An Automatic Procedure for Topic–Focus Identification

Eva Hajičová*
Charles University

Hana Skoumalová†
Charles University

Petr Sgall*
Charles University

The dichotomy of topic and focus, based, in the Praguean Functional Generative Description, on the scale of communicative dynamism, is relevant not only for a possible placement of the sentence in a context, but also for its semantic interpretation. An automatic identification of topic and focus may use the input information on word order, on the systemic ordering of kinds of complementations (reflected by the underlying order of the items included in the focus), on definiteness, and on lexical semantic properties of words. An algorithm for the analysis of English sentences has been implemented and is discussed and illustrated on several examples.

1. Topic and focus in Functional Generative Description

In the framework of Functional Generative Description (FGD), elaborated by the Prague research group of theoretical and computational linguistics, topic and focus are understood as constituting one of the hierarchies typical for the (underlying) syntactic structure of the sentence. A detailed discussion of this framework, including explicit definitions of the basic notions, can be found in Sgall, Hajičová, and Panevová (1986), Hajičová and Sgall (1987), Sgall (1987), Petkevič (1987; in preparation). In the present paper it is possible only to characterize these notions briefly and informally. However, an algorithm is included that determines the topic–focus structure of the input sentences (on their nonmarginal readings). The function of this algorithm can be checked, and its usefulness, connected with that of the underlying framework, may then be compared with other approaches.

In the prototypical case, the **topic** (theme, "given" information) can be understood as that part of the sentence structure that is being presented by the speaker as readily available in the hearer's memory, whereas the **focus** (comment, rheme) is what is being asserted about the topic. If negation or another "focalizer" (such as *only, even, also*) is present, then primarily its scope (or its "focus") is constituted just by the focus of the sentence. Thus, for example, in *The king of France is not bald,* the subject, which is the topic of the sentence on its preferred reading, is outside the scope of negation, so that if the sentence is uttered as referring to the world we live in, it is connected with a presupposition failure: the existence of the king of France is presupposed (entailed even by the negative sentence). Our notion of topic appears to have much in common with the more recently characterized concept of background or restrictor; on the other

* Institute of Formal and Applied Linguistics, Charles University, Malostranské nám. 25, 118 00 Praha 1, Czech Republic.
† Institute of Theoretical and Computational Linguistics, Charles University, Celetná 13, 110 00 Praha 1, Czech Republic.

hand, our focus comes close to nuclear scope (see, especially, Partee 1992).[1] Rochemont and Culicover (1990) analyze a notion of focus similar to ours, using the framework of Principles and Parameters theory; however, their theory runs into problems in cases in which the focus is not a single constituent (see Koktová 1993).

The topic–focus articulation (TFA) is both expressed by grammatical means (word order, morphemes or their clitic versus "strong" shapes, syntactic constructions, position of the sentence stress or "intonation center") and semantically relevant. Thus, it is impossible to account for the structure of the sentence without describing TFA.

In FGD the sentence structure is understood as based on the relation of **syntactic dependency** and is thus extremely flat. The syntactic relations in the narrow sense are handled in the form of a dependency tree, with the main verb constituting the label of its root and the branches being labeled by symbols denoting the kinds of complementation. These include, on the one hand, inner participants or arguments, such as Actor, Addressee, Objective, and, on the other hand, free modifications, such as Locative, Means, Manner, Cause, several temporal and directional modifications, those of Condition, Regard, Accompaniment, etc.[2]

Before we present our algorithm, let us illustrate the basic notions of our framework by a few examples (with uppercase letters denoting a non-final, i.e., marked placement of the intonation center; in sentences without capitalization, the intonation center is supposed to be placed at the end):

(1) (a) John talked to few girls about many problems.

(b) John talked about many problems to few girls.

(2) (a) ?John made a canoe out of every log.

(b) John made a CANOE out of every log.

(3) (a) Everybody in this room knows at least two languages.

(b) At least two languages are known by everybody in this room.

(4) (a) They arrived by car at the lake.

(b) They arrived at the lake by car.

(5) (a) She gave several children a few apples.

(b) She gave a few apples to several children.

(6) (a) They moved from Boston to Chicago.

(b) They moved to Chicago from Boston.

[1] See Rooth (1985), Krifka (1992), and works cited therein, in which similar issues are discussed. Some of the authors concentrate on focalizers and their scopes and/or foci, whereas we consider a sentence containing no focalizer to constitute the prototypical case (it is open to discussion whether in this a "covert focalizer," such as the assertive modality of the main verb, is present on some level of representation).

[2] Note that we do not discuss the relations of coordination and of apposition in this paper. In FGD, the correlates of function words in syntactic representations do not take the form of specific nodes in the tree. Rather, these correlates take the form of labels on edges (see the syntactic units previously illustrated) or of parts of complex labels on nodes (such as values of morphological categories, e.g. Plural, Feminine, Preterite, Conditional, and semantic distinction within the individual syntactic categories of adverbial modifications, such as the meanings of the prepositions *in, on, above, under* with Locative).

These pairs of sentences, as is known from previous discussions, show that TFA is relevant not only for a possible placement of the sentence in a context, but also for its **semantic** interpretation, even for its truth conditions. In (1)–(3), the semantic difference concerns the distribution of the scopes of quantifiers. On the preferred reading the quantifier belonging to the topic has a wide scope, which is in agreement with the view according to which the focus is asserted "about" the topic. For example, a paraphrase of the preferred reading of (3)(a) would be *About everybody in this room I tell you that (s)he knows at least two languages.* In (4)–(6), differences in presuppositions are connected with at least some readings of the sentences. Thus, (6)(b) presupposes that "they" moved to Chicago (since this phrase belongs to the topic). This presupposition is absent in (6)(a), in which the *to* phrase belongs to the focus; the presupposition that "they" moved somewhere from Boston is triggered only by those readings of this sentence in which the *from* phrase belongs to the topic.

Thus, as for TFA, in all such cases a characteristic difference may be found. The (a) sentences are ambiguous in that the penultimate sentence part in some readings (and thus in some dependency-based syntactic representations of these sentences) belongs to the focus and in others to the topic. In the (b) examples, this ambiguity is absent. The item now placed in the penultimate position (or that following the intonation center, which marks the most dynamic item) belongs to the topic in all readings.

Different surface means are used to express the differences in TFA in the English examples. Even in English, there are instances of "free" word order (i.e., of surface word order determined directly by TFA), as in (1), (4), and (6). In other cases a secondary placement of the intonation center is used, as in (2). In others, specific syntactic constructions allow for an appropriate shape of surface word order, such as passivization in (3) or the prepositional expression of Addressee in (5).

The distribution of TFA may be checked by such means as the **question test**. For example, (1)(a) may be a full answer to a question such as (7). But this is not the case with (1)(b). On the other hand, (1)(b) may be a full answer to (8), in the way that (1)(a), rather than (b), may be a full answer to (9):[3]

(7) What do you know about John?

(8) To whom does John speak about many problems?

(9) How does John behave towards few girls?

Thus, (1)(a) can answer two of questions (7)–(9), whereas (1)(b) can answer just one of them. This also applies to (2)–(6). As we have just seen, the (b) sentences, rather than their (a) counterparts, are restricted to one of the possible TFAs. In this sense, the order of the relevant complementations (arguments and free modifications) in the (a) sentences may be understood as primary and that in (b) as secondary. It is then possible to specify a basic, **systemic ordering** (SO) of the kinds of complementations of every verb (noun, adjective).

After several years of research in this domain, including several series of psycholinguistic experiments with Czech and with German sentences (see Pfeiffer, Půček,

[3] What is meant by "full answer" here is not only that the topic part of the answer is redundant (contained in the question and not deleted in the answer), but also that, when formulating the answer, a speaker does not assume any other "given" or "known" information than that contained in the question. A detailed discussion of the question test and its comparison with other operational criteria (based on a natural response or commentary, often connected with negation) can be found in Sgall, Hajičová, and Panevová (1986, Chapter 3).

and Sgall, 1994), as well as investigations with native speakers of English, we hypothesize that the SO of some of the main kinds of complementations in English has the following shape:[4]

Time – Actor – Addressee – Objective – Origin – Effect – Manner – Directional. *from* – Means – Directional. *to* – Locative

The core of our experiments has consisted of checking (with native informants) whether the (a) or (b) sentence in such a pair can answer a question in which neither of the two relevant complementations is mentioned, or one in which only one of them is mentioned. Thus, for example, (6)(a) is a natural answer to *What are Jane and Jim doing?* or to *Have you heard about Jane and Jim recently?* On the other hand, (6)(b) occurs much more probably as an answer (however redundant) to *From where did Jane and Jim move to Chicago?* In this sense the examples above can be understood as corroborating the cited shape of SO for some of the pairs of complementations: Example (1) illustrates that Addressee precedes Objective, since only (1)(a) is possible as an answer to (7). In the same vein, example (2) documents that Objective precedes Origin (see Section 2, in which the relevance of the secondary position of the intonation center is discussed). Other examples also have a similar significance: (3) for the pair Actor–Objective, (4) for Manner–Directional. *to*, (5) again for Addressee–Objective, and (6) for the two Directionals.

SO is one of the factors relevant for word order and for the placement of the intonation center. In the prototypical case (when the intonation center occupies the rightmost position and other conditions, discussed in Section 2, are met), SO directly determines the underlying word order in the focus part of the sentence. The following rule holds:

Rule 1
If a sentence part A precedes another one, B, under SO, and both A and B are in the focus of a sentence S, then A precedes B in the word order of S.

2. Communicative Dynamism and Word Order

To be able to characterize the procedure determining some of the main points of TFA and to illustrate the output language of our parser, we have to add a brief discussion of certain issues concerning word order.

The word order of natural languages is determined not only by SO, but also by other factors. If an item occurs in the topic, it may be placed more to the left than would correspond to SO; the specific order of the elements of the topic is influenced by the speaker's discourse strategy. There are also grammatical rules, such as those concerning the positions of the verb (e.g., in the "second position" in German), of the adjective or another modifier before or after the head noun in a noun group, and of clitics. Cases in which the intonation center has a secondary (non-final) position must also be considered.

The interplay of word order and these other factors allows for a specification of the scale of **communicative dynamism** (CD). This scale is responsible for the "dynamic" progression of parts of the sentence, from topic proper through intermediate parts to

[4] In German and in most Slavonic languages the situation differs in that Objective and Effect follow several of the adverbial modifications.

focus proper as the most dynamic element (carrying the intonation center).[5] CD is semantically relevant for the scopes of quantifiers, as illustrated by example (10).

(10) (a) It was JOHN who talked to few girls about many problems.

(b) It was JOHN who talked about many problems to few girls.

This example differs from (1) in that the two groups containing the relevant quantifiers (*few girls* and *many problems*) both are in the topic of the sentence, whereas in (1)(a) and (b), at least one of them belongs to the focus on all readings. Thus, with (1) it may be claimed that only the boundary between topic and focus is responsible for the different distribution of the scopes of quantifiers; however, (10) shows that the individual degrees in the scale of CD also influence the meaning of the sentence: even if the two quantifiers both are contained in the topic, the one contained in the less dynamic sentence part has the wide scope on the preferred reading.[6]

Another point shows the importance of including CD in syntactic representations of sentences: on the scale of CD, there is always a certain step dividing the sentence (its syntactic representation) into the topic and the focus as the less dynamic and the more dynamic parts of the sentence, respectively. Therefore, in our syntactic representations of sentences, we work with the scale of CD as with the underlying word order. An alternative choice would be to mark the scale of CD by specific indexing of the lexical occurrences in the sentence.

We can now formulate Rule 1, from Section 1, in a more precise form, as rule 1′ referring to the underlying word order (CD), rather than to the surface one.

Rule 1′
If a sentence part A precedes another one, B, under SO, and both A and B are in the focus of a sentence S, then A precedes B in the underlying word order of S.

It follows from Rule 1′ that B can be less dynamic than A (i.e., B can precede A in the underlying word order) in a sentence S only if B belongs to the topic of S. As mentioned above, in the topic part the underlying word order often differs from SO, which is conditioned mainly by the speaker's discourse strategies. The speaker chooses the topic proper (the least dynamic element) among the items assumed to be most salient in the hearer's memory. Often this is what was referred to by the focus proper of the preceding utterance.[7]

Now we can see why the (b) examples in Section 1 lack the ambiguity present in the (a) sentences. For example, in (1)(a) the underlying (and surface) order of the two

[5] Like many other linguistic notions, that of the intonation center is far from clear. Since it is not possible to discuss this issue in depth here (which has been the objective of a rich discussion), we can only characterize our standpoint as follows: (a) in a sentence having more than one sentence stress, we understand the last (rightmost) one as the intonation center, and (b) we assume that the prototypical (unmarked) position of the intonation center is (in English) at the last word of the sentence. We are aware that these formulations do not cover all the possible cases, but the more or less marginal exceptions must be left aside for the aim of the present paper.

[6] We cannot discuss here the issues concerning other possible interpretations of sentences such as (1) and (10). Their acceptability often depends on the lexical setting of the sentence and on pragmatic factors. This also concerns the cases of "group reading" (e.g., in *The three men built those two houses*) or of J. Hintikka's "branching quantifiers."

[7] More precisely, the topic proper refers to one of those items that, at the given time point, are most salient in the stock of knowledge shared by the speaker and (according to the speaker's assumption) by the hearer. The set of highly salient items (called "established" in our earlier writings) can be compared to the "focus list" of Grosz (1977).

rightmost complementations (*to few girls* and *about many problems*, i.e., Addressee and Objective) is in accordance with SO, but in (1)(b) this is not so: the Objective, which was most dynamic (rightmost in underlying word order) in (a) does not occupy this position in (b). This means that it is included in the topic of sentence (b) on all its readings (i.e., in all syntactic representations of the sentence). This is similar with examples (3)–(6), and with (2) the switch of the intonation center plays the same role as the switch of word order in the other examples.

On the other hand, the ambiguity of the (a) sentences is determined by the fact that the scale of CD is in accordance with SO here and that one of the complementations thus belongs to the topic in some of the readings and to the focus in others. For example, in (1)(a) the group *to few girls* is in such an ambiguous position: in some of the readings, the boundary between topic and focus precedes this group; in others, the boundary follows it. In both (a) and (b), the most dynamic complementation belongs to the focus on all the readings.

The dichotomy of topic and focus concerns the sentence as a whole. In sentences with items embedded more deeply than the immediate complementations of the main verb, it is necessary to characterize the positions of individual word occurrences in the sentence in a more specific way. We therefore work with the distinction of **contextually bound** (CB) and **non-bound** (NB) lexical occurrences. Operational criteria to distinguish between these two values again may be found in the question test and in similar procedures. For example, only CB items can have the shape of weak pronouns or be deleted (thus, in *He LEFT* the subject is CB, whereas in *HE left* it is NB).

A CB item is always considered to be less dynamic than its head and than its NB sister nodes (i.e., nodes depending on the same head). Thus, in *He left YESTERDAY*, the subject is CB and thus less dynamic than its head, the verb, and also than its NB sister, the adverb. This implies that the main verb is always more dynamic than all its CB complementations and less dynamic than the NB ones; i.e., in the scale of CD the verb stands immediately after or before the boundary between topic and focus.

To illustrate the notion of contextual boundness, we present two additional examples:[8]

(11) (How do you find your neighborhood?) Our(CB) new(NB) neighbor(CB) has stolen(NB) my(CB) CAR(NB).

(12) (Which teacher do you mean?) I(CB) mean(CB) our(CB) teacher(CB) of CHEMISTRY(NB).

These sentences can also be used to exemplify how, on the basis of the dichotomy of CB and NB items, the notions of topic and focus can be defined more exactly (for a more explicit formulation, see Sgall et al. 1986, Chapter 3):

(i) The main verb and its immediate complementations belong to the topic if they are CB and to the focus if they are NB.

(ii) More deeply embedded items belong to the topic (focus) if their head words (in the framework of dependency syntax) belong there.

[8] In our syntactic representations we do not handle the correlates of function words as (labels of) separate nodes; they have the shape of indices accompanying auto-semantic lexical units (see footnote 2). This appears to be more adequate, since both their semantic and syntactic properties differ substantially from auto-semantic words. Furthermore, it is not economical to enlarge the number of nodes beyond necessity, adding special nodes for prepositions or articles, which can accompany only their nouns, or for conjunctions and auxiliary verbs, which can accompany only lexical verbs and which do not accept any (other) arguments or modifications of their own.

(iii) If the verb and all its immediate complementations (in other words, all elements of the center of the sentence) are CB, then only the NB item(s) embedded under the most dynamic element of the center constitutes the focus, with the rest of the sentence belonging to its topic.

In (11) the noun *neighbor*, being CB (as a definite subject noun usually is), belongs to the topic, according to (i), and so does *new* as its modifier, according to (ii), although it is NB. The verb and the noun *car* both belong to the focus, according to (i), and so does *her*, according to (ii).

In (12) all of the CB words belong to the topic according to (i) or (with *our*)[9] to (ii); then (iii) determines the adjunct of *chemistry* as the focus of (12).

The (underlying) syntactic representations of sentences in our framework can now be illustrated (with several simplifications) in the form of linearized dependency trees. With this notation, every dependent item is included in its pair of parentheses, labeled by the corresponding syntactic symbol. This symbol occurs as the label of the edge in the tree, or as a subscript following a parenthesis in the linearized representation:[10]

(13) A neighbor gave a boy a book.

(13′) (neighbor.Indef)$_{\mathrm{Act}}$ give.Pret (boy.Indef)$_{\mathrm{Addr}}$ (book.Indef)$_{\mathrm{Obj}}$

(14) A painter arrived at a French village on a nice September day.

(14′) (painter.Indef)$_{\mathrm{Act}}$ (village.Indef (French)$_{\mathrm{Gener}}$)$_{\mathrm{Dir}}$ arrive.Pret (day.Indef (September)$_{\mathrm{Gener}}$ (nice)$_{\mathrm{Gener}}$)$_{\mathrm{Time}}$

(15) The neighbor met him yesterday.

(15′) (neighbor)$_{\mathrm{Act}}$ (he)$_{\mathrm{Obj}}$ meet.Pret^{t} (yesterday)$_{\mathrm{Time}}$

Most of our symbols (for Indefinite, Preterite, Actor, Addressee, Objective, Directional) should be self-explanatory; Gener(al Relationship) is the free modification typical for an adjectival modifier of a noun. In (15′), the superscript t denotes the verb as belonging to the topic (being CB), although this is not in an immediate correspondence with its position in the surface word order. In English, the word order is grammatically restricted; thus also in (14) the verb occupies the position after the subject, in the surface, although it is followed by a CB item. Typically, the position of the verb in TFA (and often also the position of a complementation) is ambiguous, and in the present examples we give only one of the possible readings of the sentence. The unmarked case, when the verb belongs to the focus, is left without a specific notation mark here. The TFA positions of the complementations are indicated by their positions in the underlying word order, i.e., in CD: those belonging to the focus stand to the right of the head verb, and those in the topic stand to the left of it.

Let us note that, for example, the written shape of (14) may also be pronounced with a secondary placement of the intonation center, as in (16), with another TFA. This pronunciation is not probable, but it is possible, as after such a co-text as (17):

[9] The contextual boundness of this pronoun in the given position is derived from its indexical character and its associative link with the speaker. For this reason, such an item can always be referred to as "established," or "recoverable," or "identifiable" in the terminology of Halliday (1967) or Chafe (1976).

[10] A detailed discussion of dependency trees and the labels of their edges (the syntactic values, i.e., kinds of arguments and modifications) and of their nodes (the values of morphological categories) was presented by Sgall, Hajičová, and Panevová (1986, Chapter 2). In the notation presented here, the morphological categories are handled so that only their marked values are indicated. Unmarked (prototypical) values such as Singular, Present, and Definite are assumed "by default."

(16) A painter arrived at a French VILLAGE on a nice September day.

(16′) (painter.Indef)$_{\text{Act}}$ (day (September)$_{\text{Gener}}$
(nice)$_{\text{Gener}}$)$_{\text{Time}}$ arrive.Pret (village.Indef (French)$_{\text{Gener}}$)$_{\text{Dir}}$

(17) In the autumn, painters often look for nice sceneries in most different environments.

Similarly, with (15) there is a less probable pronunciation (possible only in specific contexts) with the pronoun HIM stressed. Many, though not all, such marked cases are accounted for by the parser described in Section 3. The output language of this parser has been illustrated by examples (13′)–(16′).

Sections 1 and 2 have introduced our treatment of topic and focus. As such, the comments and the examples could not cover all the possible sentence structures. The interested reader can find a more general and precise characterization of the basic notions we work with in Sgall et al. (1986), Hajičová and Sgall (1987), and Petkevič (in preparation). The main objective of the paper is to present a procedure specifying the TFA of a sentence. However, at this stage, not all combinations of marginal phenomena are covered by our algorithm.

3. A Procedure for the Identification of TFA

An automatic identification of topic, focus, and degrees of communicative dynamism, discussed in a preliminary way by Hajičová and Sgall (1985), can be based on the following considerations:[11]

Languages with a high degree of "free" word order (such as most Slavonic ones) differ from English or French in that a secondary position of the intonation center is frequent there only in spoken discourse.[12] On the other hand, in technical texts (which typically are written), there is a strong tendency to arrange the words so that the intonation center falls on the last word of the sentence (where it need not be phonetically manifested), with the exception of course, of enclitic words. This usage, occasionally recommended by manuals and textbooks concerning, for example, the stylistics of Czech or Russian, makes it possible to read such a text aloud without paying much attention to the choice of the placement of the intonation center.

A general procedure for determining TFA in such languages can then be based on the following points:

(i) All complementations preceding the verb are CB and thus belong to the topic. As for the complementations following the verb, Rule 2 may be stated:

Rule 2

The boundary between topic (to the left) and focus (to the right) can be drawn between any two elements following the verb, provided that those belonging to the focus are arranged in the surface word order in accordance with SO (see Section 1).

[11] As usual in computational linguistics, it is impossible to handle all marginal and exceptional cases by a relatively simple, general procedure. Natural language processing always requires solutions covering first the typical (or most frequent) cases and only then more complex procedures accounting for peripheral phenomena. Thus, the present paper also does not aim at a complete solution that would handle all possible cases appropriately.

[12] Note that one can specify the position of the intonation center even with a written sentence: the sentence can be read aloud either correctly (in accordance with the author's intention) or incorrectly. The fact that there are also cases in which different placements of the intonation center are suitable for the given context is not immediately relevant.

(ii) The verb is ambiguous as to its position in the topic or in the focus.

(iii) If a spoken utterance (with its intonation center identified) is analyzed, then (i) and (ii) hold for sentences with normal intonation (intonation center at the end). However, if a non-final element carries the intonation center, then all the complementations standing after this element belong to the topic; for the rest of the sentence, (i) and (ii) hold; the bearer of the intonation center belongs to the focus.

In English the surface word order is determined by grammatical rules to a large extent, so that intonation plays a more decisive role than in the Slavonic languages. The written shape of the sentence does not suffice here to determine TFA to such a degree as it does in Czech, for example. Rule 2 also applies, but otherwise only certain important regularities can be stated here on the basis of word order and grammatical values (especially a definite noun group is often CB, and an indefinite one regularly is NB). To be able to reduce the ambiguity of the written shape of the sentence as much as possible, it is necessary to take into account certain semantic clues.

Especially with Locative and Temporal modifications, it is important to distinguish between specific information (e.g., *on a nice September day, on October 22, 1991, seven months ago*) and items containing just a general setting (e.g., *always*) or being directly determined by the utterance itself, such as indexicals, like *today* and *this year*. The latter examples usually belong to the topic, whereas the former ones typically occur in the focus.

As for the verb, it is important to have access to the verb of the preceding utterance and to use a systematic semantic classification of the verbs. If the main verb of sentence n has the same meaning as (or a meaning included in) that of sentence n – 1 (in the sense of hyponymy), then it belongs to the topic. Also, verbs with very general lexical meanings (such as *be, have, happen, carry out,* and *become*) may be handled as belonging to the topic. Otherwise (i.e., in the unmarked case), the verb typically belongs to the focus (in which case no subscript is being used in our representations).

An algorithmic procedure has been formulated by H. Skoumalová, completing the parsing of a written English sentence so as to identify its TFA. In the output of this procedure, many ambiguities remain, but sentences (even in their spoken shape) often are ambiguous as to their TFA. Thus it should be understood as a good result if the procedure identifies such an ambiguity. In its present form, however, the algorithm has several limitations. It can process only simple sentences. It determines the appurtenance of an element to topic or to focus, but does not specify CD within topic. It also handles just the verb and its complementations; deeper embedded elements are left aside for the time being.

The algorithm has been formulated as follows:

(a) After the dependency structure of the sentence has been identified by the parser, so that also the underlying dependency relations (valency positions) of the complementations (to the governing verb) are known, the verb and all the complementations are first assumed to be NB, i.e., to belong to the focus, which we denote by f.

(b) If the verb occupies the rightmost position in the sentence and its subject is

(ba) definite (including noun groups with *this*, with *one of the*, etc.), then the verb is NB, i.e., f, and its subject is CB, belonging to the topic, which we denote as t;

(bb) indefinite, then the subject is f and the verb is t. In either case, the other complementations are handled according to (cb) below.

(c) If the verb does not occupy the rightmost position, then:

(ca) the verb itself is understood as t, if it has a very general lexical meaning (see above), or as f if its meaning is very specific, or else as ambiguous (t/f);

(cb) the complementations preceding the verb are denoted as t, with the exception of an indefinite subject and of a specific (i.e., neither general nor indexical; see above) Temporal complementation; either of the latter two is characterized as t/f;

(cc) to the right of the verb,

(i) if there is a single complementation, and this is a definite noun group or a personal pronoun, it is t/f;

(ii) if the rightmost complementation is Temp or Loc, and it is specific, it is f; otherwise it is t (i.e., is understood as standing to the left of the verb in the underlying word order and is shifted there);

(iii) if A is the left item of the rightmost pair that now (after the possible change of word order carried out according to (ii) above) fails to follow SO (see Section 1 and Rule 2), then A belongs to the topic (t), and so do all the complementations between A and the verb; the rightmost complementation of the whole sentence is f (only a personal pronoun following another one (including those of the third person) is t/f in this position), all those standing between A and the rightmost one are t/f;

(iv) if neither (ii) nor (iii) is met and the rightmost complementation is indefinite, it is f;

(v) all remaining complementations to the right of the verb are t/f.

(d) If all the complementations have been determined as t or t/f, then

(da) if the verb was t/f after point (ca) and the rightmost complementation is a definite noun group, an indexical word, or pronoun, then this rightmost element gets t(f), which denotes a specific kind of ambiguity: this element is to be understood as having f only in case there is no other f in the reading of the sentence;

(db) if (da) does not apply, then both the rightmost element of the sentence and its verb get t/f.

(e) The remaining representations containing no f are deleted.[13]

We are aware that our procedure does not cover all possibilities occurring in English sentences. Deeper embedded elements have not yet been properly analyzed, and different pronominal forms should be classified in a much more detailed way. Other cases, assumed to occur with low probability (such as, for example, *The neighbor GAVE the boy a book*, or *The neighbor gave HIM the book*), are not taken into account.

[13] We assume that at least one reading of the sentence has been assigned an f (NB element) by now. The readings without a focus are not valid representations of sentences, since one of the basic assumptions is that every sentence contains a focus.

When implemented (together with a simplified parser),[14] the algorithm was checked with a set of sentences having our examples (1)–(3) as its core, and it yielded the expected results, as presented in Section 4.

4. Examples

Let us first reproduce here examples (13)–(15) from Section 2 (with a changed numbering), accompanied by the corresponding input strings of our program, in which the occurring word forms are complemented already by the lexical data. The program presupposes that each word form occurring in the text has undergone lexical (and morphemic) analysis so that it has been assigned the relevant data found in the lexicon. These include word class and sem(antic features) such as hum(an). The verbs are accompanied by their valency frames (grids), which also include data on the surface shape of the individual kinds of complementations so that it is easy to reconstruct the original sentence at the end of the procedure:[15]

(1) A neighbor gave a boy a book.

(1″)
```
verb(topic(f),touch(0),sem(interm),label(gave),
ltree(act(topic(f),touch(0),def(0),so(01),surf(np),
label(neighbor),ltree(det(a)))),rtree(addr(topic(f),touch(0),
def(0),so(012),surf(np),label(boy),ltree(det(a))),
obj(topic(f),touch(0),def(0),so(0123),surf(np),label(book),
ltree(det(a)))))
```

(2) A painter arrived at a French village on a nice September day.

(2″)
```
verb(topic(f),touch(0),sem(interm),label(arrived),
ltree(act(topic(f),touch(0),def(0),so(01),surf(np),
label(painter),ltree(det(a)))),rtree(loc(topic(f),touch(0),
def(0),sem(gen),so(012345678),surf(np),label(village),
ltree(prep(at),det(a),generic(french))),temp(topic(f),
touch(0),def(0),sem(gen),so(0),surf(np),label(day),
ltree(prep(on),det(a),generic(nice),generic(september)))))
```

(3) The neighbor met him yesterday.

(3″)
```
verb(topic(f),touch(0),sem(interm),label(met),
ltree(act(topic(f),touch(0),def(1),so(01),surf(np),
label(neighbor),ltree(det(the))),tree(obj(topic(f),touch(0),
def(1),so(0123),surf(ppers),label(him),ltree),temp(topic(f),
```

[14] It is not an objective of this paper to present a parser of English. The parser that has been used as a basis of our procedure is founded on dependency syntax and covers just the simple shapes of English sentences. Its lexical scope can be enlarged easily, if the added lexical items are accompanied by appropriate grammatical data, especially by valency (case) frames specifying the optional and obligatory arguments (Actor, Addressee, Objective, Origin, and Effect, with verbs). Prepositions are being analyzed just in one or two meanings each.

[15] The notation differs slightly here from that of Section 2; the complex symbols are reflected here by subtrees in which the nodes for function words are still present. The symbol *topic* denotes here whether the given item belongs to the topic or to the focus, *touch* stores the information if the complementation has been already determined, *sem* is the semantic information about the verb (general, specific, intermediate), and *ltree* and *rtree* are the left and right subtrees in the dependency tree. The word form is saved under *label*. *so* contains information about the position of the complementation in systemic ordering, and *surf* is the surface form (noun group, personal pronoun, indexical word, etc.). The other symbols are self-explanatory.

```
touch(0),def(1),sem(gen),so(0),surf(index),label(yesterday),
ltree)))
```

The output of the procedure characterized in Section 3 is as follows:

```
(1*)    a neighbor(t/f) gave(t/f) a boy(t/f) a book(f)

(2*)    a painter(t/f) arrived(t/f) at a french village(t) on a nice
        september day(f)

(3*)    the neighbor(t) met(t/f) him(t/f) yesterday(t(f))
```

To illustrate how our procedure works for the sentences differing from (1)–(3) in the values of delimiting features (definite–indefinite), in word order, and so on, we add a list of these sentences with simplified, perspicuous results of the procedure, i.e., with the values t and f produced by our algorithm added to the autonomous (autosemantic) lexical occurrences. Ambiguity is denoted here in an abbreviated way, so that "t/f" means "t in some readings and f in others" (in combination with the values of other words in the sentence), and "t(f)" means "obtaining f only in case there is no other f in the sentence." In this way it is easy to check whether the decision points in the algorithm, which are illustrated by the examples, have been handled adequately.

```
(1)(A)   a neighbor(t/f) gave(t/f) a boy(t/f) a book(f)
   (B)   a neighbor(t/f) gave(t/f) a boy(t/f) the book(t/f)
   (C)   a neighbor(t/f) gave(t/f) the boy(t/f) a book(f)
   (D)   a neighbor(t/f) gave(t/f) the boy(t/f) the book(t/f)
   (E)   the neighbor(t) gave(t/f) a boy(t/f) a book(f)
   (F)   the neighbor(t) gave(t/f) a boy(t/f) the book(t/f)
   (G)   the neighbor(t) gave(t/f) the boy(t/f) a book(f)
   (H)   the neighbor(t) gave(t/f) the boy(t/f) the book(t/f)
   (I)   a neighbor(t/f) gave(t/f) him(t/f) a book(f)
   (J)   a neighbor(t/f) gave(t/f) him(t/f) the book(t/f)
   (K)   the neighbor(t) gave(t/f) him(t/f) a book(f)
   (L)   the neighbor(t) gave(t/f) him(t/f) the book(t/f)
   (M)   a neighbor(t/f) gave(t/f) a book(t) to a boy(f)
   (N)   a neighbor(t/f) gave(t/f) the book(t) to a boy(f)
   (O)   a neighbor(t/f) gave(t/f) a book(t) to the boy(f)
   (P)   a neighbor(t/f) gave(t/f) the book(t) to the boy(f)
   (Q)   the neighbor(t) gave(t/f) a book(t) to a boy(f)
   (R)   the neighbor(t) gave(t/f) the book(t) to a boy(f)
   (S)   the neighbor(t) gave(t/f) a book(t) to the boy(f)
   (T)   the neighbor(t) gave(t/f) the book(t) to the boy(f)
   (U)   a neighbor(t/f) gave(t/f) it(t) to him(t/f)
   (V)   the neighbor(t) gave(t/f) it(t) to him(t/f)

(2)(A)   a painter(t/f) arrived(t/f) at a french village(t) on a
         nice september day(f)
   (B)   a painter(t/f) arrived(t/f) at a french village(t/f)
         yesterday(t(f))
   (C)   a painter(t/f) arrived(t/f) at the french village(t) on
         a nice september day(f)
   (D)   a painter(t/f) arrived(t/f) at the french village(t/f)
         yesterday(t(f))
```

(E) the painter(t) arrived(t/f) at a french village(t) on a nice september day(f)

(F) the painter(t) arrived(t/f) at a french village(t/f) yesterday(t(f))

(G) the painter(t) arrived(t/f) at the french village(t) on a nice september day(f)

(H) the painter(t) arrived(t/f) at the french village(t/f) yesterday(t(f))

(I) a painter(t/f) arrived(t/f) there(t) on a nice september day(f)

(J) a painter(t/f) arrived(t/f) there(t/f) yesterday(t(f))

(K) yesterday(t) a painter(t/f) arrived(t/f) there(t(f))

(3)(A) a neighbor(t/f) met(t/f) a boy(t) on a nice september day(f)

(B) a neighbor(t/f) met(t/f) a boy(t/f) yesterday(t(f))

(C) a neighbor(t/f) met(t/f) the boy(t) on a nice september day(f)

(D) a neighbor(t/f) met(t/f) the boy(t/f) yesterday(t(f))

(E) the neighbor(t) met(t/f) a boy(t) on a nice september day(f)

(F) the neighbor(t) met(t/f) a boy(t/f) yesterday(t(f))

(G) the neighbor(t) met(t/f) the boy(t) on a nice september day(f)

(H) the neighbor(t) met(t/f) the boy(t/f) yesterday(t(f))

(I) the neighbor(t) met(t/f) him(t/f) yesterday(t(f))

We assume that the sentences are pronounced so that the intonation center is carried by the rightmost sentence part bearing an index f. Thus, for instance, (3)(H) corresponds to the following sentences:

(3) (H1) The neighbor MET the boy yesterday.

(H2) The neighbor met the BOY yesterday.

(H3) The neighbor met the boy YESTERDAY.

5. Conclusion

As we have mentioned in Section 3, our algorithm does not cover all cases of TFA occurring in English sentences. For the present stage of research, it has been possible to account only for the primary shape of sentence structure (the verb with its arguments and free modifications) and for the prototypical cases of TFA.

Future research in the domain of automatic processing of TFA thus may concentrate on solving further problems connected with secondary cases. Above all, this concerns the following points in which a more general procedure could be formulated:

(i) The procedure should also take into account deeper embedded sentence parts (embedded verb clauses, modifiers in noun groups, etc.). Criteria to decide on these sentence parts being CB or NB will make it necessary to work with a detailed semantic classification of lexical items and to take into account the analysis of preceding co-text.

(ii) Such "focus-sensitive adverbs" or "focalizers" as *only, also, even, mostly,* negation, etc. (see Section 1 and footnote 1) should be considered, since their foci may differ from the focus of the sentence as a whole (although in the prototypical case such a difference does not occur).

(iii) If a semantic comparison of lexical items with those present in the preceding utterances of the discourse is made possible (see point (i)), then the cases of ambiguity resulting from the procedure could be considerably reduced. In any case, for practical applications it will be necessary to work with preferences, excluding the least probable readings.

(iv) One of the most promising prospects is to join a procedure of the kind described in the present paper with an acoustic analysis of spoken discourse, in which the position of the intonation center could be determined as one of the important factors.

We hope, however, that the procedure outlined in the present paper can serve as one of the starting points both for a comparison of the views on TFA based on dependency and on other syntactic theories and for achieving a relatively complete algorithmic analysis of TFA as that dimension of the sentence structure which permits a characterization of the sentence in its fundamental interactive nature.

References

Chafe, Wallace L. (1976). "Givenness, contrastiveness, definiteness, subjects, topics and point of view." In *Subject and Topic,* edited by Charles N. Li. 25–55. New York: Academic Press.

Grosz, Barbara J. (1977). "The representation and use of focus in dialogue understanding." Technical Report 151, SRI International. Doctoral dissertation, University of California at Berkeley.

Hajičová, Eva, and Sgall, Petr (1985). "Towards an automatic identification of topic and focus." In *Proceedings, Second Conference of the European Chapter of the Association for Computational Linguistics,* Geneva. 263–267.

Hajičová, Eva, and Sgall, Petr (1987). "The ordering principle." *Journal of Pragmatics* 11:435–454.

Halliday, Michael A. K. (1967). "Notes on transitivity and theme in English." *Journal of Linguistics* 3:37–81, 199–244; 4:179–215.

Koktová, Eva (1993). Review of Rochemont and Culicover (1990). *Prague Bulletin of Mathematical Linguistics,* 58:88–93.

Krifka, Manfred (1992). "Focus and quantification." In *SALT II, Proceedings, Second Conference on Semantics and Linguistic Theory,* edited by C. Barker and D. Dowty. Columbus: Ohio State University.

Partee, Barbara (1992). "Quantificational structures and compositionality." Presented at the Third Hungarian Symposium on Logic and Language. To be published by Kluwer, Dordrecht in the Proceedings, edited by L. Kálmár.

Petkevič, Vladimír (1987). "A new dependency based specification of underlying representations of sentences." *Theoretical Linguistics* 14:143–172.

Petkevič, Vladimír (in prep.). *An Extended Dependency-Based Specification of Underlying Representations of Sentences.* To be published by Charles University, Prague.

Pfeiffer, Oskar; Půček, Michael; and Sgall, Petr (1994). "Die Thema-Rhema-Gliederung im Deutschen und ihre automatische Analyse." In *Computatio linguae, 2, Zeitschrift für Dialektologie und Linguistik, Beihefte,* 83, edited by U. Klenk, 148–164.

Rochemont, Michael S., and Culicover, Peter W. (1990). *English Focus Constructions and the Theory of Grammar.* Cambridge: Cambridge University Press.

Rooth, Mats (1985). "Association with focus." Doctoral dissertation, University of Massachusetts, Amherst.

Sgall, Petr (1987). "The position of Czech linguistics in theme-focus research." In *Language Topics. Essays in Honour of Michael Halliday,* Vol. I, edited by Ross Steele and Terry Threadgold, 47–55. Amsterdam: John Benjamins.

Sgall, Petr; Hajičová, Eva; and Panevová, Jarmila (1986). *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects,* edited by Jacob L. Mey. Prague: Reidel, Dordrecht and Academia.