sentence consists not just of its predicate-argument pattern, but rather on a pattern with F predicated of T, i.e. F(T), with the negative counterpart ~F(T). This aspect of interpretation was discussed in the frame of intensional semantics by B. H. Partee in Hajičová et al. (1998:48-53; see also J. Peregrin's analyses commented on by B. H. Partee in the quoted book, and Partee 1998). TFA is semantically relevant in the general case: It not only determines in which contexts the sentence may occur, but is also decisive for the truth conditions (in this we differ e.g. from Fintel 2004). The examples (1) and (2) (see the TR of (1)(b) in Fig. 1 and its simplified linearized form in (1)(b')) point out that neither the presence of negation (or of another focalizer), nor that of quantifiers such as *many*, *few* is necessary for two sentences that differ just in their word order or in the position of the intonation center to carry different truth conditions. The operations on the addressees' states of minds clearly differ from each other.

.

(1) (a) I work on my dissertation on Sundays.
(b) On Sundays, I work on my dissertation.

(1)(b') (*Sunday*.Plur)$_{\text{TEMP}}$ (*I*)$_{\text{ACT}}$ *work*.Pres.Decl ($_{\text{OBJ}}$ (*my*)$_{\text{RSTR}}$ *dissertation*)

```
                         work.Pres.Decl
                           /    /      \
                          /    /        \
          Sunday.Plur.Indef.TEMP  I.ACT   dissertation.Sing.Def.OBJ
                                               /
                                              /
                                          my.RSTR
```

Figure 1.
A simplified TR of (1)(b).

If the impact of TFA on interpretation is reflected, then a **presupposition** can be distinguished from an **allegation** (as two kinds of entailment, met by subcollections of possible worlds for assertions carried by individual utterances): an allegation is an assertion A entailed by an occurrence of sentence S such that the **negative** counterpart of S entails neither A nor its negation (see Partee 1996). Often a definite noun group triggers a presupposition if it occurs in T, but only an allegation if it belongs to F, cf. the discussions on exx. (2) and (3).

(2) The King of France is (not) bald
(3) The exhibition was (not) visited by the King of France

In Hajičová et al. (1998) an application of **Tripartite Structures** (Operator, O – Restrictor, R – Nuclear Scope, N) on the interpretation of TFA was characterized, with R correspponding to T, N to F and O to the assertive modality of V, or to a focalizer.

The opposition of **contextually bound** (CB) and non-bound (NB) items may be handled as an elementary concept. It is to be understood as grammatically patterned, rather than in the literal sense, cf. ex. (4), in which we understand *company* as NB (*n*), while *my*, *mother* and *his* are CB (*b*); we denote the intonation center by capitals if it has a marked position. Kruijff-Korbayová and Steedman's (2003) *background* comes close to our concept CB:[1]

(4) (Tom entered together with his friends.) My.*b* mother.*b* recognized.*n/b* only.*n* HIM.*n*, but no.*n* one.*b* from his.*b* COMPANY.*n*.

In a TR, a node depends on its head either from the left or from the right (i.e. as CB or NB), only the root (typically NB) depends on no head. The speaker refers by CB items to entities assumed to be easily accessible by the hearer(s), prototypically 'given'. They refer to 'established' items, mentioned in the preceding co-text and thus still sufficiently salient, to indexicals, or to permanently established items given by culture or technical domain. NB items are presented as not directly predictable, as 'new' information. Thus, the accented *HIM* in (4) is NB, referring to a subject chosen from a set of alternatives, not directly predictable although known from the context, which admits the anaphoric pronoun to be used.

Operational **criteria** for the distinction of NB and CB items, such as the question test (see below) or the use of strong (stressed) forms of pronouns, have been discussed in the writings quoted above. **Ambiguity** between NB and CB is frequent; cf. e.g. *recognized* in ex. (4). A CB node precedes its mother node and its NB sisters, and a NB node follows them (exceptions include focalizers, see Hajičová et al. 1998, 134ff). V and its dependents belong to F iff they are NB; so does every other item subordinated to an element of F (where 'subordinated' means the transitive closure of 'depends'), certain exceptions being connected with a 'quasi focus,' i.e. a CB item differing from V to which a part of F is subordinated. All items not belonging to F belong to T.

On the **surface**, both in Czech and in English, in unmarked cases, V and its dependents that follow it belong to F, and the items preceding V are parts of T. In marked cases, V can be CB, i.e. in T, or (a part of) F may precede V; the intonation centre (sentence stress) marks F. The dependents of nouns primarily are NB.

Prototypically, NB items are included in focus (F), and CB items belong to topic (T). The order of the nodes within F is fixed, meeting the canonical **systemic ordering**, with e.g. the adjunct of Means (and also Directional.*from*) preceding Directional.*where-to* (by *n/b* we denote ambiguity of NB/ CB):

(5)(a) We.*b* went.*n/b* by car.*n/b* from Spain.*n/b* to France.*n*.
(b) We.*b* went.*n/b* to France.*b* by car.*n/b* from Spain.*n*.

Such a primary order can be found also for other pairs of dependents; see Sgall et al. (1995) for an examination of German, Czech, and partly also English.[2]

Our analysis is not restricted to cases in which either T or F would correspond to a single constituent, as illustrated e.g. by those readings of (5)(a) in which the subject (Actor) and V constitute T (being CB), while F consists of the three adverbials (which are NB).

Thus, TFA can be handled with the use of a single opposition of T and F. The discrepancy between the single relationship of aboutness as the basis of the semantic value of TFA and the two dichotomies assumed to constitute the information structure in certain other approaches is absent.

The **question test** is one of the relevant operational criteria; e.g. in some readings of (5)(a) only the subject is CB, since they can answer the question *What did they do?* In one of the readings of (5)(b) *from Spain* is understood as F (as the single NB item), since this is the counterpart of the interrogative element in the question *From where did they move to France by car?*, which (5)(b) can answer; the rest of the sentence, known from the question, is its T.

In the flow of **discourse**, (a part of) T may be referentially identical (or semantically associated) either to T or to F of the preceding sentence, or T is chosen from another part of the set of established items than from those referred to previously. T then may carry the L+H* pitch accent; we see **contrastive** items in T (or contrastive CB items),[4] which Hajičová et al. (1998, 151) discuss in connection with focalizers. However, a contrastive topic also occurs without a focalizer. Thus, in (6) *in Slovakia* is contrasted with the F of the first conjunct.[5]
:

(6) (Is Czech spoken in Bohemia or in Slovakia?) Czech is spoken in BOHEMIA, and in Slovakia, SLOVAK is spoken.

4. The simple pattern of the core and the vast periphery of language

As the linearization of TRs shows, the **core** of sentence structure can be understood as based on a simple pattern, coming close to propositional calculus, and thus to systems that correspond to common **human mental abilities**, assumed to be innate on independent reasons. This might help to explain the relative easiness of the child's acquisition of the core of language, without postulating a complex framework of specific innate features. The non-prototypical, **marked** phenomena constitute a vast and complex periphery of language, with the following layers:

(i) marked members of grammatical oppositions within the language core, such as the values of Plural, Preterite, or, within TFA, the CB phenomena;

(ii) peripheral phenomena in the TRs (e.g. coordination and apposition, marked positions of focalizers), which require more complex descriptive devices;

(iii) contextually restricted relations between TRs and morphemic, phonemic and phonetic representations of sentences – a very large domain, going from ambiguous and synonymous items in the lexicon and in morphemics (inflectional paradigms, their irregularities, etc.) to instances of surface word order not corresponding directly to CD.

The complex, large peripheral domains can be mastered by children step by step; the specific, contextually restricted rules and exceptions are internalized one after the other, on the basis of analogy. A theoretical description may capture the core of language by relatively weak means (equivalent to a context-free grammar, cf. our TRs), accompanied by models of the non-prototypical phenomena.

Notes

1 We use English as well as Czech examples to illustrate that the main features of TFA and of the TRs are identical in typologically different languages. Altough their repertoires of morphological categories (definiteness, verbal aspect, etc.) differ, most of the diversity of languages concerns the relationships between (underlying) representations and the surface layers. Most of our Czech examples are taken from PDT.

2 In a TR that comprizes no T (i.e. corresponds to a thetic judgment), V is least dynamic. This often is reflected by the surface word order in Czech, and even in English, in which the filler *there* is then used.

3 It deserves further discussion to which extent other clausal operators have similar effects as focalizers and which sentence adverbials behave in such a way.

[4] Cf. the 'topic shift' of the centering theory of B. Grosz, A. Joshi and C. Sidner. Perhaps the 'hat contour' (see Steube 2001) is always connected with a contrastive T. – It is useful to examine the prosody not just of a single word form, but of a longer part of the utterance, as Veselá et al. (2003) do.

[5] A contrastive part of T typically starts the sentence in Czech and belongs to the *Vorfeld* in German; see also Lahousse (2003) for French.

References:

Böhmová, A., J. Hajič, E. Hajičová and B. Hladká (2003). The PDT – A 3-level annotational scenario. In: *Building and Using Parsed Corpora* (A. Abeillé, ed.), pp. 103 - 127. Kluwer Academic Publishers, Dordrecht/Boston/ London.

Fintel, K. von (2004). A minimal theory of adverbial quantification. In: *Context-Dependence in the Analysis of Linguistic Meaning* (H. Kamp and B. H. Partee, eds.), pp. 177-186. Elsevier, Amsterdam.

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

Kruijff-Korbayová I. and Steedman M.: Discourse and information structure. In: *Proceedings of the ESSLLI workshop on information structure and discourse semantics.* Helsinki. Dordrecht: Kluwer, 2003, 1-12.

Lahousse, K. (2003). La complexité de la notion de topique et l'inversion du sujet nominal. In: *Adverbiaux et topiques. Travaux de linguistique* 47, 111-136 (M. Charolles and S. Prévost, eds.). Duculot, Brussels.

Partee, B. H. (1996). Allegation and local accommodation. In: *Discourse and Meaning* (B. H. Partee and P. Sgall, eds.) , pp. 65-86. Benjamins, Amsterdam/Philadelphia.

Partee, B. H. (1998). Focus, quantification, and semantics-pragmatics issues. In: *Focus. Linguistic, Cognitive and Computational Perspectives* (P. Bosch and R. van der Sandt, eds.), pp. 213-231. Cambridge University Press.

Petkevič, V. (1995). A new formal specification of underlying structures. *Theoretical Linguistics* 21, 7-61.

Sgall, P., O. Pfeiffer, W. U. Dressler and M. Půček (1995). Experimental research on Systemic Ordering. *Theoretical Linguistics* **21**, 197-239.

Steube, A. (2001). Grammatik und Pragmatik von Hutkonturen. *Linguistische Arbeitsberichte* 77, 7-30. English version: Bridge contours in German assertive main clauses. *Folia linguistica* 37, 2003, 163-190.

Veselá, K., N. Peterek and E. Hajičová (2003). Some observations on contrastive topic in Czech spontaneous speech. *Prague Bulletin of Mathematical Linguistics* **79-80**, 5-22.