A Dependency-Based Parser for Topic and Focus

Eva Hajičová
Faculty of Mathematics and Physics
Charles University
Malostranské n. 25
118 00 Praha 1
Czechoslovakia

1. Introduction

A deepened interest in the study of suprasegmental features of utterances invoked by increasing attempts at a build-up of algorithms for speech recognition and synthesis quite naturally turned attention of the researchers to the linguistic phenomena known for decades under the terms of theme-rheme, topic-comment, topic-focus. In the present paper we propose a linguistic procedure for parsing utterances in a "free word order" language, the resulting structure of which is a labelled W-rooted tree that represents (one of) the (literal) meaning(s) of the parsed utterance. Main attention will be paid to the written form of language; however, due regard will be also paid to (at least some of) the suprasegmental features and additional remarks will be made with respect to parsing strategies for written and spoken English.

2. Dependency-Based Output Structures

2.1. The procedure is based on the linguistic theory of functional generative description as proposed by Sgall (cf. Sgall, 1964,1967; Sgall et al., 1986). The representation of the meaning(s) of the sentence - i.e. the output of the analysis - is a projective rooted tree with the root labelled by a complex symbol of a verb and its daughter nodes by those of the complementations of the verb, i.e. participants (or - in another terminology - the cases, theta-roles, valency), as well as adverbials. The relation between the governor (the verb) and the dependants (its daughter nodes) is a kind of dependency between the two nodes. The complementations of the daughter nodes (and their respective complementations, etc.) are again connected with their governors by an edge labelled by a type of dependency relation. The top-down dimension of the tree thus reflects the structural characteristics of the sentences. The left-to-right dimension represents the deep word order, see Sect. 3 below. Structures with coordination may be then represented by complex dependency structures (no longer of a tree character) with a third dimension added to the tree structure (Plátek, Sgall and Sgall, 1984), or, alternatively, nodes of quite special properties can be added to the tree itself (Močkořová, 1989). Such a type of description can dispense with problems of constituency and "spurious" ambiguity and offers an effective and economic way of representing sentence meaning.

To illustrate (with several simplifications, some of which will be clarified in Sect. 3 below) the type of representations characterized till now, we present in Fig. 1 an underlying representation of the sentence (1).

(1) In August, a seminar on parsing technologies will be organized by CMU in Pittsburgh.

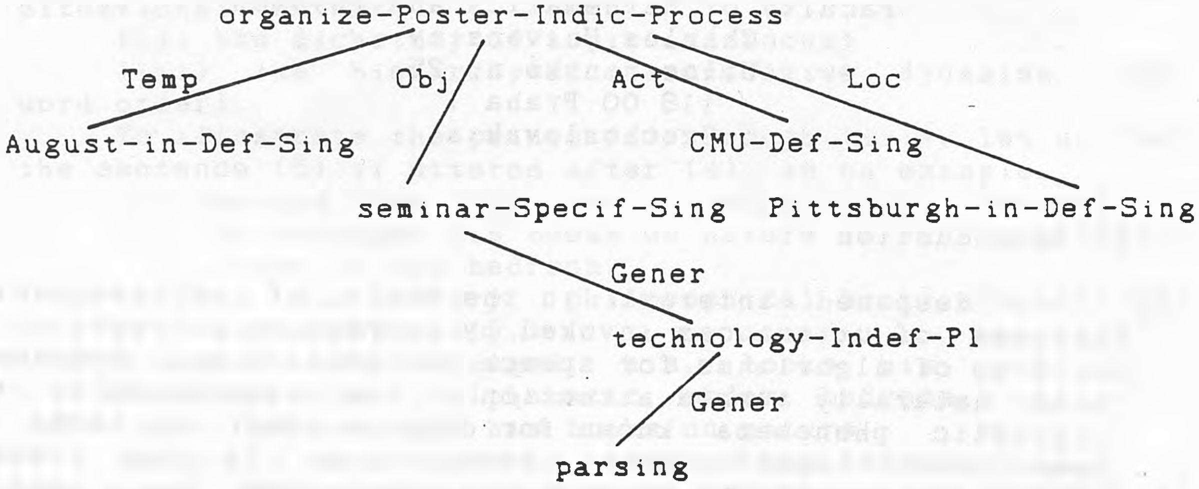

Fig.1

2.2 A dependency oriented account of syntactic(o-semantic) relations offers a rather straightforward way for a formulation of a lexically-driven parsing procedure, since a great part of the relevant information is projected from the frames belonging to the lexical entries of the heads. In the description we subscribe to, valency slots are not understood just in the sense of obligatory or regular kinds of complementation, but are classified into

(i) inner participants (theta roles, each of which can be present at most once with a single head token) and free modifications;

(ii) obligatory and optional; this distinction can be made with both kinds of complementations quoted under (i) depending on the specific heads.

As for (i), five inner participants are being distinguished (for motivation, see Panevová, 1974; Hajičová and Panevová, 1984), namely deep subject (Actor), deep object (Patient, Objective), Addressee, Origin (Source) and Effect; among free modifications, there belong Instrument, Locative, Directional, Manner, several temporal adverbials, adverbials of cause, condition, regard, General relationship, etc. As for (ii), an operational test was formulated that helps to determine which of the complementations with a given lexical head is obligatory (although perhaps deletable) and which is optional; the test is based on judgements on the coherence of a simple dialogue (see Panevová, 1974).

Both (i) and (ii) are reflected in the valency frames of individual lexical entries in the lexicon. Thus, e.g., for the verb *to change*, the valency frame consists of two obligatory slots for Actor and Objective, two optional slots for Source and Effect (*to change something from something into something*) and a list of free modifications, which can be stated once for all the verbs. If one of the free modifications is obligatory with a certain head (e.g. Directional with *arrive*, Appurtanance with *brother*, Material with *full*), this has to be indicated in the valency frame of the relevant head. 2.3 Dependency can be operationally defined on the basis of endocentricity (cf. Sgall and Panevová, 1989, following Kulagina ,1958). If in a syntactic construction one of two members of the construction can be left out, while the other retains the distributional properties characteristic for the given pair, then the member that can be omitted is considered to depend on the other: e.g., in *Jim read* a *book* the sentence part a *book* can be omitted without the sentence losing its grammaticality; thus, the verb rather than *the book* is the head of the construction. The set of word classes that is determined on independent grounds can then be used to identify the "direction of dependency" in other (exocentric) constructions: though in *Jim bought* a *book* the sentence part a *book* cannot be omitted, *buy* and *read* are assigned a single word class (on independent morphemic and syntactic criteria) and thus it may be postulated that *bought* rather than a *book* is the governor (head) of the construction *bought* a *book*. In a similar vein, a construction such as *Jim read* can be substituted in its syntactic position (as constituting a sentence) by a subjectless verb in many languages (cf. Latin *Pluit*; also in English *It rains* the surface subject *it* has no semantic value: it cannot be freely substituted by a noun or by another pronoun and is equivalent to the Latin ending).

2.4 It is not our objective in the present paper to contrast dependency structures with those of phrase structure grammar. Let us only mention in conclusion of this section, that among the main advantages of dependency trees there is the relatively small number of nodes; the basic syntactic hierarchy can be described without any non-terminal nodes occurring in the representations of sentences,although in their derivations non-terminals can be used without the limitations characteristic of Gaifman's approach to dependency. In addition, if function words are understood as mere grammatical morphemes having no syntactic autonomy, then their values can be treated as indices, i.e. parts of complex labels of nodes, as illustrated in Fig. 1 above. In this way, the component parts of syntactically autonomous units can be represented correctly as having other syntactic properties than the autonomous units themselves, and the representations do not get necessarily complicated.

3. The Semantic Impact of Topic-Focus Articulation

3.1 The topic-focus articulation of an utterance has an impact on the semantic interpretation of the given utterance. It is important to notice that (a) and (b) are two different sentences in (2) as well as in (3), though the semantic difference is much more important in (3) than in (2). With (2) the two sets of propositions to which the two sentences correspond assign the value "true" to the same subset of possible worlds, which is not the case with (3)[1]. (The intonation center is denoted by italics.).

(2)(a) Mother is *coming*.
(2)(b) *Mother* is coming.

(3)(a) I do linguistics on *Sundays*.
(3)(b) On Sundays, I do *linguistics*.

In the representations of meaning as characterized in Sect. 2, we distinguish:

(i) contextually bound (CB) and non-bound (CN) nodes, where "contextually" covers both verbal co-text and situational context;

(ii) the dichotomy of topic and focus;

(iii) the hierarchy of communicative dynamism (deep word order).

To illustrate the points (i) through (iii), let us take the sentence (5) if uttered after (4), as na example.

(4) How did John organize the books in his library?

(5) He arranged his books on nature in an alphabetic order in his bedroom.
(In his library, philosophical books are arranged chronologically.)

(i) CB nodes: he, arranged, his, books, his
NB nodes: nature, alphabetic, order, bedroom

(ii) topic: he arranged his books on nature
focus: in an alphabetic order in his bedroom

(iii)deep word order (dots stand for the modifications of the nodes explicitly mentioned)
he - ...books... - arranged - order... - ...bedroom

3.2 The impact of the three aspects (i) through (iii) can be illustrated by the examples (6) through (8), respectively:

(6)(a) (You have just listened to our night concert.) The compositions of Chopin were played by S. *Richter*. We will devote to him also our next *programme*.
him = Richter

(6)(b) (You listen to our night concert.) Chopin's compositions were played by S. *Richter*. We will devote to him also our next *programme*.
him = Chopin

(7)(a) Staff only behind the *counter*.

(7)(b) *Staff* only behind the counter.

(8)(a) It was *John* who talked to few girls in many towns.

(8)(b) It was *John* who talked in many towns to few girls.

The distinction between (a) and (b) in (6) consists in the different preference of anaphoric use of referring expressions if the possible referent is mentioned in the previous context by an NB or a CB element (as *Chopin* in (a) or in (b), respectively); in both cases, the anaphoric elements are in the topic part of the sentence.

The sentence (7)(a) differs from (7)(b) only in that *the counter* is in the focus part of (a), while *staff* is in the focus part of (b), which difference leads to a significant distinction in interpretation: (a) holds true if the members of the staff are (to stay) only behind the counter and nowhere else, while (b) holds true if the space behind the counter is (to be) occupied only by the members of the staff; in contrast to (b), the sentence (a) holds true also if there is somebody else than a member of the staff in that space. In (7), the relevant semantic distinction is rendered by a different placement of the intonation center; in (3) above, the same effect results from a word order change.

The clefting in (8) univocally points to *John* as the focus of the sentence, the rest being its topic; the two sentences (a) and (b) differ as to the (deep) order of Locative and Addressee. This distinction again has an important semantic impact: with (a), there was a group of girls who were few, and the same group was talked to in many towns, while with (b) John talked in each of the many towns with (maybe) a (different) small group of girls. This difference need not be reflected in the surface word order: the same effect is reached by a shift of intonation center, see (9)(a) and (b).

(9)(a) John talked to few girls in many *towns*.
(9)(b) John talked to few *girls* in many towns.

4. Parsing Procedure for Topic and Focus

4.1 The proposed procedure of automatic identification of topic and focus is based on two rather strong hypotheses:
(i) the boundary between topic and focus is always placed so that there is such an item A in the representation of meaning that every item of this representation that is less (more) dynamic than A belongs to the topic (focus); in the primary case the verb meets the condition on A and is itself included in the focus;
(ii) the grammar of the particular language determines an ordering of the kinds of complementations (dependency relations) of the verb, of the noun, etc., called 'systemic ordering' (SO). The deep word order within focus is determined by this ordering; with sentences comprising contextually bound items, these items stand to the left in the hierarchy of communicative dynamism and their order (with respect to their governors) is determined by other factors. An examination of Czech in comparison with English and several other languages has led to the conclusion that the SO of some of the main complementations is identical for many languages, having the form Actor - Addressee - Objective, As for Instrument, Origin, Locative, it seems that English differs from Czech in that these three complementations follow Objective in English, though they precede it in Czech. It need not be surprising that languages differ in such semantically relevant details of their grammatical structures as those concerning SO - similarly as they appear to differ in the semantics of verbal aspects, of the articles, of dual number, etc.

We assume further that every sentence has a focus, since otherwise it would convey no information relevant for communication; however, there are sentences without topic.

4.2 For an automatic recognition of topic, focus and the degrees of CD, two points are crucial:

(A) Either the input is a spoken discourse (and the recognition procedure includes an acoustic analysis), or written (printed) texts are analyzed.

(B) Either the input language has (a considerable degree of) the so-called free word order (as in Czech, Russian, Latin, Warlpiri) or its word order is determined mainly by the grammatical relations (as in English, French).

Since written texts usually do not indicate the position of intonation center and since the "free" word order is determined first of all by the scale of communicative dynamism, it is evident that the former cases in (A) and (B) do not present so many difficulties for the recognition procedure as the latter cases do.

A written "sentence" corresponds, in general, to several spoken sentences which differ in the placement of their intonation center, cf., e.g., ex. (3) above. In languages with the "free" word order this fact does not bring about serious complications with written technical texts, since there is a strong tendency to arrange the sentences in such texts so that the intonation center falls on the last word of the sentence (if this word is not enclitical).

4.31 A procedure for the identification of topic and focus in Czech written texts can then be formulated as follows (we use the term 'complementation' or 'sentence part' to denote a subtree occupying the position of a participant or free modification as discussed in Sect. 2 above):

(i)(a) If the verb is the last word of the surface shape of the sentence (SS), it always belongs to the focus.

(i)(b) If the verb is not the last word of the SS, it belongs either to the topic, or to the focus.

Note: The ambiguity accounted for by the rule (i)(b) can be partially resolved (esp. for the purposes of the practical systems) on the basis of the features of the verb in the preceding sentence: if the verb of the analyzed sentence is identical with the verb of the preceding sentence, or if a relation of synonymy or meaning inclusion holds between the two verbs, then V belongs to the topic. Also, a semantically weak, general verb such as *to be*, *to become*, *to carry out*, most often can be understood as belonging to the topic. In other cases the primary position of the verb is in the focus.

(ii) The complementations preceding the verb are included in the topic.

(iii) As for the complementations following the verb, the boundary between topic (to the left) and focus (to the right) may be drawn between any two complementations, provided that those belonging to the focus are arranged in the surface word order in accordance with the systemic ordering.

(iv) If the sentence contains a rhematizer (such as *even*, *also*, *only*), then in the primary case the complementation following the rhematizer belongs to the focus and the rest of the sentence belongs to the topic.

Note. This concerns such sentences as *Here even* a *device of the first type can be used.*; in a secondary case the rhematizer may occur in the topic,e.g., if it together with the sentence part in its scope is repeated from the preceding co-text.

4.32 Similar regularities hold for the analysis of spoken sentences with normal intonation. However, if a non-final complementation carries the intonation center (IC), then

(a) the bearer of the IC belongs to the focus and all the complementations standing after IC belong to the topic;

(b) rules (ii) and (iii) apply for the elements standing before the bearer of the intonation center;

(c) the rule (i)(b) is applied to the verb (if it does not carry the IC).

4.33 As for the identification of topic and focus in an English written sentence, the situation is more complicated due to the fact that the surface word order is to a great extent determined by rules of grammar, so that intonation plays a more substantial role and the written form of the sentence displays much richer ambiguity. For English texts from polytechnical and scientific domains the rules stated for Czech in Sect. 4.31 should be modified in the following ways:

(i)(a) holds, if the surface subject of the sentence is a definite NP; if the subject has an indefinite article, then it mostly belongs to the focus, and the verb to the topic; however, marginal cases with both subject and verb in the focus, or with subject (though indefinite) in the topic and the verb in the focus are not excluded;[2]

(i)(b) holds, including the rules of thumb contained in the note;

(ii) holds, only the surface subject and a temporal adverbial can belong to the focus, if they do not have the form of definite NP's;

(iii) holds, with the following modifications:

(a) if the rightmost complementation is a local or temporal complementation, then it should be checked whether its lexical meaning is specific (its head being a proper name, a narrower term, or a term not belonging to the subject domain of the given text) or general (a pronoun, a broader term); in the former case it is probable that such a modification bears the IC and belongs to the focus, while in the latter case it rather belongs to the topic;

(b) if the verb is followed by more than one complementation and if the sentence final position is occupied by a definite NP or a pronoun, this rightmost complementation probably is not the bearer of IC and it thusfinite NP or a pronoun, this rightmost complementation probably is not the bearer of IC and it thus belongs to the topic;

(c) if (a) or (b) apply, then it is also checked which pair of complementations disagreeing in their word order with their places under systemic ordering is closest (from the left) to IC (i.e. to the end of the focus); the boundary between the (left-hand part of the) topic and the focus can then be drawn between any two complementations beginning with the given pair;

(iv) holds.

4.34 If a spoken sentence of English is analyzed, the position of IC can be determined more safely, so that it is easier to identify the end of the focus than with written sentences and the modifications to rule (iii) are no longer necessary. The procedure can be based on the regularities stated in Sect. 4.32.

Up to now, we have taken into account in our discussion on automatic identification of topic and focus in spoken utterances only the position of the intonation center; a question naturally arises whether other features of intonation patterns such as tune and phrasing (in terms of Pierrehumbert) can help as clues for sentence disambiguation as for its topic and focus. Schmerling (1971) was the first, to our knowledge, to propose that the different interpretations of Chomsky's 'range of permissible focus' (which basically corresponds to our 'deep word order', see Hajičová and Sgall, 1975) are rendered on the surface by different intonation patterns; most recently, Pierrehumbert and Hirschberg (1989, Note 5) express a suspicion that the accented word in such cases (within an NP) need not have the same prominence in all the interpretations; they also admit that similar constraints on the accenting of parts of a VP are even less understood.

5. Parsing Sentences in a Text

To resolve some complicated issues such as the ambiguity of pronominal reference, a whole co-text rather than a single sentence should be taken into account. Several heuristics have been proposed to solve this problem; e.g., Hobbs (1976) specifies as a common heuristics for pronominal resolution the determination of the antecedent on the basis of the hearer's preference of the subject NP to an NP in the object position (in a similar vein, Sidner ,1981, in her basic rule tests first the possibility of co-specification with what she calls 'actor focus'), the other strategy including inferencing and factual knowledge. Following up our investigation of the hierarchy of activation of items of the stock of knowledge shared by the speaker and the hearer (see Hajičová and Vrbová, 1982; Hajičová, 1987; Hoskovec, 1989; Hajičová and Hoskovec, 1989), we maintain that also this hierarchy should be registered for parsing sentences in a text. We propose to use a partially ordered storage space, reflecting the changes of the activation (prominence) of the elements of the information shared by the speaker and the hearer. The rules assigning the degrees of activation after each utterance take into account the following factors:

(i) whether the given item was mentioned in the topic part or in the focus part of the previous utterance: mentioning in the focus part gives the item the highest prominence, mentioning in the topic part is assumed to assign a one degree lower activation to the given item;

(ii) grammatical means by which the given item is rendered in the surface shape of the utterance: mentioning by means of a (weak) pronoun gives a lower prominence than mentioning by means of a noun;

(iii) association with the items explicitly mentioned in the utterance: items which are associated with the items explicitly mentioned in the preceding utterance get a certain level of prominence, though lower than those mentioned explicitly; it is assumed that the association relations can be classified according to the 'closeness' of the items in question so that some types of associations receive higher degrees of activation than others (e.g., is-a relation is 'closer' in this sense than the part-of relation);[3]

(iv) non-mentioning of a previously mentioned item: an item

that has been introduced into the activated part of the stock of shared knowledge but is not mentioned in the subsequent utterances loses step by step its prominence;
(v) not only the immediate degree of activation after the given utternace is relevant for the assignment of reference but also the sequence of degrees of salience from the whole preceding part of the text; thus if an item is being mentioned subsequently for several times in the topic of the sentence, its salience is maintained on a high level and it is more likely an antecedent for pronominal reference than an item that appeared in the focus part (with no prominence history) and received thus the highest degree of activation.

6. Concluding Remarks

Since even in such languages as English or French, surface word order corresponds to the scale of communicative dynamism to a high degree (although such grammatical means as passivization, or the inversion of *make out of* to *make into* , etc., often are necessary here to achieve this correspondence), it is useful in automatic language processing to reflect the word order of the input at least in its surface form. If the effects of the known surface rules on the verb placement, on the position of adjectives, genitives[4], etc., before (or after) nouns, and so on, are handled, and if the items mentioned in the preceding utterance are stored (to help decide which expressions are contextually bound), then the results may be satisfactory.

Notes.

1 With (2) as well as with (3) the presuppositions triggered by (a) and (b) differ, so that different subsets of possible worlds get the value 'false'; e.g., (2)(b) differs from (2)(a) in presupposing that someone is coming.

2 For the solution of such cases, it again is useful to "remember" the lexical units contained in the preceding utterance, cf. the Note to (i)(b) in Sect. 4.31 above.

3 It is more exact to understand the association relationships in terms of natural language inferencing (concerning the occurrence of a single associated item) than in terms of the activation of the whole set of items associated with an occurrence of a possible 'antecedent'.

4 This has been done, at least to a certain degree, in the experimental systems of English-to-Czech and Czech-to-Russian translation, implemented in Prague.

References

Hajičová, E. (1987), Focussing - A Meeting Point of Linguistics and Artificial Intelligence. In: Artificial Intelligence II - Methodology, Systems, Applications (ed. by Ph. Jorrand and V. Sgurev), Amsterdam, 311-322.

Hajičová, E. and T. Hoskovec (1989), On Some Aspects of Discourse Modelling. In: Fifth Int. Conference on Artificial Intelligence and Information-Control Systems of Robots (eds. I. Plander and J. Mikloško), Amsterdam.

Hajičová, E. and J. Panevová (1984), Valency (Case) Frames of Verbs. In: Sgall (1984), 147-188.

Hajičová, E. and P. Sgall (1975), Topic and Focus in Transformational Grammar, Papers in Linguistics 8, 3-58.

Hajičová, E. and J. Vrbová (1982), On the Role of Hierarchy of Activation in the Process of Natural Language Understanding, in Horecký (1982), 107-113.

Hobbs, J. R. (1976), Pronoun Resolution. Rep. 76-1, Dept. of Computer Science, City College, City Univ. of New York.

Horecký, J., ed. (1982), Coling 82 - Proceedings of the Ninth Int. Conf. on Computational Linguistics, Prague - Amsterdam.

Hoskovec, T. (1989), Modelling a Pragmatical Background of Discourse. In: AI '89, Prague, 289-296.

Kulagina, O. S. (1958), Ob odnom sposobe opredelenija grammatičeskich ponjatij, Problemy kibernetiki 1, 203-214.

Močkořová, Z. (1989), Generalizivané podkladové závislostni struktury (Generalozed Underlying Dependency Structures), diploma theses

Panevová, J. (1974), On Verbal Frames in Functional Generative Description I, Prague Bulletin of Mathematical Linguistics 22, 3-40; II, 23 (1975), 17-52.

Pierrehumbert J. and J. Hirschberg (1989), The Meaning of Intonational Contours in the Interpretation of Discourse.

Plátek M., Sgall, J. and P. Sgall (1984), A Dependency Base for a Linguistic Description. In: Sgall (1984), 63-97.

Schmerling ,S. F. (1971), Presupposition and the Notion of Normal Stress. In: Papers from the Seventh Regional Meeting. Chicago Linguistic Society , 242-253.

Sgall, P. (1964), Generative Beschreibung 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.

Sgall, P. (1967), Functional Sentence Perspective in a Generative Description. In: Prague Studies in Mathematical Linguistics 2, 203-225.

Sgall, P., ed. (1984), Contributions to Functional Syntax, Semantics, and Language Comprehension, Amsterdam - Prague.

Sgall, P., Hajičová, E. and J. Panevová (1986), The Meaning of the Sentence in Its Semantic and Pragmatic Aspects, Dordrecht - Prague.

Sgall, P. and J. Panevová (1989), Dependency Syntax - A Challenge, Linguistics 15.

Sidner, C. L. (1981), Focusing for Interpretation of Pronouns. American Journal of Computational Linguistics 7, 217-231.