On Intermediate Structures and Tectogrammatics*

Petr Sgall

an amplified version of a paper presented at
the European Association for Machine Translation Workshop
in Prague, April 22-23, 1999

1. Introductory remarks

Multilingual machine (assisted) translation, the need of which will soon be underlined by the increase of the number of EU languages, requires intermediate structures of a new level. It may be assumed that procedures for analysis and synthesis of the individual languages can be formulated relatively easily on the basis of the methods that have been already tested, mainly on languages from western Europe. However, the transfer procedures should find a new, more economical and perspicuous shape. Moreover, most of the central and east European languages typologically differ from those languages that have already been handled more systematically; displaying a high degree of "free" word order, they exhibit more perspicuous means of the topic-focus articulation (TFA), which represents one of the basic aspects of their grammatical structure. Therefore it appears useful to prepare an alternative structure that could serve as a basis for the treatment of grammatical structures of these new languages, or, eventually, of all languages involved.

To work with a single interlingua seems to be impossible, but what may be understood as feasible is a set of intermediate structures relatively close to each other, especially in what concerns grammar (syntax and morphology); the elaboration of lexical issues (including word formation and phraseology) in this or another style depends on large electronic resources, which now start to come into existence for a growing set of languages (such as WordNet or EuroWordNet). We do not discuss lexical issues in this contribution.

The theoretical basis for the grammar of intermediate structures should be as perspicuous and economical as possible, making use of a metalanguage into which most different theoretical approaches can be reatively easily translated. Such a framework can be found in dependency grammar with complex node labels, describing sentence structure in the form of bracketted strings of indexed symbols. This means that, contrary to many approaches to parsing, we work with an extremely flat structure (using nonterminals only during the derivation of sentence representations, rather than to include them in the descriptions), but, on the other hand, we proceed in the analysis (parsing) to the underlying structures, attempting at a full disambiguation of sentences (although the removal of indistinctness of meaning is impossible on this level, since it requires knowledge based inferencing and the specification of reference). These structures may be used with advantage as intermediate structures for MT, since they reflect basic properties of the languages involved, but at the same time come much closer to each other than other often used output layers of parsers, so that the task of the transfer procedure is made much easier.

2. Dependency syntax

2.1 Overview

Many of the dependency based approaches have been oriented mainly - but by far not only -

at formulating theoretical bases for computational implementation. The late B. Vauquois and his Grenoble group have gained much experience in a detailed description of syntax; I. Mel'chuk, J. Apresjan and others have brought a substantial new insight also in issues of the structure of the lexicon; P. Hellwig has combined very systematically dependency with unification both in theory and in natural language processing; R. Hudson has analyzed different aspects of the grammatical background of an implemented system, J. Kunze, S. Starosta, J. Sleator and many other linguists and mathematicians have devoted attention to dependency based parsers and to a mathematical elaboration of respective formal descriptions. Also our research group of formal and computational linguistics at Charles University in Prague has been convinced for long decades that the dependency based approach is adequate for a theoretical, formal description of language based on a functional view, as well as for automatic processing of language. The framework elaborated by this group, Functional Generative Description (FGD) has been first introduced by Sgall (1967) and discussed in detail also by J. Panevová and E. Hajičová. A synthesis was published by Sgall et al. (1986); further, cf. Sgall (1992) on the levels of language structure, and Platek et al. (1978; 1984) and Petkevič (1995) on a formal specification (both generative and declarative) of the syntactic representations.

In the rest of Sect. 2 it is pointed out how the dependency based view is supported by functional considerations that have been fundamental for European structural linguistics and how the basic ingredients of a formalization of this view can be achieved. In Sect. 3 it is discussed how TFA and the relationships between sentence and coherent discourse can be handled; Sect. 4 deals with the relationships between syntax and morphemics, and shows how a great deal of grammatical information can be included in lexical entries. Sect. 5 aims at substantiating the view that FGD admits a specification of the representations of sentences by means of a very economical mechanism, based on a few general and natural principles.

2.2 Functional Background of Dependency Syntax

As systems based on human interaction, natural languages have developed in the conditions given by human communication and their basic properties have been shaped under the impact of these conditions. One of these properties is the anthropocentrism of syntax. The way of thinking proper to European structural linguistics, esp. to the classical Prague School may contribute to understand this aspect of sentence structure. As V. Mathesius and V. Skalička have shown, the prototypical sentence patterns have either the shape of a (human) action (N1 V N2 or N V), or that of a property (N is A). Referring to a relationship between two participants other than a simple action of one of them affecting (or effecting) the other, grammar often makes it necessary to use the action scheme, although from a cognitive viewpoint this is not fully appropriate. Thus, e.g. the sentence *The boy sees the girl* is structured along the action scheme, although it contains no element having the cognitive role of Agentive. The fact that an action constitutes the prototypical content of an assertion corresponds to the conditions in which language came into being, and also to those in which language acquisition normally starts. An event is primarily referred to by a verb, and one of the points in which the communicative function of language has been decisive for sentence structure is the central role of the verb in the sentence. The grammatical categories corresponding to modal, temporal and aspectual parameters of actions generally accumulate on the verb, and the valency of the verb determines the possible and the necessary ingredients of the sentence. The lexical units filling these valency slots display their own valency and, together with their complementations, they specify, step by step, the content of the sentence.

A further domain showing the impact of communicative factors may be clearly seen if the

sentence is described with due regard to its position in the context, including the anthropocentric layer of TFA (illustrated in Sect. 3), which is expressed by grammatical means and is semantically relevant, so that it should be described within grammar.

2.3. The levels of the system of language

Basing the description of the system of language on its partition into levels such as (underlying) syntax, morphemics, phonology and phonetics, we divide the relationship between the signifiant and the signifié, i.e. that of F. de Saussure's sign, into several steps. A morpheme is then regarded as a function of a morph (string of phonemes) and, at the same time, as having its own syntactic and semantic functions. The distinction between primary and secondary functions and markedness rules help in describing the relationships between units of adjacent levels.

The choice of (surface) subject is semantically relevant in some cases (e.g. with passivization, cf. Sect. 4.1), so that the concept of subject (as differing from Actor, or underlying subject) is necessary. The identification of subject may be handled as immediately concerning underlying (tectogrammatical) structure; the valency slot (argument) chosen as subject with passivization can be marked as such and, in languages that are similar to Czech in this point, during the transition to surface (morphemics) this argument either is changed into subject (esp. if with the active verb it would have the form of Accusative), or the passive is impersonal (e.g. *Jirka byl spatřen* 'Jirka was seen' or ...*obdivován* 'admired' vs. *O Jirkovi bylo hovořeno* 'Jirka was spoken about'). This view would not represent a support for the alleged necessity of a separate level of surface syntax. Surface word order can be understood to belong to the level of morphemics, where the representation of the sentence has the shape of a string without parentheses, rather than that of a tree (or even a more complex network, or its linearization). The immediate transition between (underlying) syntax and morphemics brings a possibility how to handle the issue of non-projective constructions. They are strictly limited and the question is how to account for them as for exceptions. They may be described as such by means of shallow rules changing the underlying projective order under certain specific conditions (see Sect. 4.2 below). Also word order shifts concerning a marked position of the intonation center can be handled by similar shallow rules, cf. Sect. 3.1.

The tectogrammatical representations (TRs) constitute a patterning of the cognitive content. While this patterning is language dependent in what concerns its repertoire of units (lexical items, grammatical categories and their values, valency slots), the content itself is not determined by an individual language.

2.4. Valency as the Core of Syntax

Not only complementations determined (subcategorized) by individual heads in the sense of configurationality or in a similar vein, but also the whole classes of adjuncts can be specified in the valency frames of lexical items. However, a classification of individual complementations as such (as types) into arguments and adjuncts has to be distinguished from the relationships of their tokens to their individual heads (for which they can be obligatory or optional). Thus, in our approach, Actor, Addressee, Objective, Origin and Effect, all of which are illustrated by the Czech ex. (1), are understood as arguments; the main criterion is that each of them can occur at most once with a head-verb token (if neither coordination nor apposition is present), be they obligatory or optional with a given verb.

(1) Matka j¡ zmØnila £Ÿes z copu na chlapeckì.
Lit.: Mother her changed hairdo from braid into boy's.
Actor Addressee Objective Origin Effect
[Mother changed her hairdo from a braid into bobbed hair.]

The identification of subject may be handled as immediately concerning underlying (tectogrammatical) structure; the valency slot (argument) chosen as subject with passivization can be marked as such and, in languages that are similar to Czech in this point, during the transition to surface (morphemics) this argument either is changed into subject (esp. if with the active verb it would have the form of Accusative), or the passive is impersonal (e.g. Jirka byl spatýen 'Jirka was seen' or ...obdivov n 'admired' vs. O Jirkovi bylo hovoýeno 'Jirka was spoken about').

Adjuncts in some (marked) cases are obligatory with individual head words, such as the complementation of Directional-to with to arrive, that of Manner with to behave, that of Appurtenance with brother, and so on. Thus there are four possibilities, three of which have to be specified (by indices identical for small groups of head words), whereas the fourth one (that of optional adjuncts) can be handled uniformly for a whole word class (with a single index common to the lexical entries of the class). A tentative (by far not complete) list of adjuncts can be found in Sgall et al. (1986, Chapter 2).

Surface deletion of certain complementations which are present (or even obligatory) in TRs may occur if the speaker assumes that the specific context makes them easily recoverable for the hearer. A suitable operational criterion can be seen in Panevov 's (1974) 'dialogue test'.

The following examples of valency frames (cf. Sect. 3 below for their more complete description) illustrate our classification (with the subscript o for obligatory items).

change V Acto Objo Or Effo glass N Material
give V Acto Addro Objo man N
rain V full A Materialo
brother N Appurto green A

Function words do not have specific positions in sentence structure; except for certain peripheral cases, articles and prepositions are always connected with nouns, auxiliary verbs and conjunctions with verbs, and they cannot be freely modified. Thus their counterparts in the TRs can be handled as parts of complex symbols (of node labels), i.e. as indices of lexical symbols themselves; these indices denote the values of categories such as definiteness, number, tense, modality, and the syntactic relations, be they expressed by morphs (endings, affixes, function words), alternations, or by word order.

2.5. Dependency trees

The basic pattern of sentence structure can have the form of a dependency tree, cf. a simplified model of the preferred TR of sentence (2) in Fig. 1, with morphological indices (grammatemes) of Preterite, Declarative, Singular, Definite, and syntactic symbols (labels of edges) for Actor, Locative, Appurtenance, Restrictive adjunct, Identity:

(2) My friend Jim works in the domain of archeology.

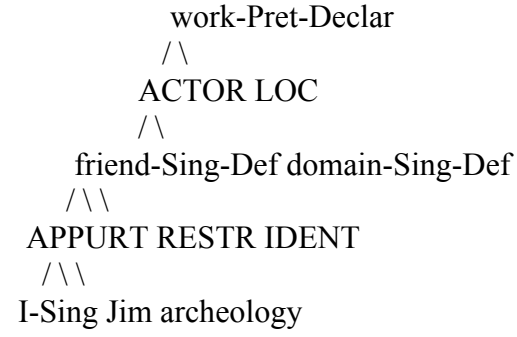

Fig. 1.

The complex labels of the nodes of our trees indicate (a) lexical meanings (which should be denoted by abstract symbols reflecting their inner lexico-semantic composition, but are just substituted here by the graphemic shape of words), and (b) values of morphological categories such as tense, aspect, number, etc. The labels of edges indicate the valency slots or kinds of the dependency relation (Actor, Addressee, Objective, Means, Locative, etc.).

2.6. Coordination and projectivity

Along with dependency, the syntactic representations include a specification of several further relations. One of these is TFA, expressed mainly by an interplay of word order and sentence prosody (esp. the position of intonation centre); in our trees it is represented by the left-to-right order of the nodes with the boundary line between topic, standing to the left, and focus, to the right of this boundary line.

Other kinds of syntactic relations are those of coordination (conjunction, disjunction and others) and of apposition. Their interplay with dependency cannot be accounted for with full adequacy by trees; more than two dimensions are needed. However, it is important that the relationships of the different dimensions are strongly restricted by such conditions as that of projectivity (adjacency) and similar restrictions holding for the relationships between coordination and the basic two dimensions of the tree. Thanks to these restrictions, the representations can be handled by limited means; they can be denoted by a linearized version of the network, namely by a string of complex symbols with every dependent node being included into a pair of parentheses and every string of items connected by a relation of coordination or apposition having such a pair. The valency slots and the kinds of coordination are written as indices of parentheses. The possibility to use such a framework to describe most different combinations of the two kinds of relations can be illustrated by ex. (3), where (3)(b) is a simplified TR of the sentence (3)(a).

(3)(a) Ann and Jim, Martin's brother, who are a nice pair, moved from a town to a village.

(b) ((Ann (Jim (Martin)Appurt brother.Def.Sing)Appos
)Conj (Descr (Rel)Actor be.Pres.Declar
(Obj pair.Specif.Sing (Restr nice))))Actor

move.f.Pret.Declar (Dir.1 town.Specif.Sing)
(Dir.3 village.Specif.Sing)

The syntactic symbols for Actor, Appurtenance, Objective, Descriptive and Restrictive adjunct, Directional.1 and .2 as kinds of dependency are written as subscripts to parentheses, indicating by their position at the left or the right parenthesis the direction from the dependent item to its head; others denote Conjunction (versus Disjunction, Apposition and other values) as a kind of Coordination, and Definite, Specifying, Singular, Present, Declarative, etc., as grammatemes; Rel denotes the (prototypical case of a) relative pronoun; the difference between the underlying and the surface word order positions of the noun pair and the adjective nice is due to factors discussed in Sect. 3.2.

On the morphemic level, the sentence representation has the form of a string of symbols (morphemes and their parts, the 'semes', such as cases, values of Tense, Gender, Number, Definiteness). Therefore, the condition of projectivity is absent on this level.

3. Topic-Focus Articulation (TFA)

3.1. TFA as one of the syntactic hierarchies

TFA can be systematically described with the help of the following ingredients of TRs:

(a) Introducing the primitive notion of 'contextual boundness', we understand as contextually bound (CB) those items that refer to entities supposed by the speaker to be easily accessible to the hearer and "spoken about", whereas the non-bound (NB) nodes contain "new information", or, at least, bring the established entitities into relationships new for the hearer. (b) Underlying word order (communicative dynamism, CD) is represented as the order of nodes in a TR and distinguished from the surface (morphemic) word order. CD is expressed by the position of intonation center (indicated by capitals in our examples, whenever it is not at the end of the sentence), by surface word order, and by further syntactic or morphemic means. Within such a framework, the definition of topic (T) and focus (F) can be anchored in the concepts of contextual boundness and of CD. The verb itself and any of its immediate dependents belong to T iff they are CB. If a node n belongs to T, so do all nodes dependent on it (with a specific proviso for cases where all these items are CB). The rest of the TR is the focus of the sentence. An outline of an automatic identification of Topic and Focus was presented by Hajičová et al. (1995). The TFA of a given sentence can be operationally tested by the so-called question test and by similar criteria.

In languages with a high degree of "free" word order, such as Slavic, Latin, and so on, the word order reflects CD more faithfully than in English or French; not only the opposition of T and F (with T prototypically preceding F), but also subtle differences concerning e.g. a contrastive (part of) T are then more accessible to analysis, cf. the following Czech examples:

(4) Karel Marii NEVID·L, Martin ANO.
Lit. Charles-Nom. Mary-Acc. not-saw, Martin-Nom. yes.
[Charles HASN'T seen Mary, (but) Martin HAS.]

(5) Marii Karel NEVID·L, Milenu ANO.
Lit. Mary-Acc. Charles-Nom. not-saw, Milena-Acc. yes.
[Mary HASN'T been seen by Charles, (but) Milena HAS been.]

(6) Tady studuje chemii TOMµæ a v BrnØ MARIE.

Lit. Here studies chemistry-Acc. T-Nom., and in Brno M-Nom.
[Here TOM studies chemistry, and in Brno MARY does.]

(7) Chemii tady studuje TOMµæ a fyziku MARIE.
Lit. Chemistry-Acc. here studies T-Nom. and physics M-Nom.
[Chemistry is studied by TOM here, and physics by MARY.]

Typically, the contrasted part of T (the most dynamic part of the sentence, T proper) gets the leftmost position.

3.2. Systemic ordering

The repertoire of the types of complementation (valency slots) displays a certain basic, or systemic ordering (SO), which, within the focus of a sentence (more exactly, within its NB parts) is reflected by its CD. Only if one or more of the complementations occur in a sentence as CB, then their position in CD can switch more to the left than what would correspond to SO. This can be illustrated by the following examples, with which the order B - A is only possible if B belongs to T (is CB), although with A - B in some readings A belongs to T and in others to F; we attach Czech versions of the (b) examples under (b').

(8)(a) They went (A) by car (B) to the RIVER.
(b) They went (B) to the river (A) BY CAR.
(b') Jeli (B) k ýece (A) AUTEM.
(9)(a) Ron cannot sleep (A) quietly (B) in a HOTEL.
(b) (B) In a hotel Ron cannot sleep (A) QUIETLY.
(b') Ron nem…§e (B) v hotelu (A) klidnØ SPµT.

Since in English the word order variation is limited, it happens here more often than in Slavic languages that the secondary situation (with B less dynamic than A) is expressed by a left dislocation as in (9)(b), by a specific syntactic construction as in (10)(b) or (11)(b), or by a marked position of the intonation center as in (12)(b); note that due to differences in English and Czech SO, in (11) and (12) the positions of A and B differ in the two languages.

(10)(a) (A) Dutch companies published (B) many books on LINGUISTICS.
(b) (B) Many books on linguistics were published (A) by Dutch COMPANIES.
(b') (B) Mnoho knih o lingvistice vydala (A) nizozemsk NAKLADATELSTVÖ.
(11)(a) Jim made (A) a canoe (B) out of a LOG.
(b) Jim made (B) a log (A) into a CANOE.
(b') Jim udØlal (B) k noi (A) z KLµDY.
(12)(a) Jim dug (A) a ditch (B) with a HOE.
(b) Jim dug (A) a DITCH (B) with a hoe.
(b') Jim kopal (A) motykou (B) STROUHU.

These isues have been analyzed by means of several series of psycholinguistic tests (for Czech, German, and partly also for English, see Sgall et al. 1995), with the result that SO differs from one language to the other. It appears that for some of the main complementations of English the scale of SO is as follows:

Temp - Actor - Addressee - Objective - Origin (Source) - - Effect - Manner - Dir.from - Means - Dir.to - Locative

Czech and other Slavic languages differ from English in that in them most of the adverbial complementations precede Objective and Effect under SO. Also in German, Means probably precedes Objective, so that (12')(a) is ambiguous as for the position of Means (CB or NB), whereas (12')(b) lacks this ambiguity (the Objective always is CB here):

(12')(a) Jim hat mit einer Hacke eine RINNE gegraben.
(b) Jim hat eine Rinne mit einer HACKE gegraben.

3.3. The semantic relevance of TFA

TFA is not only a matter of contextual positions of sentences, of pragmatics, but is semantically relevant, even from the view point of truth conditions. This is true not only of sentences with such overt complex quantifiers as many and few, but also of other sentences, cf. (13), and (14) vs. (15):

(13)(a) I exercise in the mornings.
(b) In the mornings I exercise.
(14)(a) John saw an explosion.
(b) Mary saw an explosion.
(c) John and Mary saw an explosion.
(15)(a) An explosion was seen by John.
(b) An explosion was seen by Mary.
(c) An explosion was seen by John and Mary.
(d) An EXPLOSION was seen by John and Mary.

With (14)(c) John and Mary could have seen different explosions, which is not the case with (15)(c), at least on its preferred reading.

Thus, also the semantic interpretation immediately concerns TFA, not only in connection with focus sensitive operators (the scopes of which we studied in cooperation with B. H. Partee, see Hajičová et al. 1998) and with other differences in truth conditions proper (where the opposition of truth versus falsity is at play), but also in further issues. The question whether a definite noun group triggers a genuine presupposition or just a weaker kind of entailment (allegation, see Hajičová 1972; 1984) depends on whether the noun group is CB or not.

Since TFA is semantically relevant, it is appropriate to look for means how to represent TFA in the interface representations. In the prototypical case the word order in most different languages corresponds to the scale of CD, and thus it appears as optimal to have the word order in the TRs identical to this scale, with every left daughter being CB, and to mark the main verb as CB (belonging to T) or not.

3.4. Sentence and discourse

The pragmatically conditioned reference assignment plays a crucial role in the cohesion of a discourse, and it is then important to ensure the possibility to identify the specification of reference in correspondence with the speaker's intention. As pointed out by Hajičová et al.

(1982; 1995; 1998), it appears that in natural language such a mechanism is based on the degrees of salience of the items contained in what the speaker assumes to be the hearer's stock of information at the given time point.

Prototypically, the referent of an expression included in the focus of the preceding utterance is the most salient item. What was referred to in the topic of that utterance, is one degree less salient. In the presence of such a small difference in salience, a weak pronoun is not sufficient to disambiguate, but a strong pronoun is; thus (16) can be followed by (17) with it being ambiguous, whereas this can only refer to cloth (or to (16) as a whole).

(16) The table was covered by a green cloth.
(17) It/This was old and shabby.

If an activated item is not mentioned in the next utterance, its salience is reduced more than by one degree, so that a reference by a weak pronoun is less probable. It has also been found that if an object is referred to by a CB item in an utterance, then the subsequent reduction of its salience is slower than in case of objects referred to only by NB items.

As soon as more than the structure of a single sentence is taken into account in MT, a possible strategy would be to work with a simplified representation of the degrees of salience, perhaps in the shape of a 'focus list' or stack. Among the referents listed after (or even during) the occurrence of a sentence S, there should be (a) the just mentioned objects with those used as NB in S in the highest positions, (b) objects or entitites that are 'at hand' for the hearers due to their domain specific or situation conditioned layers of knowledge, and (c) indexical and other elements that always (at least in the given culture) can be understood as being 'at hand' (with us, in Europe, Newton, and so on).

4. Status of Underlying Structure

4.1. Two levels of syntax, or syntax and morphemics?

The underlying structure, constituting the patterning of the cognitive content as reflected by a particular language, can be characterized as the level of linguistic (literal) meaning. In several linguistic approaches, a level of surface syntax is used side by side with (or instead of) such an interface level. However, it is not certain whether strictly synonymous syntactic constructions exist, i.e. whether it is appropriate to distinguish between such two levels. Let us reconsider some of the relevant examples:

(18)(a) After he arrived, we started to discuss this.
(b) After his arrival we started to discuss this.
(19)(a) She allowed them to sing.
(b) She allowed them that they sing.
(20) After his arrival we will start to discuss this.

In (18)(b) the after-group does not express tense; the fact that the time point of arrival precedes that of the utterance is only inferred from the combination of the meanings of after and of (the Preterite in) started. On the other hand, e.g. in (20) such an inference cannot take place and the temporal relationship then remains indistinct; i.e., (20) corresponds semantically either to After he arrives, we will start... or to (Now,) after he has arrived, we will start...

The synonymy is dubious also with (19), where in (a) the coreference between them and

the deleted subject of the infinitive is grammatical (intrasentential, a case of 'control'), whereas in (b) the coreference is textual, expressed by a pronoun the reference of which, in general, is indistinct, so that (as for restrictions determined by grammar) the referential identity is not ensured, cf. (21), where in (b) them refers to Mary and Paul, whereas they refers to their children:

(21)(a) She allowed Mary and Paul that their children sing.
(b) (Mary and Paul asked the hostess if their children may sing during the break.) She allowed them that they sing.

Passivization is another case in which synonymy is absent. In languages in which (in contrast to English) an active and a passive sentence can display the same word order (so that often they share their TFA), synonymy of the two verb forms might be looked for; however, if a sentence contains such an adverbial as inadvertently, or with great pleasure, then again the TRs of the active and the passive sentences do differ, due to the specific relationship of this adverbial to the subject. The choice of subject is also relevant in what concerns the control relation, since it is the subject of the infinitive (be it active or passive) that is the controllee. Therefore, the opposition between active and passive, i.e. the choice of subject has to be reflected in the TRs and does not require the distinction of underlying and surface syntax (cf. Sgall 1992 and Sect. 2.3 above).

Moreover, if surface word order is interpreted as belonging to the level of morphemics, where the sentence representation is a string without parentheses, then an immediate transition from the underlying level to that of morphemics is possible. This offers a possibility to cope with the difficult issue of seemingly non-projective constructions. They are strictly limited and may be described as such by means of shallow rules reflecting the difference between the two layers of word order that is present under certain specific conditions. Thus, e.g. sentence (22) can be derived by a rule bringing the heavy relative clause to the end of the sentence in a similar way as prepositions are brought to the beginning of their nominal groups and conjunctions to that of their verb's groups in (23). The conjunction is derived from the index that characterizes the verb want as occupying the position of (the head of) an adverbial of Cause.

(22) Mary saw a friend there, who told her the nice news.
(23) Jim visited Claire, since he wanted to ask her for advice.

Word order shifts concerning a marked position of the intonation center can also be handled by similar shallow rules, to which we turn in Sect. 4.2; cf. the difference between the primary morphemic shape of (8)(a) above and of (24); in the latter example, the to-group belongs to T in all readings of the sentence, similarly to (8)(b).

(24) They went by CAR to the river.

4.2. Syntax and morphemics

If the surface word order is to be derived from CD by movement rules, at least the following issues have to be accounted for, cf. Hajičová (1991), Sgall (1997).

(a) In sentences with a secondary placement of the intonation center, cf. (12b) and (24) above, the righmost item of the TR may be transferred to another position and marked as the bearer of the intonation center of the sentence (which then, in phonemics, gets a falling

stress).

(b) The grammatically determined positions of verbs, adjectives, clitics, and so on, are to be reflected (the adjective primarily is more dynamic than its head noun, but it is placed to its left in English or Czech surface word order). In Czech, clitics usually are placed in the "second" (Wackernagel's) position (with exceptions, similar e.g. to those concerning the 'Vorfeld' in German), even if they depend on an infinitive standing after the main verb; (cf. Fig. 2, in which we attach the syntactic indices to nodes; see point (d) below as for the seeming non-projectivity):

(25) Martin mi ji tu hodlal ukázat.
Martin to-me her here intended to-show
E. Martin intended to SHOW me her here.

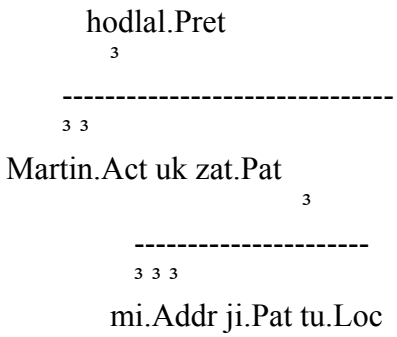

Fig. 2.
A simplified TR of sentence (25).

This may also concern other CB dependents of an infinitive:

(26) Martin mi Milenu zamìçlel uk zat. Martin to-me Milena-Acc intended to-show
E. Martin intended to SHOW me Milena.

Wackernagel's position is just after the leftmost node of the uppermost part of the tree (and, if this node differs from the verb, also after all nodes subordinated to the leftmost one).

(c) Function words are usually placed at the beginning of their word groups in the morphemic string; the rules describing these movements are to be combined with changing the indices in the complex node labels into symbols of articles, prepositions, etc. (cf. ex. (23) above).

(d) Also other cases of apparent non-projectivity (i.e. peripheral, strongly limited cases) can be described by movement rules concerning morphemics:

(27) I met a man yesterday, who looked for your ADDRESS.
The order in the TR: I - yesterday - met - a man, who... (see Fig. 3).

(28) a larger town than Boston:
The order in the TR: larger - than Boston - town.

Similarly as with (25) and (26) above, the TR then has a projective shape; the seeming

deviations from projectivity are reflected by the differences between CD (order of the nodes in the TR) and the surface word order. Specific problems are connected with long distance dependencies, with which e.g. a wh- item can land at the beginning of a clause to the verb of which it is subordinated.

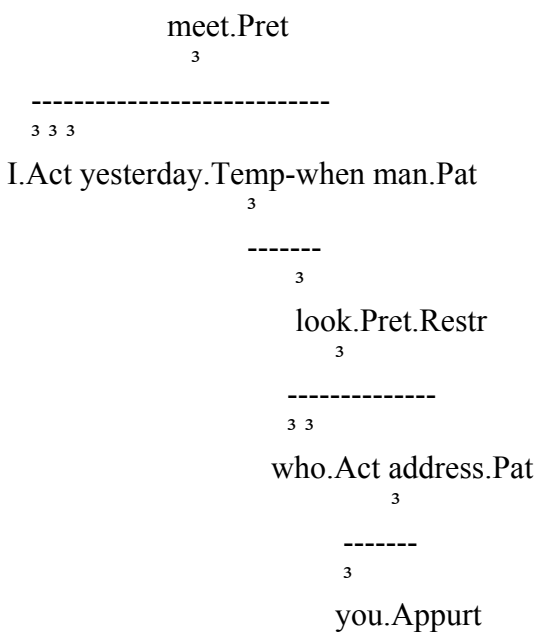

Fig.3.
A simplified TR of sentence (27).

The transition from surface (morphemics) to (underlying) syntax can be handled on the basis of the following considerations:

In the unmarked case, morphemes (endings, prepositions, other function words) express dependency relations and values of morphological categories (number, definiteness, tense, aspect, modalities, degrees of comparison), whereas surface (morphemic) word order expresses the scale of CD. Prototypically these two orders are identical, the differences are relatively rare and can be described by a limited number of movement rules. In certain non-prototypical cases, word order expresses syntactic (dependency) relations; however, even in case of a highly 'configurational' language (as English with the order Actor, verb and Objective), with most dependency relations it is necessary to take morphemes into account (prepositions in English, also case endings in most Continental languages, suffixes in the agglutinating languages of Asia and so on).

The assignment of underlying (i.e. CD) positions to function words during parsing is relatively simple if their underlying counterparts are handled as indices within the complex node labels of the corresponding autosemantic units. To cope with the other differences between the surface word order and the scale of CD (such as marking the underlying counterpart of the bearer of the intonation center as the most dynamic part of the TR, or that of an adjective as more dynamic than its governing noun, etc.), we can use numerical indices for the underlying positions (cf. Starosta 1993). It would require a systematic investigation to see whether such an approach allows for a natural account not only for the subtle relationships between surface word order, CD and SO, but also for those concerning the different positions of quantifiers or of 'focalizers', cf. Hajičová et al. (1998).

The core of the transition from morphemic strings to syntactic representations (preliminary formulations of implemented procedures for which can be found with Kirschner 1987; Hajičová et al. 1995) then is to describe the transition from endings and function words to their functions. Thus, Nominative and Accusative express primarily the Actor and Patient, respectively, Genitive typically depends on the adjacent noun, Dative expresses the Addressee, a prepositional group usually expresses an adverbial (dependent on the verb), prepositions either express certain kinds of dependency relation (e.g. E. with: Accompaniment or Means, from: Directional.1, through: Directional.3), or they have more subtle semantic functions (values of 'syntactic grammatemes'), as is typical of those concerning location and direction, see Table 1 with Czech examples and with DIR2 standing for 'through which place', DIR3 for 'where to' and DIR1 for 'from where'.

LOC	DIR2	DIR3	DIR1	EXAMPLES
na+L	pýes+Acc	na+Acc	z/s+Gen	na zdi 'on a wall', na stole 'on a table'
v+L	skrz+Acc	do+Gen	z+Gen	v koçi 'in a basket'
	Inst	zd¡		'through a wall'
u+Gen		pod,l+Gen	k+Dat	od+Gen k lesu 'to a forest'
nad+Inst	nad+Inst	nad+Acc	nad	t¡m 'over it-Loc'
	pýes+Acc	pýes	nØj	'across it'
pod+Inst	pod+Inst	pod+Ak	pod	t¡m/to 'under it'
pýed+Inst	pýed+Acc	pýed		t¡m/to 'before it'
za+Inst	za+Acc	za		t¡m/to 'behind it'
mezi+Inst	mezi+Acc	z+Gen	mezi	nimi 'among/ /between them'

Table 1.

Another task is to proceed from surface word order (and, in speech analysis, from the position of the intonation center of the sentence, and of phrasal stress) to the CD scale. The cases discussed above under (a) - (d) require specific movement rules to be included into the parser.

Furthermore, it is necessary to restore the deleted items (in coordinated structures, in case of the deletion of an obligatory valency slot, and so on). It is not easy to draw the boundary line between those cases in which the deleted lexical unit is to be inserted as such (e.g. with blue [flag] and yellow flag), and cases in which just an anaphoric or indexical item is appropriate (e.g. with George has arrived on Sunday either here or there is present in the TR, i.e. the sentence is ambiguous in this point).

4.3. Lexical and Grammatical Information

The dependency-based approach allows for the valency frames and lists of free complementations to specify the sentence structure on the basis of properties of individual words and word classes (the latter being determined e.g. by means of indices in the individual lexical entries).

In our approach, a lexical entry contains the following parts:

(a) the underlying representation of the lexical unit itself, i.e. of its lexical meaning; (b)

a specification of the morphological categories (grammatemes) relevant for the given word class (e.g. number and definiteness for nouns, or tense, aspect, different kinds of modalities, etc., for verbs, degrees of comparison for adjectives); restrictions on the combinations of the categories are listed for every word class as a whole, only exceptions being registered in individual lexical entries;

(c) the valency frame, the basis of which is the list of complementations that may (or must) depend on the given word (arguments, i.e. Actor, Addressee, Objective, Effect, Origin, and adjuncts, such as Locative, Instrument, Manner, Cause, etc.), ordered in accordance with SO, cf. Sect. 3.2 above; obligatory complementations are indicated by means of specific indices and it is denoted by an index whether a complementation is deletable with the given head (as e.g. DIR2 with to arrive) or not (as Objective with to create); also the optional or obligatory function of an item as controller is specified here (e.g. Actor is an obligatory controller with to try, an optional one with to decide; Addressee is an optional controller in the case of to advise, to forbid); indices characterize individual complementations as being able to occupy certain specific positions (e.g. that of subject, or of a wh-element);

(d) subcategorization conditions, e.g. the Objective of a verb possibly having (or not) the shape of a noun group, of a verb clause, etc.

5. Specification of Underlying Representations

The class of TRs can then be specified either by means of a generative procedure, or by a corresponding declarative definition, either of which can use a small number of general principles to describe the core of grammar.

The generative procedure for TRs is based on the following points:

(i) To generate a node n means

(a) to create the node n either as the root of a TR, or as a node that is dependent on another one and is placed to the right of all its sister nodes, and also

(b) to choose n's lexical value and the values of its grammatemes, taking into account the subcategorization conditions of the mother node of n and the restrictions on the combinations of grammatemes (specified in the lexical entry of the head or in the data concerning the respective word class); the technique used to realize these conditions and restrictions is unification; e.g. if n is the Objective of a verb that subcategorizes its Objective as a verb, then the lexical unit in the label of n has to be accompanied by the symbol identifying its word class as verb;

(c) if n is a root, the lexical part of its label is a verb, and its grammatemes determine it as a finite verb form of the main clause; n is then specified either as CB (i.e., as belonging to the topic) or, in the non-marked case, as NB (i.e., as belonging to the focus);

(ii) if the symbol of a complementation (argument or adjunct) is present in the frame of the node n, then it is possible to generate either a left or a right daughter of n;

(ii)(a) in case a left daughter is being generated, it is assigned a CB marker and a complementation value chosen from the frame of n;

(ii)(b) if a right daughter is being generated, it is assigned a NB marker and a complementation from the left end of the frame of n.

(iii) If the chosen complementation is an argument, it is deleted in the frame of n (as having been saturated).

NOTE: Choosing a complementation "from the left end" means that optional complementations can be skipped (it should be recalled that the complementations are

ordered in accordance with SO in the frames), and deleted in the frame of n; if the last one present there has been deleted, no more daughter nodes can be generated in this step, and point (iv) below is carried out. Analogously to the primary values of the grammatemes, the NB marker can be understood as primary, i.e. as the absence of a marker.

(iv) If no complementation is present in the frame of n, then the procedure goes back to the mother node of n, which now is to be considered as node n; if no mother node is present, the procedure is finished.

(v) Only representations that contain a focus are understood as TRs, more exactly, only those whose branch that contains only rightmost daughters starting from the root includes a NB node; this condition can be handled in such a way that the first occurrence of the NB marker (in a label of a node the mother of which has no more dynamic sister) is registered as saturating this point.

A declarative specification of underlying sentence representations can be formulated in accordance with these lines, using unification. The set of representations meeting the conditions specified in the lexical entries of the head words (including the order of the contextually non-bound complementations under SO) can be specified in this way. To this aim, the notion of unification has been enriched so as to allow to check the order of nodes and to make a distinction between saturated and non-saturated items (see PetkeviŸ 1995). The deletion of a saturated argument, mentioned in the Note above, ensures that an argument occurs at most once in a clause. This restriction does not concern adjuncts, cf. the three Temporal adverbials in Yesterday she came late to the office in the morning).

As was mentioned above, this specification of TRs covers only the core of sentence syntax. It has to be completed in several respects, especially in what concerns coordinated structures (corresponding to a third dimension of the network) and the position of such specific items as the operator of negation and other focalizers (only, even, also, etc., cf. Hajičová et al. 1998).

6. Concluding remark

Our experience with illustrating FGD and checking its principles supports the view that the classification of syntactic relations, valency slots, and grammatemes, as well as the view of TFA discussed here can be seen as highly appropriate for languages of different types. If this is so, the TRs may serve as a basis for the syntactic patterning of the output of the parsers of different languages if we want to make the transfer procedures in multilingual translation as simple as possible.

Acknowledgement:

* The research underlying this paper was carried out in the frame of the project VS-96151.

References:

Hajičová Eva (1972): Some remarks on presuppositions. Prague Bulletin of Mathematical Linguistics 17:11-23; printed in F. Papp and G. Szépe (eds.): *Papers in Computational Linguistics*. Budapest 1976:189-197.

Hajičová Eva (1984): Presupposition and allegation revisited. Journal of Pragmatics 8:155-

167; amplified as On presupposition and allegation. In: Sgall (1984:99-122).

Hajičová Eva (1991): "Free" word order described without unnecessary complexity. *Theoretical Linguistics* 17:99-1O6.

Hajičová Eva, Partee Barbara H. and Petr Sgall (1998): *Topic-focus articulation, tripartite structures, and semantic content*. Dordrecht:Kluwer.

Hajičová Eva, Petr Sgall and Hana Skoumalová (1995): An automatic procedure for topic-focus identification. *Computational Linguistics* 21:81-94.

Hajičová Eva and Jarka Vrbov (1982). On the role of the hierarchy of activation in the process of natural language understanding. In: *COLING* 82. Ed. by J. Horecký. Amsterdam: North Holland, 107-113.

Kirschner Zdeněk (1987): APAC3-2: An English-to-Czech machine translation system. In: *Explizite Beschreibung der Sprache und automatische Textbearbeitung* XIII. Prague: Faculty of Mathematics and Physics, Charles University.

Panevová Jarmila (1974). On verbal frames in Functional Generative Description. Prague Bulletin of Mathematical Linguistics 22:3-40, 23(1975):17-52; a revised version in Prague Studies in Mathematical Linguistics 6, 1978, 227-254.

Petkevič Vladimír (1995): A new formal specification of underlying structures. Theoretical Linguistics 21:7-61.

Plátek Martin, Sgall Jiří and Petr Sgall (1984): A dependency base for a linguistic description. In Sgall (1984:63-97).

Plátek Martin and Petr Sgall (1978): A scale of context sensitive languages: Applications to natural language. *Information and Control* 38:1-20.

Sgall Petr (1967): Functional sentence perspective in a generative description. *Prague Studies in Mathematical Linguistics* 2:203-225.

Sgall Petr, ed. (1984): *Contributions to functional syntax, semantics and language comprehension.* Amsterdam/Philadelphia:Benjamins,

Sgall Petr (1992): Underlying structure of sentences and its relations to semantics. *Wiener Slawistischer Almanach*. Sonderband 33. Ed. by T. Reuther. Vienna: Gesellschaft zur Förderung slawistischer Studien, 273-282.

Sgall Petr (1997): On the usefulness of movement rules. In: Caron B. (ed.), Actes du 16e Congr†s International des Linguistes (Paris 20-25 juillet 1997), Oxford: Elsevier Sciences.

Sgall Petr, Pfeiffer Oskar E., Dressler Wolfgang U. and Michael Půček (1995): Experimental research on Systemic Ordering. *Theoretical Linguistics* 21:197-239.

Sgall Petr, Hajičová Eva and Jarmila Panevová (1986): *The meaning of the sentence in its semantic and pragmatic aspects*, ed. by J. L. Mey, Dordrecht:Reidel - Prague:Academia.

Starosta Stan (1993): Word order and focus in constrained dependency grammar. In: Hajičová E. (ed.), Functional description of language. Prague: Charles University, 237-252.