STRUCTURE OF SENTENCE AND INFERENCING IN QUESTION ANSWERING

Eva Hajičová and Petr Sgall

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

ABSTRACT

In the present paper we characterize in more detail some of the aspects of a question answering system using as its starting point the underlying structure of sentences (which with some approaches can be identified with the level of meaning or of logical form). First of all, the criteria are described that are used to identify the elementary units of underlying structure and the operations conjoining them into complex units (Sect.1), then the main types of units and operations resulting from an empirical investigation on the basis of the criteria are registered (Sect.2), and finally the rules of inference, accounting for the relevant aspects of the relationship between linguistic and cognitive structures are illustrated (Sect.3).

1. A system of natural language understanding may gain an advantage from using the underlying structure of sentences (which with some approaches can be identified with the level of meaning or of logical form) as one of its starting points, instead of working with word specific roles. Arguments for such a standpoint, which were presented in Hajičová and Sgall (1980), include the following two main points:

(a) natural language is universal, i.e. its structure makes it possible to express an unlimited number of assertions, questions, etc., by finite means; once its underlying (tectogrammatical) structure is known, it is possible to use it as an output language of natural language analysis in man-machine communication and thus, without any intellectual effort on the side of the user, to ensure the functioning of automatic question answering systems (or of systems of dialogues with robots, etc.); even if many simplifications have been included into such a system, it is then known what has been simplified and it is possible to remove the simplifications whenever necessary (e.g. if the system is to be used for another set of tasks, including the analysis of a broader set of input texts, questions, etc.);

(b) linguistic meaning is systematic, so that the configurations of "deep cases" (valency), tenses, modalities, number, etc. make it possible to find fully reliable information; on the other hand, such systems as those based on scenarios or scripts work in most cases with rules that are valid for the unmarked cases (in a marked case e.g. lunch in a restaurant can be taken by an employee of the restaurant, who does not reserve a table, order the meals and pay for them ...).

To find out which of the semantic and pragmatic distinctions are reflected in the system of language (or, in other words, to find out in what respects the underlying structure of sentences differ from their surface patterns) testable operational criteria are needed; these criteria should help to distinguish:

(i) whether two given surface units are strictly synonymous (i.e. share at least one of their meanings), or not;

(ii) whether a single surface unit has more than one meaning (is ambiguous), or whether a sibgle meaning is concerned, which is vague or indistinct (cf. Zwicky and Sadock, 1975; Kasher and Gabbay, 1976; Keenan, 1978);

(iii) whether a given distributional restriction belongs to the tectogrammatical level, or whether it is given only by the cognitive content itself, i.e. by extralinguistic conditions;

(iv) between a case of deletion (of a tectogrammatical unit by surface rules) and the absence of the given unit in the underlying structure;

(v) between different kinds of tectogrammatical units (e.g. inner participants of cases, and free or adverbial modifications);

(vi) which tectogrammatical unit has been deleted, in case more of them can occupy the deleted position (cf.

the tectogrammatical difference between the elements of the topic and those of the focus of the sentence, or more exactly, between contextually bound and non-bound elements of the meaning of the sentence).

As for (i), a criterion has been elaborated that works similarly as Carnap´s intensional isomorphism, but is adapted for the structure of natural language, the surface grammatical means of which also exhibit synonymy: *He expected that Mary comes* and *He expected Mary to come* are considered synonymous, since with any lexical (and morphological) cast such two sentences correspond to a single proposition (a single truth value is assigned to any possible world).

On the other hand *John talked to a girl about a problem* is not considered to be synonymous with *John talked about a problem to a girl*, since the known (Lakoff´s) examples with a specific quantification do not share their truth conditions; also our simple examples differ in their tectogrammatical structures (having different topic-focus articulations).

For points (ii), (iii) and (v) the classical criteria known from European structural linguistics are used, such as the diagnostic contexts, possibility of coordination, or Keenan´s (1978) criterion of the necessary knowledge of the speaker whether s/he uses an ambiguous item in this or that of its meanings. It should be noted that perhaps each of the criteria has its weak points (often the implications work in one direction only, in some cases not only surface features, but also the tectogrammatical character of the context has to be taken into account, etc.).

Point (iv) can be systematically tested by means of the so-called dialogue test (cf. Hajičová and Panevová, in press): e.g. in *John came* the direction (rather than the starting point or the time point) has been deleted, so that the speaker necessarily knows where John came and can answer such a question (though s/he may not know from where of when John came).

With respect to point (vi) the question test or the tests concerning negation can be used, as far as the topic-focus articulation is concerned; thus e.g. in *John sent a letter to his SISTER* the verb as well as the Objective are ambiguous, since the sentence can (in different contexts) answer e.g. such questions as *What did John do?* (only *John* being included in the topic of the answer, all the rest belonging to its focus), *What did John send where?* (also the verb belonging to the topic of the answer), *What did John do with the letters?* (*a letter* rather than the verb being included in the topic), etc.; the criterion shows that *John* belongs to the topic in all readings of the sentence (since *John* is contained in all relevant questions, if such improbable or secondary pairs are excluded as our sentence answering the question *What happened?*without *John* referring to one of the most activated elements of the stock of shared knowledge at the given time point), and that *his sister* belongs to the focus (not occurring in any relevant question).

2. The framework resulting from an application of the criteria characterized in Sect. I can be briefly outlined as follows:

The elementary units of the underlying structure are of three kinds:

(a) lexical elements (semantic features); in the present paper we do not deal with operations or relations concerning the combining of features into more or less complex lexical meanings;

(b) elementary gramatical meanings (grammatemes), which can be classified as values belonging to various categories or parameters (delimitation, number, tense, aspect, different kinds of modalities, etc.);

(c) syntactic elements (functors) such as Actor, Addressee, Instrument, Directional, etc.

The underlying structure of a sentence can be conceived of as a network (which can be linearized, see Plátek, Sgall and Sgall, in press) the nodes and edges of which are labelled. A label of a node consists of a lexical meaning and a combination of grammatemes from different categories (the set of relevant categories is determined by the word class of the lexical meaning). A label of an edge consists in a functor, which is interpreted either as a Dependency relation, or as one of the relations of Coordination (corresponding to the meanings of *and*, *or*, *but*, *etc.*) or of Apposition. The dependency relations are combined (in the underlying structure of a sentence without coordination or apposition) into a projective rooted tree, the nodes of which are ordered (from left to right) according to the scale of communicative dynamism, which is decisive for the topic-focus articulation of the sentence. The relations of Apposition and Coordination are combined with those of Dependency according to certain rules described in the last quoted paper and illustrated by Fig. 1 to 3.

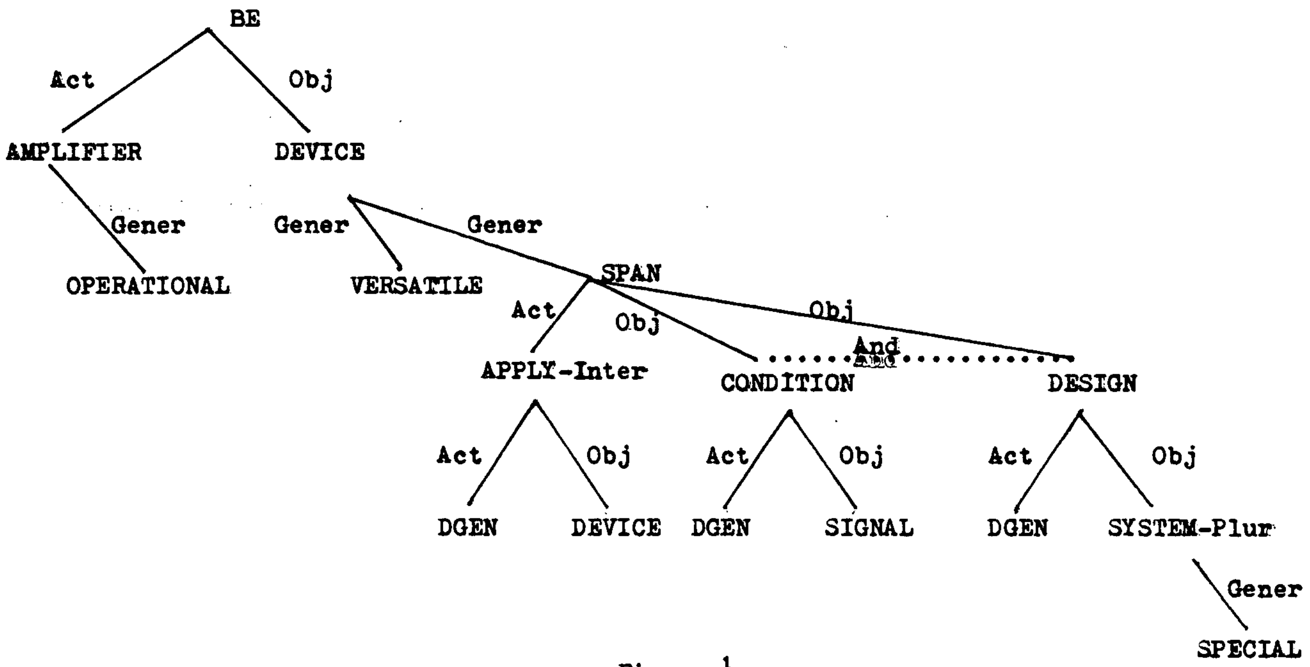

Figure 1.

A simplified underlying representation of *Operational amplifier is a versatile device with applications spanning signal conditioning and special systems design*; Gener is the functor of general relation (the kind of dependency often found between a noun and its modifications), the other symbols are self-explanatory; the grammatemes are written only if they are marked, i.e. Present, Indicative, Singular, Specifying are understood as determined 'by default'.

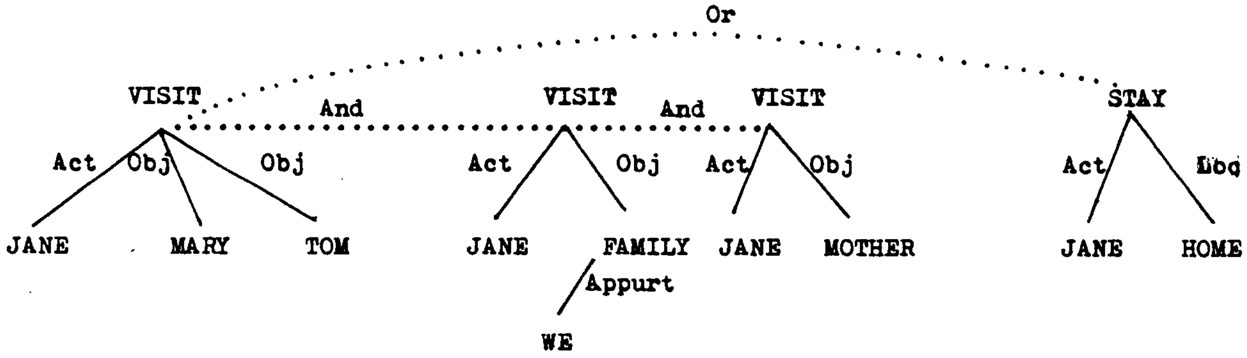

Figure 2.

A simplified underlying representation of *Jane either visits Mary and Tom, our family, and Mother, or she stays at home*.

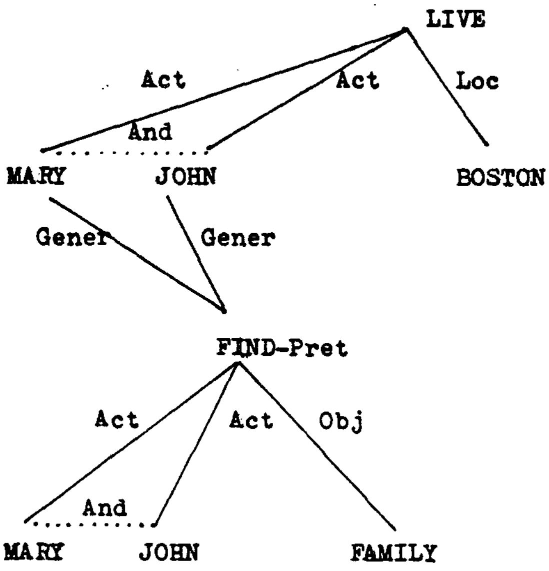

Figure 3.

A simplified underlying representation of *Mary and John, who founded a family, live in Boston*.

Fig. 1 points out how phrasal coordination is handled; in Fig. 2 a configuration of two sentence coordinations (with deletions) appears; Fig. 3 illustrates cases where two coordinated nodes have an expansion (relative clause) in common.

If interjectional sentences, vocative sentences and pseudosentences consisting only in a noun phrase are not discussed, then it can be stated that the root of every tree of the mentioned kind is labelled by a symbol the lexical part of which belongs to the word class of verbs. The kinds (and to a certain part also the order) of the dependency edges going from a node to those dependent on it are determined by the valency frame of the governing word (included in the lexical entry of the given lexical meaning). The kind of dependency relation are specified in two respects,which are relevant for their combinatorial properties: (a) they are classed either as (inner) participants, namely Actor (i.e. Actor/Bearer, or Tesnière's *premier actant* rather than Fillmore's *Agentive)*, Objective, Addressee, Origin and Effect, or as (free) modifications, i.e. Instrument, Manner, Locative, several kinds of Directional and Temporal modifications, Cause, Condition (real and irreal), etc.; (b) they are either obligatory, or optional. Every participant (which occurs only with some governing words, and at most once as dependent on the same token of the governing word) is included in the valency frames of all words on which it can depend; the free modifications are the same for all words belonging to the same word class (on the level of underlying structures), so that they can be listed once for all; only those modifications that are obligatory with a given lexical unit are quoted in its frame.

Two specific cases are important for the empirical investigations: (i) a dependent word present in the underlying structure but deleted in the surface should be distinguished from the absence of the given element on the underlying structure; (ii) with the inner participants it is also necessary to distinguish between the absence of an (optional) participant and a general participant of the fiven kind (this does not concern only the general Actor, typically expressed by *one* in English, but also the Objective, *cf.* Hajičová and Panevová, in press).

3. With this approach, the underlying structures are relatively close to the surface structure of sentences. This is connected with the advantages granted by the universal character of natural language (ensuring that the framework is nottoo narrow and can be generalized if applied to a larger class of texts, etc.). On the other hand, with such a framework it is necessary to use a model of natural language inferencing, if we want the procedure of language understanding to go beyond purely linguistic relationships. If e.g. in a question-answering system based on such a framework not only such answers should be identified that were literally present in the input text, but also those yielded by simple (mostly unconscious) inferencing normally carried out by the reader of the text, then rules of inference can be added. A first tentative set of such rules is being checked in the experiments with the system prepared on the basis of the method TIBAQ in Prague. These rules range from general ones to more or less idiosyncratic cases concerning the relationships between specific words, as well as modalities, hyponymy, etc.

A rather general rule changes e.g. a structure of the form (V-act(N_{Actor})...) into (V-act(D_{Actor}(N_{Instr})...), where V-act is a verb of action, D is a dummy (for the general actor) and N is an inanimate noun; thus *The negative feedback can servo the voltage to zero* is changed into *One can serve the voltage to zero by ...* . A rather specific rule connected with a single verb is that changing (*use* ($S_{Patient}$) (X_{Accomp}) ... into (*use* (X_{Regard}) ($Y_{Patient}$) ...), e.g. *An operational amplifier can be used with a negative feedback* = *With an operational amplifier a negative feedback can be used*. Other similar rules concern the division of conjunct clauses, the possible omission of an adjunct under certain conditions (i.e. if not being included in the topic, e.g. from "It is possible to maintain X without employing Y" it follows that it is possible to maintain X), or several shifts of verbal modalities, esp. a sentence having the main verb with a Possibilitive modality (*can*, *may*) is derived from a positive declarative sentence; in some cases (when the name of a device occupies the position of the Actor of the main verb) also a reverse rule is available, deriving e.g. *The device X is used with a negative feedback* from *The device X can be used with a negative feedback*. Further rules yield a conjunction or a similar connection of two statements; e.g. *X is a device with the property Y* and *X can be applied to handle Z* are combined to yield *X is a device that has the property Y and can be applied to handle Z*; also explicit definitions (including e.g. the verb *call*) are identified and the inference rules allow for replacements of the definiendum by the definiens and vice versa in other assertions.

Besides these kinds of rules it is necessary to study (i) rules standing closer to inference as known from logic (deriving specific statements from general ones, etc.), (ii) rules of "typical" (unmarked) consequence as given e.g. by a ´script´ and (iii) rules of "probable consequences", e.g. if John worked hard in the afternoon and he is tired in the evening, then the latter fact probably was caused by the former (if no other cause was given in text). In our experiment of question answering we do not use these types of inference, but they will be useful for more general systems.

Another direction in which the system probably can be made more flexible concerns the absence of overt quantifiers and marking of their scopes in our underlying structures. One of our next aims consists in the construction of a procedure transducing the underlying structures into a mixed language, which would include means for marking quantifiers and their scopes (similarly to many formal languages of logic), while it would share all other aspects of its structure with the level of underlying representations of natural language.

Colmerauer´s Q language is used for the implementations of the main procedures of the question-answering system, so that e.g. A(B,C(D,E)) represents a tree the head of which is A, which has two sister nodes, B, C, the latter being again expanded by D and E. The tree structure is used in our syntactico-semantic analysis of Czech (prepared by J.Panevová and K.Oliva) and of English (by Z.Kirschner) to represent the dependency relation between nodes. Due to the fact that Q language works only with elementary labels, the complex labels of our description have to be decomposed (i.e.the features and grammatemes of individual work forms occupy similar positions as their daughter nodes). Also the procedures for the application of inference rules and for the identification of (full and partial or indirect) answers to a question given by the user (on the basis of the corpus of input texts that have been analyzed) are programmed in Q language. The synthesis of Czech and morphemic analysis are implemented in PL/1. For a more general system the set of inference rules should be substantially enlarged, and various heuristics, strategies and filters should be formulated in order to keep the number of derived assertions in fixed limits. For these aims the experience gained in the first experiment will be used.

REFERENCES

Hajičová E. and J. Panevová (in press), Valency (Case) Frames of Verbs, in Luelsdorff and Sgall (eds.)

Hajičová E. and P. Sgall (1980), Linguistic Meaning and Knowledge Representation in Automatic Understanding of Natural Language, Prague Bull. of Mathematičal Linguistics 34, 5-19

Kasher A. and D.-M. Gabbay (1976), On the Semantics and Pragmatics of Specific and Non-Specific Indefinite Expressions, Theoretical Linguistics 3,145ff.

Keenan E. (1978), Some Logical Problems in Translation, in Meaning and Translation (ed. by F. Guenthner and M. Guenthner-Reutter), London, 157-189

Luelsdorff P. and P. Sgall (eds.), Contributions to Functional Syntax, Semantics and Language Comprehension, to be published by Benjamins and Academia

Plátek M., Sgall J. and P. Sgall (in press), A Dependency Base for a Linguistic Description, in Luelsdorff and Sgall (eds.)