TOWARDS AN AUTOMATIC IDENTIFICATION OF TOPIC AND FOCUS

Eva Hajičová and Petr Sgall

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

ABSTRACT

The purpose of the paper is (i) to substantiate the claim that the output of an automatic analysis should represent among other things also the hierarchy of topic-focus articulation, and (ii) to present a general procedure for determining the topic-focus articulation in Czech and English.

(i) The following requirements on the output of an automatic analysis are significant:
(a) in the output of the analysis it should be marked which elements of the analyzed sentence belong to its topic and which to the focus;
(b) the scale of communicative dynamism (CD) should also be identified for every representation of a meaning of the analyzed sentence, since the degrees of CD correspond to the unmarked distribution of quantifier scopes in the semantic interpretation of the sentence;
(c) the analysis should also distinguish topicless sentences from those having a topic, which is relevant for the scope of negation.

(ii) For an automatic recognition of topic, focus and the degrees of CD, two points are crucial:
(a) either the input language has (a considerable degree of) the so-called free word order (as in Czech, Russian), or its word order is determined mainly by the grammatical relations (as in English, French);
(b) either the input is spoken discourse (and the recognition procedure includes an acoustic analysis), or written (printed) texts are analyzed.

In accordance with these points, a general procedure for determining topic, focus and the degrees of CD is formulated for Czech and English, with some hints how the preceding context can be taken into account.

1. We distinguish between the level of linguistic meaning (de Saussure's and Hjelmslev's "form of content", Cosieru's "Bedeutung", others "literal meaning") and its interpretation in the sense of truth-conditional, intensional logic (see Materna and Sgall, 1980; Sgall, 1983).

For some purposes of automatic treatment of natural language (including machine translation) it is sufficient if the output of the procedure of analysis is more or less identical with the representation of the (linguistic) meaning of the sentence. For other purposes, such as that of full natural language comprehension, it is necessary to go as far as the semantic (truth-conditional) interpretation, using a notation that includes variables, operators, parentheses and similar means.

The topic-focus articulation (TFA) is understood as one of the hierarchies of the level of meaning, whose other two hierarchies are that of dependency syntax (close to case grammar) and that of coordination (and apposition) relations. The basic task of a description of TFA is to handle the differences between such sentences as John gave Mary a BOOK. John gave a book to MARY. It was MARY John gave a book. It was a BOOK what John gave to Mary. It was JOHN who gave Mary a book. It was JOHN who gave a book to Mary. (Capitals denote the position of the intonation center.)

In Fig. 1 we present a simplified representation of one of the meanings shared by the sentences in which the intonation center is placed on BOOK; Fig. 2 corresponds to one of the meanings shared by the sentences with the intonation center on MARY.

The following requirements on the output of an automatic analysis are significant, whatever approach to the description of the structure of the sentence has been chosen:

a) In the output of the analysis it should be marked which elements of the analyzed sentence belong to its topic

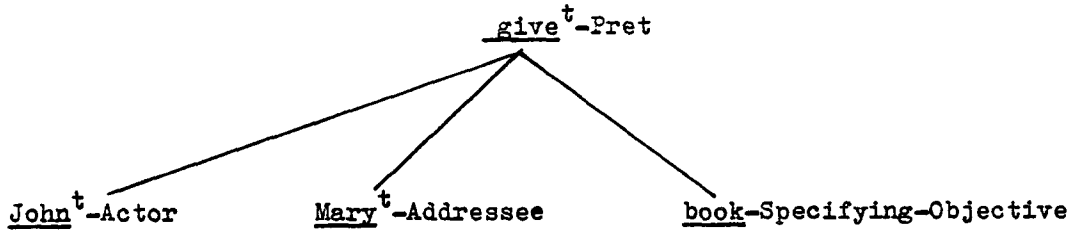

Figure 1. A simplified representation of one of the meanings of the sentences John gave Mary a BOOK. It was a BOOK what John gave to Mary. The superscript t indicates that the given occurrence of the lexical unit is included in the topic; the left-to-right ordering of the nodes corresponds to the scale of communicative dynamism.

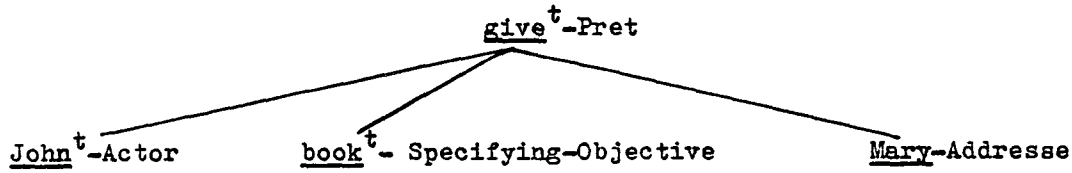

Figure 2. A simplified representation of one of the meanings of John gave a book (= one of the books) to MARY. It was MARY (to whom) John gave a book.

and which to its focus, since this is relevant as for which questions can be answered by the sentence; thus e.g. (1) can answer (2) and (3), while (4) can be answered by (5) rather than by (1).

(1) John talked to few girls about many PROBLEMS.
(2) How did John behave? (What about John?)
(3) About what did John talk to whom?
(4) To whom did John talk about many problems?
(5) John talked about many problems to few GIRLS.

Note that in (1) the verb as well as the Addressee belong to the focus in some of the readings, while in others they are included in the topic; in (5) the Addressee belongs to the focus and Objective to the topic in all readings, only the position of the verb here being ambiguous (this can be checked by tests including negation and by the question test, see Sgall et al., 1973).

b) The scale of communicative dynamism or CD (ibid.) should also be identified for every representation of a meaning (underlying structure, etc.) of the analyzed sentence, since the degrees of CD correspond to the unmarked distribution of quantifier scopes in the semantic interpretation of the sentence; thus in (1) and (6) the Addressee includes a quantifier with a wider scope than that of the Objective (on the primary reading), while in (5) and (7) the quantifier of the Objective includes the Addressee in its scope:

(6) It is JOHN, who talked to few girls about many problems.
(7) It is JOHN, who talked about many problems to few girls.

c) The analysis should also distinguish topicless sentences (corresponding, in the prototypical cases, to Kuno's neutral description or to the thetic judgements of classical logic) from those having a topic; this difference is relevant for the scope of negation: only (8)(b) is semantically equivalent to It is not true that fog is falling, whereas in (9)(b) the subject, included in the topic, is outside the scope of negation; a more appropriate paraphrase is: About Father it is not true that he is coming.

(8)(a) FOG is falling.
(b) No FOG is falling.
(9)(a) Father is COMING.
(b) Father is not COMING.

2. It has been found (Hajičová and Sgall, 1980 and the writings quoted there) that the scale of CD coincides to a great part with Chomsky (1971) calls "the range of permissible focus".

With the elements belonging to the focus the scale of CD is determined by the kinds of complementation (deep cases), the order always being in accordance with what we call systemic ordering; for the main participants of the verb in

English this ordering is Time - Actor - - Addressee - Objective - Origin - Effect - Manner - Instrument - Locative (see Seidlová, 1983). The scale of CD differs from this ordering only if at least one of the elements in question belongs to the topic (this is true about the Addressee in (10)(b), about the Origin in (11)(b), and about the Effect in (12)(b) below):

(10)(a) I gave several children a few APPLES.
(b) I gave a few apples to several CHILDREN.
(11)(a) John made a canoe out of a LOG.
(b) John made a CANOE out of a log.
(12)(a) John made a log into a CANOE.
(b) It was a LOG John made into a canoe.

Thus, in the (b) sentences a few apples, a log and a canoe are contextually bound, standing close to a few of those apples, one of the logs, the canoe we spoke about, respectively. In the (a) examples the rightmost complementations belong to the focus (they carry the intonation center), while the complementations standing between them and the verb are ambiguous in this respect: in some meanings of the sentence they are contextually bound and belong to the focus; a similar ambiguity concerns also the verbs in all the examples.

Systemic ordering is language specific; Czech differs from English e.g. in that it has Time after Actor and Object after Instrument; for more details, see Sgall, Hajičová and Panevová (in press, Chapter 3).

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

(A) Either the input is 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), or its word order is determined mainly by the grammatical relations (as in English, French).

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

A written "sentence" corresponds, in general, to several spoken sentences which differ in the placement of their intonation center. Thus, if an adverbial of time or of place stands at the end of the sentence, as in (13), then at least two sentences may be present, see (14)(a) and (b), where the intonation center is marked by the capitals; the TFA clearly differs:

(13) We were swimming in the pool in the afternoon.
(14)(a) We were swimming in the pool in the AFTERNOON.
(b) We were swimming in the POOL in the afternoon.

In languages with the so-called free word order this fact does not bring about serious complications with technical texts, since there is a strong tendency to arrange the words so that the intonation center falls on the last word of the sentence (if this is not enclitical).

3.2 A procedure for the identification of topic and focus in Czech written (first of all technical) texts can then be formulated as follows:[1]

(i)(a) If the verb is the last word of the surface shape of the sentence (SS), it always belongs to the focus.
(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 practical systems) on the basis of the features of the preceding sentence: if the verb V of the analyzed sentence is identical with the verb of the preceding sentence, or if a relation of synonymy or 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, 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

[1] The term complementation (sentence part) is used in the sense of a subtree occupying the position of a deep case or an adverbial. - We disregard the so-called subjective order here, since (in contrast to English) in a Czech written technical text such sentences as The SWITCH was off are extremely rare.

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.[2]

3.22 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).[3]

3.31 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 3.21 should be modified in the following ways:

(i)(a) holds the surface subject of the sentence is a definite NP; if the subject noun has an indefinite article, then the subject 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;

(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 adverbial, then it should be checked whether its lexical meaning is specific (its being a proper name, a narrower term, or a term not a belonging to the subject domain of the given text) or general (a pronoun, a broader term); in the former case it is probable that the adverbial bears IC and belongs to the focus, as in (15) and (16) while in the latter case it rather belongs to the topic, as in (17) or (18), where the word method probably carries the IC:

(15) Several teams carried out experiments with this method during a single week.

(16) Several teams carried out experiments with this method in Berkeley and Princeton.

(17) Several teams ... method during the last decades.

(18) Several teams ... method in this country.

(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 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.

3.32 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

---

[2] 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 the sentence part in its scope is repeated from the preceding co-text.

[3] However, an analysis of spoken discourse should pay due respect not only to the preceding co-text, but also to the situation of the utterance, i.e. to its context in the broader sense.

modifications to rule (iii) are no longer necessary. The procedure can be based on the regularities stated in 3.22.

4. Whenever the preceding context can be taken into account in the analysis, the salient (activated) items (see Hajičová and Vrbová, 1982) should be registered. A "pushdown" principle can be used: the item mentioned as the (last part of the) focus of the last utterance is the most salient in the given time--point of the discourse, while the elements that were mentioned in other positions of this utterance get a lower status in the activated part of the stock of shared knowledge, and those that have not been mentioned in one or several subsequent utterances may fade away (if they do not have a specific position of a "hypertopic", which concerns e.g. those mentioned in a heading). Thus it can be decided in some of the unclear cases (e.g. with temporal adverbials, see point (iii) in 3.31) whether a complementation belongs to the topic or to the focus. This method has its limits: the set of activated items should include not only items mentioned in the text, but also their parts, counterparts and other items connected with them by associative relations; on the other side, if a specific kind of contrast is involved, it is possible that also an item included in this set is mentioned as a part of the focus of the next utterance.

5. A procedure of automatic identification of topic and focus has been incorporated, to a certain degree, in the parser for Czech implemented in Colmerauer s Q-language on EC-1040 (compatible with IBM 360) and briefly described by Panevová and Sgall (1980), Panevová and Oliva (1982). The prototypical cases of topic-focus articulation are also covered by the English parser designed and implemented (using the same programming tools and hardware) by Kirschner (1982) as a part of the English-to-Czech machine translation project.

Both parsers account also for more complex sentences than the examples quoted in this summary; in some such cases there are embedded sentence parts that have (partial) topics and foci of their own.

Both in translation and comprehension the topic-focus articulation should be paid due respect to, since - as we have seen in Sect. 1 - it is relevant not only for an appropriate use of a sentence in different contexts, but also for truth--conditional interpretation, especially of sentences with certain kinds of quantification and with negation.

REFERENCES

Chomsky, N. (1971), Deep Structure, Surface Structure and Semantic Interpretation, in: Semantics (ed. by D.D.Steinberg and L.A.Jakobovits), Cambridge, 193-216

Hajičová, E. and P. Sgall (1980), A Dependency-Based Specification of Topic and Focus, SMIL, No.1-2, 93-140.

Hajičová, E. and J. Vrbová (1982), On the Role of the Hierarchy of Activation in the Process of Natural Language Understanding, in: COLING 82, ed. by J. Horecký, Amsterdam, 107-113.

Kirschner, Z. (1982), A Dependency--Based Analysis of English for the Purpose of Machine Translation, in: Explizite Beschreibung der Sprache und Automatische Sprachverarbeitung IX, Prague.

Materna, P. and P. Sgall (1980), Functional Sentence Perspective, the Question Test and Intensional Semantics, SMIL, No. 1-2, 141-160.

Panevová, J. and K. Oliva (1982), On the Use of Q-Language for Syntactic Analysis of Czech, in: Explizite Beschreibung der Sprache und Automatische Sprach verarbeitung VIII, Prague, 108-117.

Panevová, J. and P. Sgall (1980), On Some Issues of Syntactic Analysis of Czech, Prague Bulletin of Mathem. Linguistics 34,21-32.

Seidlová, I. (1983), On the Underlying Order of Cases (Participants) and Adverbials in English, Prague Bulletin of Mathem. Linguistics 39, 53-64.

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

Sgall, P. (1983), On the Notion of the Meaning of the Sentence, Journal of Semantics 2, 319-324.

Sgall, P., Hajičová, E. and E.Benešová (1973), Topic, Focus, and Generative Semantics, Kronberg/Taunus.