MTT 2007, Klagenfurt, May 21 – 24, 2007
Wiener Slawistischer Almanach, Sonderband 69, 2007

The Position of TFA (Information Structure) in a Dependency Based Description of Language

Eva Hajičová

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University in Prague
hajicova@ufal.mff.cuni.cz

Abstract

Information structure of the sentence (its topic-focus articulation, TFA in the sequel), though disguised under different terms and slightly different interpretations, belongs nowadays among the most frequently discussed issues of the sentence structure and its semantic and communicative function (Section 2). Based on our long-term study of these issues influenced heavily by the Prague School approach to functional linguistics and, first of all, by Petr Sgall's pioneering insights into the semantic relevance of TFA, we present an answer to three basic questions: (1) on which layer of linguistic description the basic features of TFA are to be represented (Section 3), (2) which basic oppositions are to be captured (Section 4), and (3) which formal description of language structure is best suited for such a representation (Section 5). In passing, we compare our standpoints to those embodied in other models, paying a special attention to Igor Mel'čuk's Meaning Text Theory. In the concluding section (Section 6), we briefly sketch how a carefully-designed linguistically-based annotation of a corpus may serve as a testbed for a linguistic theory.

Keywords

dependency, information structure, underlying level, Meaning Text Theory, Functional Generative Description, Prague Dependency Treebank, corpus annotation

1 Introduction

In the Introduction to his book on Communicative Organization in Natural Language (Mel´čuk 2001, CONL in the sequel), Igor Mel´čuk formulates six questions the answers to which substantiate why the book was written and what it is (and, deliberately, is not) about. The first of these questions concerns why „his hands are trembling" whenever he sits and works on the book. His uncertainty stems from the fact that „so many brilliant linguists have written and are still writing so many works on the extremely difficult subject" (CONL, p.1).

However, the autor still finds a good excuse for writing CONL: he considers the communicative organization of sentence „strictly from the perspective of constructing actual sentences from a representation of their meanings within a particular framework: the

Meaning-Text Theory“ (CONL, p.2). He notes that this viewpoint has not been explored before (at least not systematically enough).

I share the view that information structure is nowadays a hot topic in linguistics and that for the past forty years much has been published on this issue. Among those writings, many explore TFA from the point of view of the framework I subscribe to, namely the Functional Generative Description (FGD). Therefore, the reasons why I have chosen the information structure of the sentence as the central topic of my invited talk, are a bit different from those Igor Mel´čuk considers. One of the aims of the present paper is to offer starting points for a comparison of different theories.

As some of the fundamental points of the Meaning Text Theory (MTT) and FGD are close to each other (for a comparison, see e.g. Žabokrtský 2005), I was curious to see whether and in which points the two theories differ in the understanding and description of TFA. For this purpose, I will briefly summarize the arguments for our treatment of TFA as a semantically relevant phenomenon which is to be captured at the underlying (tectogrammatical, in our terms) level having the form of dependency tree structures and illustrate on a couple of hypotheses formulated within our theory how a carefully-designed linguistically-based annotation of a corpus may serve as a testbed for linguistic theories.

2 Some historical milestones

The writings on what we refer to as TFA and what is more generally (and recently) covered by the term information structure date back centuries ago; the issue is treated under different terms and there is not always possible to find a one-to-one mapping between them; also, they receive a slightly different interpretation. However, they share the underlying idea: a description of the structure reflecting the functioning of language in communication, which is different from the subject-verb-object structure (described in any formalism). One of the oldest and most stimulative, at least for its time, is Weil’s (1844) comparison of the means expressing information structure in languages of different types. Of great interest is his proposal to distinguish two types of ‘progressions’ of sentences in a discourse, in relation to which part of a given sentence serves as a starting point for the subsequent one. Sentences may follow each other in a parallel mode, i.e. they share their starting points (marche parallèle), or in a sequential mode, i.e. the starting point of a given sentence follows up the second (final) part of the preceding sentence (progression). In more modern terms, one can say that in the parallel mode, the sentences share their themes (topics), in the sequential mode the theme (topic) of one sentence relates to the rheme (focus) of the preceding sentence. (It should be noted that more than one hundred years later, a similar, though a more subtle approach was developed by Daneš 1970 in his paper on thematic progressions).

It is not our intention here to present a historical survey; let us only mention that though the first hints for a systematic treatment of these issues within structural linguistics were given by Prague scholars in the second quarter of the last century (initiated by Vilém Mathesius and later continued by Jan Firbas), one should not forget that the topic was, so to say, hanging in the air, receiving the attention esp. in German linguistics (for a more detailed discussion, see Sgall et al., 1973, 1980 and 1986).

With the entrance of formal linguistics on the scene, it is not surprising that the first suggestions for an inclusion of TFA into an integrated formal description of language came from Prague; Sgall's Functional Generative Description (Sgall 1964; 1967a) working with a tectogrammatical (underlying, deep) level of sentence structure incorporated the TFA opposition into the description of this level (Sgall 1967b).

It should be noted that the examples serving as arguments during the split of generative transformational grammar into interpretative and generative semantics reflected the difference in TFA (actually, on both sides of the dispute, though not recognized as such; see e.g. Chomsky 1971 and Lakoff 1971a, to name just the main figures). A "breakthrough" on that side of Atlantic was Mats Rooth's doctoral dissertation on association with focus (Rooth 1985), in which the author (referring i.a. to Jackendoff 1972) quite convincingly argues for the "semantic effect of focus" in the sentence offering the explanation of this effect in terms of a domain of quantification (p. 197); his starting arguments were restricted to the presence in the sentence of the so-called focusing particles such as only, even, but he extended his proposal also to the so-called adverbs of quantification (often, always) and cases such as cleft constructions in English.

The interest was aroused, and after Barbara Partee's (who was one of Mats Rooth's supervisors) involvement in the discussion of the semantic consequences of different TFA structures (see e.g. Partee 1991) the TFA issues took up an important position in the discussions of formal semanticists (for a Czech contribution to that discussion see Peregrin 1994; 1996), but not only within that domain (quite noticeable is the interest in the TFA issues in German linguistics).

One of the crucial contributions of the above mentioned discussions was the due respect to the reflection of the differences in TFA in the prosodic shape of the sentences (which view, actually, has been present in the Praguian studies of TFA). Let us mention here only Jackendoff's (1972) introduction of the difference in A and B prosodic contour and Rooth's (1985) consistent regard to the placement of the intonation pitch in his example sentences.

3 Where to represent TFA

3.1 Semantic relevance of TFA

To give an answer to the question posed in the title of this section, let us start with some examples (maybe notoriously known). The capitals denote the intonation centre, the names in brackets indicate the source of the examples.

(1)(a) Everybody in this room knows at least two LANGUAGES.
(b) At least two languages are known by everybody in this ROOM.
(Chomsky 1957; 1965)

(2)(a) Many men read few BOOKS.
(b) Few books are read by many MEN. (Lakoff 1971a)

(3)(a) Londoners are mostly at BRIGHTON.
(b) At Brighton, there are mostly LONDONERS. (Sgall 1967b)

(4)(a) I only introduced BILL to Sue.
(b) I only introduced Bill to SUE. (Rooth 1985)

(5)(a) I work on my dissertation on SUNDAYS.
(b) On Sundays, I work on my DISSERTATION.

(6)(a) English is spoken in the SHETLANDS.
(b) In the Shetlands, ENGLISH is spoken. (Sgall et al. 1986)

(7)(a) Dogs must be CARRIED.
(b) DOGS must be carried. (Halliday 1967)
(c) Carry DOGS. (a warning in London underground, around 2000)
(d) CARRY dogs.

It is not difficult to understand that the pairs of sentences under each number differ not only in their outer shapes or in their contextual appropriateness, but also in their meanings, even in their truth conditions. This difference may be attributed to the presence of quantifiers and their order (with an explicit quantification in (1) and (2) and a more or less explicit in (3) and (4)), but from (5) on, such an explanation is not possible. Also, an exclusive reference to the surface order of the sentence elements would not be correct, as illustrated by (4) and (7).

A more adequate explanation is that based on the relation of *aboutness*: the speaker communicates something (the Focus of the sentence) about something (the Topic of the sentence), i.e. F(T), the Focus holds about the Topic. In case of negative sentences, the Focus does not hold about the Topic: ~F(T).

A supportive argument for the semantic relevance of TFA can be traced in the discussions on the kinds of entailments starting with the fundamental contributions of Strawson. Strawson (1952, esp. p. 173ff.) distinguishes a formal logical relation of entailment and a formal logical relation of presupposition; this distinction – with certain simplifications - can be illustrated by (8) and (9):

(8) All Johns' children are asleep.
(9) John has children.

If John's children were not asleep, the sentence (8) would be false; however, if John did not have children, the sentence as well as its negation would not be false but meaningless. Thus (9) is a presupposition of (8) and as such it is not touched by the negation of (8).

Returning to the relation of aboutness, we can say that (8) is about John's children, and for (8) to be meaningful, there must be an entity John's children the speaker can refer to.[1]

The close connection between the notion of presupposition and TFA can be documented by a more detailed inspection of the notion of presupposition, exemplified here by sentences (10) and (11).

---

[1] This need not mean that the entity the sentence is 'about' should exist in the real world, but it should be referentially available; cf. the discussion of the notion of referential vs. existential presuppositions in Hajičová 1976, 55-58, reflected also in Sgall et al. 1986).

(10) The King of France is (not) bald.
(11) The exhibition was (not) visited by the King of France.

It follows from the above mentioned discussions on presuppositions that Strawson's (1964) ex. (10) is about the King of France and the King's existence (referential availability) is presupposed, it is entailed also by its negative counterpart; otherwise (10) would have no truth value, it would be meaningless. On the other hand, there is no such presupposition for (11): the affirmative sentence is true if the King of France was among the visitors of the exhibition, while its negative counterpart is true if the King of France was not among the visitors. The truth/falsity of (11) does not depend on the referential availability of the entity "King of France". This specific kind of entailment was introduced in Hajičová (1972) and was 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 Hajičová 1984; 1993, and the discussion by Partee 1996). Concerning the use of a definite noun group in English one can say that it often triggers a presupposition if it occurs in Topic (see sentence (10)), but only an allegation if it belongs to Focus (see sentence (11)).

These considerations have led us to the attempt at a more systematic analysis of the relations between affirmative and negative sentences (Hajičová 1972, 1984, 1993). The scope of negation can be specified, in the prototypical case, as constituted by the Focus, so that the meaning of a negative declarative sentence can be interpreted as its Focus (F) not holding of it, i.e. ~F(T). In this way it is possible to understand the semantic difference present in (10) and (11).

In a secondary case, the assertion holds about a negative Topic: F(~T), see (12) on the reading when answering the question "Why didn't he come?".

(12) He did not come because he was afraid.

Here again, the scope of negation is dependent on TFA: it is restricted to the Topic part of the sentence. The assertion entailed (on this reading) by the *because*-clause in Focus is not touched by negation.[2]

3.2 Which layer of linguistic description does TFA belong to?

The analysis summarized in Sect. 3.1. points out very clearly that TFA undoubtedly is a semantically relevant aspect of the sentence and as such should be represented at a level of an integrated language description capturing the meaning of the sentence (whatever interpretation we assign to the notion of 'meaning'). For the formal description of language we subscribe to, namely the Functional Generative Description, this is the underlying, *tectogrammatical* layer; the tectogrammatical representations of sentences (TRs) are specified as dependency tree structures, with the verb (of the main clause) as the root of the tree. While the labels of the nodes of the tree are counterparts to the autosemantic words of the sentence,

---

[2] On another possible reading of (12), e.g. if the sentence is followed by *but because he was on his leave of absence,* his being afraid is neither entailed nor negated, i.e. the assertion belongs to the allegations of the sentence, i.e. he might have come for some other reason. The negation concerns Focus, schematically: ~F(T).

counterparts of function words as well as of grammatical morphemes are just indices of the nodes and the edges of the tree: the morphological values of number, tense, modalities, and so on, are specified by indices of the labels of the nodes. For each node of the TR it is specified whether it is contextually bound or non-bound (see Sect. 4.1 below). The edges of the tree are labeled by underlying syntactic relations (such as Actor/Bearer, Addressee, Patient, Origin, Effect, several Local and Temporal relations, etc.). In the corresponding surface shapes of the sentences, there are several means (the extent of their exploitation is different in different languages) rendering the TFA distinctions: the word order, the placement of the intonation center and the intonation contours, specific syntactic constructions (such as clefting in English), or specific morphemic means (such as the particle *wa* in Japanese).

A similar question can be posed as for the MTT approach: as a multilayered description of language, this theory postulates a series of layers of representation, the "highest" (or deepest) of which, *SemR*, is declared to be the meaning, specifying the meaning of a set of synonymous sentences. The concept of meaning is thus based on the concept of same meaning; however, as Mel'čuk (1988, pp. 14, 25ff, 52) claims, a certain degree of approximation in the semantic representation is necessary, if linguistically interesting results are to be obtained. Mel'čuk (2001, pp. 25ff.) develops this idea further: Two utterances are called 'more or less synonymous' if they 'mean roughly the same', i.e. "they must express (almost) the same set of semantic units organized in the same configurations" (CONL, p.14). This relation is distinguished from the notion of strict (full) synonymy: "two utterances are called strictly (fully) synonymous if a native speaker cannot find any semantic distinction whatsoever between them, including what can be referred to as 'communicative nuances'" (ibid).[3]

SemR has as its central component SemS, i.e. a representation of propositional, or situational meaning of a family of more or less synonymous sentences, the other three components being Semantic Communicative Structure (SemCommS), the Rhetorical Structure (RhetS), and the Referential Structure (RefS) (Mel'čuk 2001, p.4).

The particularity of the theory, however, is the postulation of an intermediate deep-syntactic level, DSyntR, between the semantic (SemR) and the surface syntactic representation (SSyntR). Thus the module called semantics comprises two layers or representation, SemR and DSyntR. The latter layer is aimed at representing the lexico-grammatical organization of a concrete sentence; its central component is the deep syntactic structure, the other three components being DSynt-Communicative Structure (different from the SemCommS, see below), the DSynt-Anaphoric Structure and DSynt-Prosodic Structure. Formally, the DSyntS is an unordered dependency tree the nodes of which are labeled with full lexical units of the language and the arcs are labeled with symbols for deep syntactic relations (such as I, II

[3] Kahane (2003) summarizes this idea as follows: "During the process of sentence construction … lexical and syntactic choices carried out by the Speaker very often lead to the modification of … the initial semantic representation, making it more precise and specific. … The initial semantic representation is taken to be rather approximate … The meaning can become more precise – or less precise…". In FGD, we distinguish between strict synonymy and quasi-synonymy; with the latter, the difference in truth-conditions is relevant only in specific contexts and citations.

through VI and attributive relation ATTR; coordination is also understood as one of the dependency relations).[4]

It is not the objective of the present paper to discuss the overall organization of the MTT model. Therefore, we restrict ourselves to saying that we do not consider a two-layer semantic module the MTT approach postulates to be necessary unless one of them is proclaimed as a layer of language system and the other as a (partial) reflection of the content of the sentence. We consider the main boundary line to separate linguistic meaning from the basically extra-linguistic content. Since the SemR of MTT is described as to include also the representation of the communicative (information) structure, we constrain ourselves to the issues relevant for this representation.

4 Which basic oppositions are to be captured?

4.1 Contextual boundness as the basic opposition

In the approach of the Functional Generative Description to TFA, the underlying structure is represented as containing the attribute of contextual boundness: for every autosemantic lexical item in the sentence (i.e. for every node of its tectogrammatical representation) it is specified whether it is (a) contextually bound (*cb*), 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, or (b) contextually non-bound (*nb*), 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 by (13): both Tom and his friends are 'given' by the preceding context (indicated here by the preceding sentence in the brackets), but they are structured as non-bound (which is reflected in the surface shape of the sentence by the position of the intonation center).

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

In the prototypical case, the head verb of the sentence and its immediate dependents (arguments and adjuncts) constitute the Topic of the sentence if they are contextually bound, whereas the Focus consists of the contextually non-bound items in such structural positions (and of the items syntactically subordinated to them). Also the semantically relevant scopes of focus sensitive operators such as *only, even*, etc. can be characterized in this way.

The bipartition of the sentence into the Topic and Focus (reflecting the aboutness relation) can then be specified by the following set of the rules determining the appurtenance of a

---

[4] In FGD, coordination is not considered to be a dependency relation; rather, it is understood as a third dimension of the structure, making the resulting representation of the sentence a network of three orderings, rather than a two-dimensional tree.

lexical occurrence to the Topic (T) or to the Focus (F) of the sentence (see Sgall 1979; Sgall et al. 1986, pp. 216ff)

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

(b) every item *i* 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 k_j directly depending on it carry index *cb*, then those items k_j to which some items l_m carrying *f* are subordinated are called 'proxy foci' and the items l_m together with all items subordinated to one of them belong to F, where $1 \leq j,m$;

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

There are two reasons why to distinguish the opposition of contextual boundness as a primary (primitive) one and to derive the Topic-Focus bipartition from it: (i) the Topic/Focus distinction exhibits – from a certain viewpoint - some recursive properties which are best described by the *cb/nb* distinction, and (ii) Topic/Focus bipartition cannot be drawn on the basis of an articulation of the sentence into constituents but requires a more subtle treatment (for arguments, see Sect. 5 below).[5]

The semantico-pragmatic interpretation of sentences (for which the TRs represent suitable input) may then include an application of Tripartite Structures (Operator - Restrictor - Nuclear Scope), as outlined by B. H. Partee in Hajičová et al. (1998). Let us briefly recall some of the characteristic sentences discussed there (with their relevant TRs) and specify (in a maximally simplified notation) which parts of their individual readings belong to the Operator (O), Restrictor (R) and Nuclear Scope (N) of the corresponding tripartite structures. We assume that in the interpretation of a declarative sentence, O corresponds to negation or to its positive counterpart (the assertive modality) or to some other operators such as focusing particles, R corresponds to Topic (T), and N to Focus (F).

(14) *John only sits by the TELEVISION.*
(14') O *only,* R *John,* N *sits by the TELEVISION.*
(14") O *only,* R *John sits,* N *by the TELEVISION.*

In (14), the particle occupies its prototypical position in the TR, so that the focus of the particle is identical with the F of the sentence on either reading, i.e. with the verb included in F in (14'), and in T in (14"). If a focusing particle occupies a secondary position, rather than that of the main operator, we use the ASSERT operator (introduced by Jacobs 1984) in the interpretation. If the operator is included in T, its own focus (which differs from the sentence

[5] It is a matter of course that the distinctions made have to be checked by means of some operational tests or criteria. Several such test have already been discussed in literature the remarkable point being that they all more or less return very similar results. Not going into detail on this point, we just refer to the question test (Sgall et al. 1986), the test by means of an acceptable negative continuation (response) (Chomsky's 1971 'natural response', Posner's 1972 'Kommentieren' test) and tests based on the scope of negation or other focus sensitive operators (e.g. Hajičová, Partee and Sgall 1998).

F in such marked cases) does not cross the boundary between the T and the F of the sentences, see (15):

(15) (What did even PAUL realize?) Even Paul realized that Jim only admired MARY.
(15') O ASSERT, R *(*O *even,* R *realized,* N *Paul),* N *(*O *only,* R *Jim admired,* N *Mary)*

Let us just note that in the cases in which T or F is complex, as illustrated by (15), it is the opposition of contextual boundness that is responsible for the difference: while contextually bound items then belong to the local (partial) R, the non-bound ones belong to the corresponding N.

4.2 Communicative oppositions in SemComm of MTT

In MTT, the main structure of semantic representation, the semantic structure (SemS) is supposed to be formally a connected directed graph, with nodes (semantemes) corresponding to disambiguated lexicographic word senses of the language in question. The arcs are labeled with numbers specifying predicate-argument relations.

The communicative organization of sentences is represented on the level of meaning by the Semantic-Communicative Structure (SemComm).[6]

This structure is supposed to organize the meaning of the sentence into a message "from the viewpoint of its transmission by the Speaker and its reception by the Addressee" (CONL, p. 23); this is done by (i) specification of the division of the SemS of the sentence into (possibly overlapping) parts, called communicative areas, and by (ii) specification of a dominant node of each communicative area marked by one of a set of mutually exclusive values (one of the communicative oppositions, Comm-oppositions). Eight Comm-oppositions are distinguished (CONL, p. 49, with a more detailed specification on pp. 74ff): (i) thematicity (rheme vs. theme vs. specifier), (ii) givenness (given vs. new vs. irrelevant), (iii) focalization (focalized vs. neutral), (iv) perspective (foregrounded vs. backgrounded vs. neutral), (v) emphasis (emphasized – neutral), (vi) presupposedness (presupposed vs. non-presupposed (= asserted or neither), (vii) unitariness (unitary vs. articulated), (viii) locutionality (signaled vs. performed vs. communicated). These SemComm categories can combine with each other within one SemS and may be obligatory ((i) and (ii)) or optional (the rest). In addition, they are assumed to form a hierarchy (indicated in the list by the order in which they are listed); Mel'čuk (CONL, p. 74) admits that the hierarchy is not strict and that he is unable to sufficiently motivate the order of all the oppositions, which therefore is more or less arbitrary.

On the level of DSyntR, the communicative structure of the sentence (its DSynt-Communicative Structure) is to a great extent determined by its Sem-CommS (pp. 64ff) though due to lexical or grammatical constraints of the language involved. it can require

---

[6] It should be appreciated that Mel'čuk emphasizes the involvement of meaning in his approach; however, he is not precise when he ascribes (CONL, p. 22) to Lambrecht's book (1994) an exceptional status in drawing a systematic distinction between semantico-pragmatic (in MTT terms: communicative) categories and their formal expression in sentences, see our discussion and references in Sect. 2 above.

restructuring of the Sem-Comm-data: this is illustrated by a comparison of the Russian sentence *Iz-za ego ot'ezda my polnost'ju pomenjali nashi plany* (literal E. translation: *Because of his departure, we completely changed our plans*). with its assumed analysis on the semantic level and the deep-syntactic level. While on the semantic level, Sem-theme of the SemR is 'he left' and the 'cause' semanteme is a part of the semantic rheme, there is no easy way in Russian to express causation verbally as a separate lexeme; therefore the preposition *iz-za* is to be used, which, on the deep syntactic level, is supposed to be a part of DSynt-Theme. This argumentation necessarily leads to a question how deep the deep syntactic representation is if it depends so much on the means available in a given language. This point, however, is not crucial for our present discussion since the representation of the fundamental aspects of TFA (or communicative structure, in MTT terms) are present at the SemR in any case.

What is more pertinent to our discussion where TFA or communicative structure are to be represented, are issues related to the scopes of quantifiers and negation. As Kahane (2003) notes, the scope of quantifiers is not directly encoded in standard MTT semantic representations; Mel'čuk (CONL, pp. 358-360) in his discussion on the scope of *only* (the ambiguity of the written sentence *Word-final position only admits voiceless consonants*) considers that a specification of this must be indicated in both the SemS and the SemCommS. In SemS, this is to be done by the choice of arguments of the semanteme *only* (*only voiceless* vs. *only consonants*) and the consequent different assignment of 'rhematic focus' (*voiceless* vs. *voiceless consonants*). As for DSyntR, it is the DSynt-Prosodic Structure which marks the lexical expression as for its accent (*voiceless* vs. *consonants*).[7]

4.3 Contextual Boundness as a basis for more subtle distinctions

At the first sight, it may seem that the MTT account offers more possibilities to cover a larger display of subtle distinctions than the FGD approach. The question thus arises whether the effort to propose as economic description of the relevant phenomena as possible does not cost too much in terms of adequacy and coverage.

We claim that this is not the case with FGD. From the eight oppositions listed above, (ii), (vii) and (viii) lie beyond the linguistic structuring of meaning and pertain to the cognitive layer, rather than to the structure of sentences. The opposition of givenness (ii) can be understood as a cognitive background of the opposition of cb/nb. It is not clear whether the opposition of emphasis ((v) in the list) is semantically relevant and should be as such captured by the representation of (linguistic) meaning. The same holds of the opposition of perspective (listed as (iv) and obtaining the values of foregrounded vs. backgrounded vs. neutral): it is characterized as that piece of information that has, from the viewpoint of the speaker, a special or reduced psychological prominence, or is in the central or peripheral part of the situation. Contrary to Mel'čuk (CONL, p. 202), we take it as a relevant feature where the

---

[7] Mel'čuk does not mention other possibilities of the scope of *only* in his example: if the context after the given sentence is "… It does not indicate all positions relevant for the given structure" the theme would include only *word-final position* and the rest of the sentence would be the rheme. It would be interesting to see how this interpretation would be captured in the Sem-CommS differently from the case when the scope of *only* includes just *voiceless consonants*.

'peripherality' comes from: it is important to examine whether it is based on some internal semantic considerations, or from discourse organization, or from another source; only in the first case, it should be considered as relevant for the representation of meaning, and accounted for as such.

We thus remain with oppositions (i), (iii), and (vi). The opposition (i) directly relates to the bipartition of the sentence into Topic and Focus derived on the basis of *cb*/*nb* nodes of the representation (see Sgall's rules given above); in our discussion of presupposition and allegation in Sect. 3.1 we have also indicated how the opposition of 'presuposedness' ((vi) in the list) relates to TFA. Focalization ((iii) in the list) covers, in our opinion, several phenomena: it maybe be characterized as a **'**narrow focus**':** usually understood as a Focus consisting in a single node (plus attribute), which may be exemplified by a case of the prototypical position of a focus sensitive particle. Alternatively, it relates to a contrastive node in the Topic part of the sentence (a focusing particle can occur also in the Topic part and it may indicate a 'local focus', see the analysis of (15) above).

The recognition of a primary distinction of contextual boundness, on the other hand, offers a possibility to represent other distinctions related to TFA and referred to in present-day linguistic writings. First of all, it allows for an account of the possible recursivity of TFA, exemplified first of all in sentences which contain embedded (dependent) clauses. The dependent clause D functions as a sentence part of the clause containing the word on which D depends, so that the whole structure has a recursive character; one of the questions discussed is whether the T-F articulation should be understood as recursive, too. Several situations obtain: (i) one of the clauses may be understood as the F of the whole sentence, though each of the clauses displays a T-F articulation of its own ("local" topics and foci: *(the) market*, *unused possibilities*, and *saturated*, *transmission*, respectively), as in (16); (ii) in a general case the boundary between T and F may lie within one of the clauses (as in (17)). (The examples marked by PDT are taken from the Prague Dependency Treebank, see Sect. 6 below.)

(16) Zatímco trh s rozhlasovým signálem už je nasycen, nevyužité možnosti stále má televize zejména při regionálním a lokálním vysílání. (PDT)
Lit. While (the) market with radio signal already is saturated, unused possibilities still has television-Nom. especially with regional and local transmisson.

(17) Na základě návrhu zákona, který vláda projednala na své poslední schůzi, se budou daně snižovat až o 10 procent. (PDT)
Lit. On the basis of the propsal of the law, which the government has approved on its latest meeting, the taxes will be lowered up to ten percent.

Also the notion of *verum* focus can be captured by means of *cb*/*nb* distinction, with the yes-no modality constituting the whole Focus (see (18)).

(18) (Have you attended the last meeting?) No, I did not attend the last meeting.

The approach based on the *cb*/*nb* opposition makes it possible to account for the TFA of coordinated structures, in which each of the clauses connected in a coordinative construction, i.e. in a compound sentence, exhibits a Topic/Focus articulation of its own. In (19), the Topic of the first clause is *je bez něho* (it is without him), the Focus being *nemyslitelná* (unthinkable), while in the second clause, the Topic is *zároveň má samozřejmě v sobě* (it has

at the same time in itself) and the Focus is *spoustu nejrozmanitějších nebezpečí* (many multifarious dangers).

(19) Je bez něho nemyslitelná, ale zároveň má samozřejmě v sobě spoustu nejrozmanitějších nebezpečí. (PDT)
Lit.: it-is without him unthinkable but at-the-same-time it-has in itself many multifarious dangers.

An interesting issue is that of contrast in Focus. There are different kinds of contrast: a ‚neutral' one as in (20) or (21), a correction as in (22), or the case called by V. Mathesius the second instance (as in (23)).[8]

(20) (Preceding context: Kde se mluví česky?) Česky se mluví v ČESKU.
(Where is Czech spoken?) Czech is spoken in CZECHIA.

(21) (Preceding context: Mluví se česky v Česku nebo na Slovensku?)
Česky se mluví v ČESKU.
(Is Czech spoken in Czechia or in Slovakia?) Czech is spoken in CZECHIA.

(22) (Preceding context: Mluví se česky ve Slovinsku nebo na Slovensku?)
Česky se mluví v ČESKU.
(Is Czech spoken in Slovenia or in Slovakia?) Czech is spoken in CZECHIA.

(23) (Preceding context: Na Slovensku se mluví česky.)
(In Slovakia one speaks Czech)

(a) (Ne.) Česky se mluví v ČESKU
(No.) Czech is spoken in CZECHIA.
(b) (Ne.) Na Slovensku se mluví SLOVENSKY.
(No.) In Slovakia one speaks SLOVAK.
(c) V ČESKU se mluví česky.
In CZECHIA one speaks Czech.

It can be said (in agreement with Rooth 1985) that by default, every Focus involves a choice of alternatives, so that the 'contrast' present in (20) or (21) is just what underlies Focus. However, even in the topic part of the sentence one may see a choice of alternatives if some element of Topic is contrastive (and marked as such, cf. Hajičová, Partee and Sgall, 1998). A typical example, known from Lakoff (1971b) and the older discussions (with *he* bearing a rising contrastive stress in both examples; *her* is stressed with the typical sentence final falling pitch contour in (a); in (b) the intonation centre is on *insulted*) is in (24) and (24'):

(24) (She called him a Republican.) *Then he insulted. HER.*

(24') (She called him a Republican.) *Then he INSULTED her.*

---

[8] Since the same sentence (with a narrow focus) can be used in any of the contexts of (19) - (22), we assign this sentence the same TR in the different uses.

5 Which type of formal description

In early discussions on the integration of the topic-focus articulation into a formal description of grammar, the proponents intended to specify this aspect of the structure of the sentence in terms of the type of formal description they subscribed to. Within the framework of generative transformational grammar, Chomsky (1971, p. 205) defined focus as "a phrase containing the intonation center", i.e. in terms of constituency (phrase-structure) based description (see also Jackendoff 1972, p. 237). Such a description served as a basis also for several studies on the relationship between syntax and semantics (e.g. Schmerling 1976; Selkirk 1984; 1985): the boundaries between topic and focus or some more subtle divisions were always supposed to coincide with the boundaries of phrases. Sgall and his followers (see already Sgall 1967b) work within a framework of dependency grammar define the boundary between the two parts on the basis of syntactic dependency, of the opposition of contextual boundness and of the left-to-right order of nodes. The boundary between Topic and Focus can then be characterized as intersecting an edge between a governor and its dependent (the latter may be a single node or a subtree), with the provision that whatever is to the right of the given dependent in the tectogrammatical dependency tree, belongs to the Focus, the rest to the Topic (see Sgall's definition above in Sect. 4.1). A similar strategy can be traced in MTT: formally, the CommS is encoded by spotting out some areas of the semantic (dependency) graph and labeling each of them with a communicative marker. A dominant node of each area (of theme or rheme), which is characterized as the node that summarizes the semantic content of the area, is marked by underlining (Polguère 1990).

However, the definition of Focus (and of presupposition, in Chomskyan terms) as a phrase is untenable since it is not always possible to assign the focus value to a part of the sentence that constitutes a phrase. This claim is supported by examples as those adduced by Hajičová and Sgall (1975): in the given context, the Focus of the sentence is *for a week to Sicily*, which would hardly be specified as a constituent under the standard understanding of this notion. These examples, however, bring no difficulties for a dependency-based description.

(25) John went for a week to Sicily. (He didn't go only for a weekend to his parents.)

It was convincingly argued by Steedman (1991; 1996; 2000) that it is advisable to postulate a common structure for accounting both for the syntactic structure of the sentence as well as for its information structure. For that purpose, he proposes a modification of categorial grammar, the so-called combinatory categorial grammar. A syntactic description of a sentence ambiguous as for its information structure should be flexible enough to make it possible to draw the division line between Topic and Focus also in other places that those delimiting phrases; in Steedman (1996, p.5), the author claims that his "theory works by treating strings like *Chapman says he will give*, *give a policeman*, and *a policemen a flower* as grammatical constituents" and thus defining "a constituent" in a way that is different from the "conventional linguistic wisdom". In other words, Steedman proposes to work with non-standard constituents, as can be illustrated by (26) with the assumed intonation center at the last element of the sentence: the division of (26) into Topic and Focus is ambiguous because the verb may belong either to the topic or to the focus part of the sentence.

(26) Fred ate the BEANS.

The representation of such an ambiguity in a dependency framework like that of the Praguian Functional Generative Description causes no difficulty. In case the root of the tree (the verb) is *cb*, then it depends on the *cb*/*nb* feature of its dependents whether *Fred ate* or just *ate* are the elements of the Topic (answering the question *What did Fred eat*?, or *Who did eat what*?, respectively. If the verb is *nb*, then again two divisions are possible: either the whole sentence is the Focus (*What happened*?), or the verb and the object are the elements of the Focus (*What did Fred do*?). In the underlying tree structure, the *cb* nodes depend on the verb from the left, the *nb* nodes from the right. A division line between Topic and Focus is then drawn as characterized above.

In (26), we assumed the (normal) placement of the intonation center on the object *beans*. However, as also discussed by Steedman, the sentence may have different intonation patterns, and this may reduce its ambiguity: if the intonation center is on *Fred*, then *Fred* is the sentence Focus and the rest is the Topic (*Who ate the beans*?). If the intonation center is on the verb, then only the verb is the Focus the rest being the Topic (*What did Fred do with the beans*?). This again can be easily captured in the dependency representation of the meaning of the sentence by the assignment of the primary opposition of *cb*/*nb* nodes.

6 Annotated corpora as testbeds for linguistic theory

Any modern linguistic theory has to be formulated in a way that it can be tested by some objective means; one of the ways how to test a theory is to use it as a basis for a consistent annotation of large language resources, i.e. of text corpora. Annotation may concern not only the surface and morphemic shape of sentences, but also (and first of all) the underlying sentence structure, which elucidates phenomena hidden on the surface although unavoidable for the representation of the meaning and functioning of the sentence, for modeling its comprehension and for studying its semantico-pragmatic interpretation.

Without going into any detail, we illustrate here on the example of the Prague Dependency Treebank (PDT, see e.g. Hajič 1998), based on the framework of the Functional Generative Description (FGD), how such a testing may be done.

PDT is an annotated collection of Czech texts, randomly chosen from the Czech National Corpus (CNK), with a mark-up on three layers: (a) morphemic, (b) surface shape "analytical", and (c) underlying (tectogrammatical). The current version (publicly available on address http://ufal.mff.cuni.cz/pdt2.0), annotated on all three layers, contains 3168 documents (text segments mainly from journalistic style) comprising 49442 sentences and 33357 occurrences of word forms (including punctuation marks).

On the tectogrammatical level, which is our main concern in the present paper, every node of the tectogrammatical representation (TGTS, a dependency tree) is assigned a label consisting of: the *lexical value* of the word, of its *'(morphological) grammatemes'* (i.e. the values of morphological categories), of its *'functors'* (with a more subtle differentiation of syntactic relations by means of *'syntactic grammatemes'* (e.g. 'in', 'at', 'on', 'under'), and the TFA attribute of containing values for *contextual boundness*. In addition, some basic intersentential links 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; TGTSs differ from the theoretically adequate TRs in that coordinating conjunctions are represented as head nodes of

the coordinated structures, which makes it possible for the TGTSs to constitute two-dimensional trees.

Every node of the TGTS is assigned one of the three values of the attribute specifying TFA: *t* for a contextually bound non-contrastive node, *c* for a contextually bound contrastive node, *f* for a contextually non-bound node.

A very simplified (preferred, given by the context) TGTS of the sentence (27) is in Fig 1.

(27) Nenadálou finanční krizi podnikatelka řešila jiným způsobem. (PDT)
Lit.: (The) sudden financial crisis-Accus. (the) entrepreneur-Nom. solved by other means. (context: the enterpreneur had to solve several problems before)

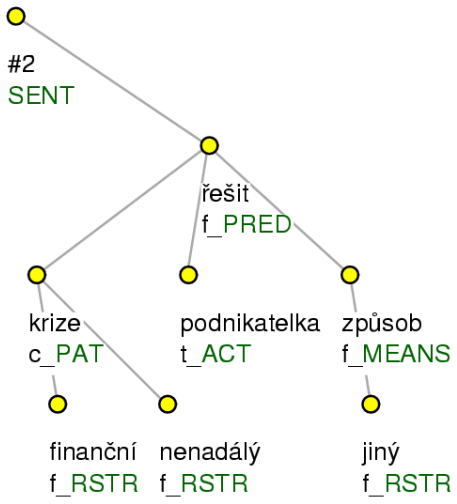

Figure 1: The preferred TGTS of sentence (27).

One of the hypotheses of the TFA account in TFG was the following:

Hypothesis: The division of the sentence into its T and F can be derived from the contextual boundness of the individual lexical items contained in the sentence (see above, Sect. 4.1).

An application of the rules in Sect. 4.1 gives the following result:

Topic: *Nenadálou finanční krizi podnikatelka* [the sudden financial crisis the enterpreneur]
Focus: *řešila jiným způsobem* [solved by other means]

The implementation of the algorithm has led to a differentiation of five basic types of Focus and it significantly supported the hypothesis that in Czech the boundary between T and F is signalized by the position of the verb in the prototypical case (the boundary between T and F: immediately before the verb in 95 % of the cases) and it has also been confirmed that the TFA annotation leads to satisfactory results even with rather complicated "real" sentences in the corpus.

Another hypothesis that has already been tested on our annotated corpus concerns the order of elements in the Focus. It is assumed that in the focus part of the sentence the complementations of the verb (be they arguments or adjuncts) follow a certain canonical

order in the TRs, the so-called systemic ordering (not necessarily the same for all languages). In Czech, also the surface word order in Focus corresponds to the systemic ordering in the prototypical case.

For Czech, the following systemic ordering is postulated (see Sgall et al. 1986): Actor – Time:*since-when* – Time:*when* – Time:*how-long* – Time:*till-when* – Cause – Respect – Aim – Manner – Place – Means – Dir:*from-where* – Dir:*through-where* – Addressee – Origin – Patient – Dir:*to-where* – Effect.

Systemic ordering as a phenomenon is supposed to be universal; however, languages may differ in some specific points: the validity of the hypothesis has been tested with a series of psycholinguistic experiments (with speakers of Czech, German and English); for English most of the adjuncts follow Addressee and Patient (Sgall et al. 1995). However, PDT offers a richer and more consistent material; preliminary results have already been achieved based on (a) the specification of F according to the rules mentioned above, (b) the assumed order according to the scale of systemic ordering (functors in TGTS), and (c) the surface word order (Zikánová 2006).

7 Conclusion

A deeper empirical analysis of sentences (in their context) in various languages convincingly shows that the issues referred to as belonging to TFA (or communicative structure, information structure, theme-rheme or whatever terms are used) are semantically relevant. Therefore, their description should be integrated into the description of the underlying, deep syntactic (tectogrammatical) level of linguistic description. It is this level that is suitable as the input to semantico-pragmatic interpretation. The phenomena connected with TFA on other levels (word order, or also particles, clefting etc.) serve as means expressing TFA.

We have argued that the primary opposition to be distinguished is the opposition of contextual boundness, from which the bipartition of the sentence into its Topic and Focus and other related notions can be derived. Such a description offers an adequate, effective and economic way of capturing the corresponding semantically relevant distinctions. A well-suited way of testing the theoretical assumptions and hypotheses is the present-day availability of corpora annotated in a systematic and linguistically-based manner, as the experience of the Prague Dependency Treebank indicates.

Acknowledgements

The author gratefully acknowledges the most useful comments and suggestions given by Jarmila Panevová and Petr Sgall after having read the pre-final version of the manuscript. The present paper has been written under the support of the grant MSM0021620838.

Bibliography

Bosch, P. & R. van der Sandt. (eds). 1994. *Focus and Natural Language Processing. IBM Working Paper 7*, Heidelberg: IBM Deutschland.

Chomsky, N. 1957. *Syntactic Structures*, The Hague: Mouton.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, Mass.: The MIT Press.

Chomsky, N. 1971. *Deep Structure, Surface Structure and Semantic Interpretation*. In Steinberg and Jakobovits, 193–216.

Daneš, F. 1970. Zur linguistischen Analyse der Textstruktur. In *Folia linguistica,* 4:72–78.

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

Hajičová, E. 1984. Presupposition and Allegation Revisited. In *Journal of Pragmatics,* 8:155–167. Amplified as Sgall, P. (ed). 1984. On presupposition and allegation. In *Contributions to Functional Syntax, Semantics and Language Comprehension,* Amsterdam: Benjamins, Prague: Academia, 99–122.

Hajičová, E. 1993. Issues of Sentence Structure and Discourse Patterns. Prague: Charles University.

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

Hajičová, E. & P. Sgall. 1975. Topic and Focus in Transformational Grammar. In *Papers in Linguistics*, 8:3–58.

Hajičová, E. & P. Sgall. 2004. Degrees of Contrast and the Topic-Focus Articulation. In Steube, A. (ed). *Information Structure – Theoretical and Empirical Aspects.* Berlin and New York: Walter de Gruyter, 1–13.

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

Jackendoff, R. 1972. *Semantic Interpretation in Generative Grammar*, Cambridge, Mass.: MIT Press.

Jacobs, J. 1984. Funktionale Satzperspektive und Illokutionssemantik. In *Linguistische Berichte,* 91:25–58.

Kahane, S. 2003. The Meaning-Text Theory. In *Dependency and Valency Handbook of Linguistics and Communication Science,* 25:1–2, Berlin and New York: de Gruyter.

Katz, J. J. & P. M. Postal. 1964. *An Integrated Theory of Linguistic Descriptions,* Cambridge, Mass.: The MIT Press.

Lakoff, G. 1971a. On Generative Semantics. In Steinberg and Jakobovits, 232–296.

Lakoff, G. 1971b. Presupposition and Relative Well-Formedness. In Steinberg and Jakobovits 329–340.

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

Mathesius, V. 1939. O tak zvaném aktuálním členění větném. In *Slovo a slovesnost* 5:171–174; translated as Kuno, S. (ed). 1975. On information-bearing structure of the sentence. In *Harvard Studies in Syntax and Semantics*, 467–480.

Mel'čuk I.A. 1988. *Dependency Syntax: Theory and Practice*, State University of New York Press, New York.

Mel'čuk I.A. 2001. *Communicative Organization in Natural Language: The Semantic Communicative Structure of Sentences*. Amsterdam and Philadelphia: John Benjamins (referred to in the text as CONL).

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

Partee, B.H. 1996. Allegation and Local Accommodation. In Partee, B.H. & P. Sgall (eds). 1996, 65–86.

Partee, B.H. & P. Sgall (eds). 1996. In *Discourse and Meaning: Papers in Honor of Eva Hajičová,* Amsterdam and Philadelphia: Benjamins.

Peregrin, J. 1994. Topic-Focus Articulation as Generatlized Quantification. In Bosch & van der Sandt (eds). 379–388.

Peregrin, J. 1996. Topic and Focus in a Formal Framework. In Partee, B.H. & P. Sgall (eds). 1996, 235–254.

Polguère, A. 1990. Structuration et mise en jeu procédurale d'un modèle linguistique déclaratif dans un cadre de génération de texte. Thèse de l'Université de Montréal.

Posner, R. 1972. *Theorie des Kommentierens*, Frankfurt am M.

Rooth, M. 1985. *Association with Focus*. PhD Thesis, Univ. of Massachusetts, Amherst.

Selkirk, E. 1984. Phonology and Syntax: The Relation between Sound and Structure. Cambridge, Mass.: MIT Press.

Selkirk, E. 1995. Sentence Prosody: Intonation, Stress and Phrasing. In Goldsmith, J.A. (ed). *Handbook of Phonological Theory*, London: Blackwell, 550–569.

Sgall, P. 1964a. Zur Frage der Ebenen im Sprachsystem.In *Travaux linguistiques de Prague,* 1:95–106.

Sgall, P. 1964b. Generative Bschreibung und die Ebenen des Sprachsystems, presented at the Second International Symposium in Magdeburg, printed in *Zeichen und System der Sprache* III, 1966, Berlin, 225–239. Reprinted in Sgall, P. (ed). 2006, 164–181.

Sgall, P. 1967a. Generativní popis jazyka a česká deklinace [Generative description of language and Czech declension], Prague: Academia.

Sgall, P. 1967b. Functional Sentence Perspective in a Generative Description of Language. In *Prague Studies in Mathematical Linguistics,* 2: 203–225. Reprinted (shortened) in Sgall, P. 2006. 275–301.

Sgall, P. 1979. Towards a Definition of Focus and Topic. In *Prague Bulletin of Mathematical Linguistics,* 31:3–25; 32:24–32; reprinted in *Prague Studies in Mathematical Linguistics*, 1981, 78:173–198.

Sgall, P. 2006. *Language in Its Multifarious Aspects.* In Hajičová, E. & J. Panevová. (eds). Prague: Karolinum.

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

Sgall, P., E. Hajičová & E. Buráňová. 1980. *Aktuální členění věty v češtině* [Topic-focus articulation of the sentence in Czech]. Prague: Academia.

Sgall, P., E. Hajičová & J. Panevová. 1986. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. In Mey, J.L. (ed). Dordrecht: Reidel and Prague:Academia.

Sgall, P., O. Pfeiffer, W.U. Dressler & M. Půček. 1995. Experimental Research on Systemic Ordering. *Theoretical Linguistics,* 21:197–239.

Steedman, M. 1985. Dependency and Coordination in the Grammar of Dutch and English, In *Language,* 61:523–568.

Steedman, M. 1991. Structure and Intonation, In *Language,* 67:260–296.

Steedman, M. 1996. *Surface Structure and Interpretation*. Cambridge, Mass. and London: The MIT Press.

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

Steinberg, D.D. & L.A. Jakobovits, (eds). 1971. *Semantics – An Interdisciplinary Reader*. Cambridge: Cambridge University Press.

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

Strawson, P. 1964. Identifying Reference and Truth Values. In *Theoria* 30:96–118. Reprinted in Steinberg & Jakobovits, (eds). 1971, 86–99.

Weil, H. 1844. *De l'ordre des mots dans les langues anciennes comparées aux langues modernes*, Paris. Translated as *The Order of Words in the Ancient Languages Compared with That of the Modern Languages*, Boston 1887, reedited, Amsterdam: Benjamins.

Zikánová, Š. 2006. What Do the Data in PDT Say about Systemic Ordering in Czech? *Prague Bulletin of Mathematical Linguistics,* 86:39–46.

Žabokrtský, Z. 2005. Resemblances between Meaning-Text Theory and Functional Generative Description. In J.D. Apresjan, L.L. Iomdin, (eds). *Proceedings of the 2nd International Conference of Meaning-Text Theory*, Moscow, 549–557.