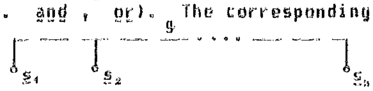
NEW DEPENDENCY BASED SPECIFICATION OF UNDERLYING REPRESENTATIONS OF SENTENCES

Vladimír PETKEVIČ
VÚMS Research Institute
Lužná 2
Prague 6, Czechoslovakia

In this lecture a pushdown store generator of deep, underlying structures (abbrev. US) of sentences is defined. This generator (a pushdown store generative grammar) developed inside the functional generative description (FGD) of language in Prague is the first part (from the viewpoint of the generation of a sentence) of the whole stratificational FGD: the output of the generator (a generated US) is transduced to the lower levels of FGD in the direction from function to form (from meaning, generally, to its representation) in order to achieve the final phonetic (graphemic) representation of the sentence. Our framework comprises the three basic dimensions of the semantics of the sentence: a) valency frames (theta-roles, types of the dependency relation), b) coordination and apposition, and c) topic-focus articulation. The generative grammar reflects the interplay of the mentioned dimensions endeavouring to simulate more closely the process of the formulation of the sentence by a real speaker.

1. In this lecture we present a new generative procedure generating underlying structures of sentences, i.e. semantic representations of meanings, within the framework of FGD of language developed by the Prague research group /Sgall et al. 1969/. This stratificational description divides the relation between the meaning and its phonetical (graphemic) expression into 5 levels: tectogrammatics (level of US's, abbrev. TR), surface syntax (SR), morphemics (MR), phonemics and phonetics (graphemics). Each of the levels is interpreted as a set of representations each of which fully describes a sentence on the pertaining level. The generative component (GC) produces the TR representation of a sentence first which is further transduced to SR, then from SR to MR and so on until the phonetical /graphemic) representation of the sentence is achieved.
Our GC is a pushdown store generator G /Sgall 1980/ generating "complex dependency structures" /Plátek et al. 1984/, i.e. linearized labelled graphs incorporating the following semantic dimensions: dependency relation (DR), coordination and apposition constructions (CA) and topic-focus articulation (TFA). Each semantic word form (semanteme, node) of a generated US depends (except for the main verb) by a semantic DR (as a complementation, modification, expansion) either on a single node or on a group of coordinated nodes. The dependent node is enclosed in < and >n where integer n encodes the type of DR, e.g. Actor, Objective, Time, Place (see Table I), and < stands between the governor and its dependent node which, in the corresponding graph, would be connected by an edge labelled n . The relation of CA is denoted by:
[s_1 ; s_2 ; ... s_n]q where s_i form a CA group, q encodes a CA type (e.g. and , or). The corresponding graph looks like:



As to TFA, each node is considered either as contextually bound (CB) or non-bound (NB), the CB (NB) node is marked by the superscript t (f). In the graph the CB (NB) nodes depend on their governor from the left (right). The CB (NB) property of a node x is related only to x itself, while the topic and focus of the sentence are global terms /Sgall et al. 1986, Hajičová 1980/. Prototypically, CB (NB) nodes belong to the topic (focus), although CB nodes can also belong to the topic and NB nodes to the topic.
Any node generated by G is a complex terminal symbol of the shape (a^z,GR_a), where a stands for a lexical (meaning) unit belonging to a certain word class (CL_a), e.g. Verb, Noun, etc., z ∈ {t,f} , and GR_a is a subset of grammatemes (=values of such categories as definiteness, number and some preposisitons with nouns, or tense, aspect and modality with verbs) appropriate for CL_a. An output string of G reflects the dependency structure of the corresponding sentence: the governor always precedes its dependent daughter nodes, the CB nodes preceding the NB ones on the same level of embedding. Those individual daughter nodes are ordered according to increasing communicative dynamism: CB sister nodes always being less dynamic than NB ones, each CB (NB) daughter node being considered less (more) dynamic than its governor, NB nodes being subject to the "systemic ordering" (see Definition below). Such an approach reflects the degree of salience or activation in the stock of knowledge the speaker is assumed to share with the hearer during the discourse /Sgall 1986/.

Example. The meaning of the sentence
(1) Jane and my brother, who created a family, live in Boston.
is generated by G as follows:
<(live^f,GR_{live})<[(Jane^t,GR_{Jane});(brother^t,$GR_{brother}$)
<(I^t,GR_I)>35]and<(create^f,GR_{create})<(who^t,GR_{who})>4
<(family^f,GR_{family})>26)>1>4<(Boston^f,GR_{Boston}))>11>)
Here the node for 'create' depends on the CA group 'Jane and my brother'. The integers denote the types of DR:
1 = General relationship (noun adjunct), 4 = Actor, 11 = Place, 26 = Objective, 35 = Appurtenance.
2. We present now the formal definition of G.

Definition. The pushdown store grammar G is defined as follows:
G = (K,V_o,V_s,K_o,$\bar{K}$,F) , where:
V_o = A ∪ N' ∪ Q' ∪ {NEG_t,NEG_f} is an output vocabulary; here A is a set of complex terminal symbols (semantic word forms, semantemes, nodes) of the shape (a^z,GR_a), where a stands for a lexical (meaning) unit from a certain word class (CL), z ∈ {t,f} (i.e. CB or NB), GR_a ⊆ GR_{CL} , where GR_{CL} is a set interpreted as the set of grammatemes appropriate for the word class CL, where a ∈ CL;
N' is the set of symbols having one of the two shapes:
< , >n , where n ∈ N, N is the set of integers encoding the kinds of DR. Exactly, N = PT ∪ FN , where
PT = {4,23,24,26,28} is a set of inner participants, i.e. complementations none of which can expand a semanteme more than once (cf. Table I); FN is a set of free (adverbial) modifications; PT ∩ FN = ∅ . For any lexical unit a the sets PT_a (inner participants of a), FN_a (free, adverbial modifications of a), and OC_a (obligatory complementations of a) are distinguished, where OC_a ⊆ PT_a ∪ FN_a , PT_a ∩ FN_a = ∅ . As FN_a = FN_b for any a, b ∈ CL , the symbol FN_{CL} can be used.
Q' is the set of symbols having one of the following shapes: [,]q , where q ∈ Q , Q being the set of symbols denoting the variants of CA, Q = {cop(=and), disj(=or), adv(=but), contr(=while),ap} ; thus Q = Q_{it} ∪ Q_{nonit} , where Q_{it} = {cop, disj, ap} and Q_{nonit} = {adv, contr} denote the subsets of the CA variants that can and cannot be iterated, respectively;
NEG_t denotes the negation of the CB verb;
NEG_f denotes the negation of the (partial) focus.
K is a set of inner states of G, each state is composed of an ordered triple,
K = K' ∪ $\bar{K}$ ∪ {K_o,(t,W^1(0),∅),(∅,COORD,∅)} ; here
K' = {(K_1,K_2,K_3)}, where
K_1 ∈ {t,f} , K_1 denoting whether the node just being expanded is being expanded by CB (K_1=t) or NB (K_1=f) complementations;
K_2 is a nonterminal representation of the modified semanteme, i.e. a complex symbol denoting the word class of the modified semanteme (negation of a verb being denoted by the superscript 1 or 2) followed by the

properties of this semanteme expanded during the derivation of an output string, i.e.: $K_2 = CL^{\underline{x}}(GR, \underline{h}^{\underline{e}}, \underline{g}, \underline{cn})$.
Here

CL is the word class of the modified semanteme;

$\underline{x} \in \{0,1,2\}$, if $\underline{x} = 1$ then either a CB verb (i.e. CL = V - the word class is Verb) has already been negated or the impossibility of such a negation is indicated; if $\underline{x} = 2$ then the negation of (possibly partial) focus dependent on the verb (or including it) has been performed;

GR is a set of grammatemes pertaining to the modified node; $\underline{h} \in \{\underline{t}, \underline{f}\}$, $\underline{h}$ denoting whether the modified node itself represented by the complex symbol K_2 is CB or NB;

$\underline{e} \in \{0,1\}$, $\underline{e} = 1$ iff the modified semanteme has an embedded (possibly partial) focus;

$\underline{g} \in \{0\} \cup Q$, $\underline{g} \in Q$ iff the given complex symbol K_2 represents a CA of the type $\underline{g}$, whose members are just being derived or expanded; if $\underline{g} = 0$ then the value of the variable $\underline{cn}$ (see below) is meaningless, in this case we omit (in Table II) the zero value of $\underline{g}$ and the value of $\underline{cn}$;

$\underline{cn} \in \{0,1\}$, $\underline{cn}$ ensures that at least two members of CA will be derived;

K_3 denotes a sequence of elements belonging to the set $\bar{N} \cup \{0\}$, where $\bar{N} = \{\underline{n}^{\underline{j}}_{\underline{k}} ; \underline{n} \in N, \underline{j}, \underline{k} \in \{0,1\}$ and $\underline{j} = 1$ iff $\underline{n} \in OC_{\underline{a}}$ for the modified node $\underline{a}$, $\underline{k} = 1$ iff $\underline{n} \in PT_{\underline{a}}$ for the modified node $\underline{a}\}$;

$K_0 = (0,0,0) \in K$ is an initial state;

$\bar{K} = \{(\underline{t}, W^1(0^1), 0), (\underline{f}, W^1(0^1), 0)\} \subseteq K$ is a set of final states;

$(\underline{t}, W^1(0), 0)$ is the evasive (wrong) state;

$(0, COORD, 0)$ is an auxiliary state for the proper derivation of CA's;

$V_S = \bar{N} \cup V'_S \cup \{W\}$ is a pushdown store (PS) vocabulary, where elemeunts of V'_S have the same shape as K_2 enriched with the supersript standing with the word class symbol and denoting the DR type by which the corresponding semanteme uas expanded last;

W is an auxiliary symbol denoting the inaccessible end of PS.

If $\underline{n} \in N$ occurs as a subscript, then $\underline{n}$ may also equal 0.

G works along with the lexicon comprising entries each of which consists of:

a) representation of a lexical meaning $\underline{a}$ accompanied by its semantic and syntactic features;

b) elements of the shape $\underline{n}^{\underline{j}}_{\underline{k}}$ ($\in \bar{N}$), where $\underline{n} \in PT_{\underline{a}} \cup OC_{\underline{a}}$.

For each word class CL there are further attached to the lexicon:

i) a set of all possible grammatemes appropriate for CL;

ii) a set of free modifications appropriate for CL (denoted FM_{CL}) - they can be associated with the entire word class.

The set of all possible complementations of $\underline{a}$ consists of $PT_{\underline{a}} \cup FM_{CL}$. This set is ordered according to increasing primary communicative dynamism of its elements (this ordering being called "systemic ordering", abbrev. SO, see Table I) and the resulting sequence is called a case frame of $\underline{a}$. SO is valid for the NB complementations only, i.e. in a sentence all NB complementations of a node must be ordered according to SO, but no such ordering is defined for the CB complementations of a node. SO in Table I is specific for Czech /Sgall 1986/. Individual entries in the lexicon can be chosen by G e.g. by means of the random generation.

Other symbols used for the description of PS elements:

$\underline{v}$ is a variable for a verbal lexical unit;

$\underline{a}$ is a variable for a lexical unit;

$\underline{f}_{\underline{v}}$, $\underline{f}_{\underline{a}}$ denote the case frames of $\underline{v}$, $\underline{a}$.

The list of all complementations along with their assigned integer codes ($\in \bar{N}$) ordered according to SO is displayed in Table I.

Table I

Codes of complementations ordered according to the systemic ordering

1 - *General relationship* (*black* table, *two* men)
2 - *Identity* (the city of *London*)
3 - *Descriptive property* (*golden* Prague)
4_1 - *Actor* (*John* made it; John *slept*)
5 - *Time: when* (He did it *yesterday*)
6 - *Time: since when* (*Since his arrival* we have not been working)
7 - *Time: till when* (I was there *till Sunday*)
8 - *Time: how long* (It lasted *two hours*)
9 - *Time: for how long* (He will stay *for two weeks*)
10 - *Time: contemporariness* (He was reading *during the pause*)
11 - *Place* (He lived *in Paris*)
12 - *Manner* (He studies *well*)
13 - *Regard* (*Regardless of what you're doing*, I am always with you)
14 - *Extent* (He studies *intensively*)
15 - *Standard* (He wrote it *according to the rules*)
16 - *Substitution* (He was appointed President *instead of me*)
17 - *Accompaniment* (He went there *with her*)
18 - *Restriction* (All were rescued *except for him*)
19 - *Instrument* (He wrote *with a pen*)
20 - *Difference* (He is *two inches taller*)
21 - *Comparison* (He is better *than I*)
22 - *Direction: through which place* (He ran *through the bushes*)
23_1 - *Addressee* (He gave *him* a pen)
24_1 - *Origin* (It is made *out of wood*)
25 - *Direction: from where* (He crept *out of the tent*)
26_1 - *Patient* (*Goal, Objective*) (I saw *it*)
27 - *Direction: where to* (She penetrated *into the woods*)
28_1 - *Effect* (He made a log *into a canoe*)
29 - *Beneficiary* (He bought a flower *for her*)
30 - *Condition* (I will not leave *if you do not give me money*)
31 - *Aim* (He did it *in order to ...*)
32 - *Cause* (He smiled *for it was too ridiculous*)
33 - *Result* (He did it *so that I could be free*)
34 - *Concession* (*Although he was clever*, I ...)
35 - *Appurtenance* (a leg *of the table*)
36 - *Partitive* (a bunch *of flowers*)

The numbers having the subscript 1 belong to PT, others belong to FM.

PS is written in such a way that the leftmost symbol is that one stored at the accessible end of PS.

Before the definition of the defining function F of G the meaning of other symbols used is presented:

V (the word class of the lexical unit of $\underline{v}$) and N denote the word class of verbs and nouns, respectively;

A, D are variables for a word class;

Symbol C stands (symbols U, D' stand) for complex nonterminal symbols of the shape K_2, where $\underline{g} \in Q$ ($\underline{g} = 0$); i.e. C stands (U, D' stand) for the complex symbols which represent (do not represent) the CA's whose members are just being derived or expanded;

$\underline{i} \in \{\underline{t}, \underline{f}\}$;

$\underline{p} \in \{0,1\}$;

$\underline{s}$ denotes a sequence of elements of $\bar{N}$;

$\underline{m}$ stands for a symbol of the shape $\underline{n}$, iff on the left-hand side of the same rule the variable $\underline{k}$ has the value 0; otherwise $\underline{m}$ stands for an empty sequence;

$\underline{y}$ stands for 0 iff $\underline{h}^{\underline{e}}$ on the left-hand side of the same rule has the value $\underline{t}^0$, otherwise $\underline{y}$ stands for 1;

the prime and bar symbols (e.g. $\underline{e}'$, $\bar{\underline{e}}$) have a similar meaning as their simple counterparts (i.e. $\underline{e}$) (here it means that also $\underline{e}'$, $\bar{\underline{e}} \in \{0,1\}$).

If a superscript or a subscript of a variable has the value 0, it may be absent in the notation used.

F is the defining function of G. It has two parts: Table II and Limiting Conditions, i.e. conditions limiting the possibility of using individual rules as given in Table II. F consists of 14 (schemes of) rules denoted 1.A, 1.B, 2, 3, 4, 5, 6.A, 6.B, 6.C, 6.D, 7, 8.A, 8.B, 8.C. Each rule consists of the left and the right part. The left part consists of an (input) state (IS) and a PS symbol (RS). The right part consists of an (output) state OS, the sequence of PS symbols (WS) and the sequence of output symbols ($\underline{O}$). The functioning of G consists in a computation, i.e. in a sequence of steps in each of which a rule is applied. The rule $\underline{r}$ can be applied if during the computation of G the current state of G equals IS of $\underline{r}$

TABLE II

Num	R S	I S			O S			W S	O
		K_1	K_2	K_3	K_1	K_2	K_3		
1.A		Ø	Ø	Ø	t	V (GR,h)	Ø	r_v,W	< (v^h,GRv)
1.B		Ø	Ø	Ø	Ø	COORD	Ø	C,W	<[
2	n_k^j	t	V (GR,t)	Ø	t	V^1 (GR,t)	Ø	n_k^j	NEG_t
3	n_k^j	t	D^x (GR,h^e)	s	t	D^x (GR,h^e)	s,n_k^j		
4	U	t	D^x (GR,h^e)	s	f	D^x (GR,h^e)	Ø	s,U	
5	n_k	f	D^x (GR,h^e)	Ø	f	D^x (GR,h^e)	Ø		
6.A	n_k^j	i	D^p (GR,h^e)	s	t	A ($\overline{GR}$,i)	Ø	r_a,D',s,m	< (a^i,GR_a)
6.B	n_k^j	i	D^p (GR,h^e)	s	Ø	COORD	Ø	C,D',s,m	<[
6.C	C	Ø	COORD	Ø	t	A ($\overline{GR}$,i)	Ø	r_a,C	(a^i,GR_a)
6.D	C	Ø	COORD	Ø	Ø	COORD	Ø	$\bar{C}$, C	[
7	U	f	V^p (GR,h^e)	Ø	f	V^2 (GR,h^e)	Ø	U	NEG_f
8.A	U	f	D^x ($\overline{GR}$,$\bar{h}^{\bar{e}}$)	Ø	h	U^1	Ø		>n
8.B	C	f	D^x ($\overline{GR}$,$h^{\bar{e}}$)	Ø	Ø	COORD	Ø	C'	;
8.C	C	f	A^x ($\overline{GR}$,$\bar{h}^{\bar{e}}$)	Ø	t	A (GR,h^e)	Ø	r_a	]q

and simultaneously the current PS's symbol is RS. By applying $\underline{r}$, IS switches to the corresponding OS, RS being read (=removed) from the top of PS while WS is written onto the top of PS (in our notation of WS the leftmost symbol of WS becomes a new top symbol of PS) and $\underline{O}$ is written at the output tape (in the left-to-right direction). G starts in K_0, the rules are applied in an arbitrary order, the only condition being that the current state and the PS's top symbol agree with the left-hand side of the applied rule. In some cases a choice between the rules is possible (i.e. G is non-deterministic). If G reaches the state $\underline{s}$ to which no rule can be applied, then either $\underline{s} \in \bar{K}$, i.e. an US has been achieved on output, or $\underline{s} \notin \bar{K}$, i.e. the resulting string differs from a proper US in that it contains no (non-empty) focus.

Limiting Conditions

Ad 1.A: a. The set of grammatemes of the lexical unit $\underline{v}$ is $GR_{\underline{v}}$, which is a set of grammatemes appropriate for a main verb, $GR_{\underline{v}}$ = GR.

Ad 1.B: a. C = V(GR,$\underline{h}$,g,0) .
b. g ∈ Q.

Ad 2,3: None.

Ad 4: a. U = $A_{\underline{n}}$(GR',$\underline{h}$'$^{\underline{e}'}$) or U = W.

Ad 5: None. (Notice the absence of the superscript $\underline{j}$ over the symbol $\underline{n}$, i.e. $\underline{j}$ = 0.)

Ad 6.A: a. D' = $D_{\underline{n}}$(GR,$\underline{h}$) .
b. A,D,GR,$\overline{GR}$ meet specifically listed restrictions of subcategorizations and others (not discussed here).
c. The set of grammatemes of the lexical unit $\underline{a}$ is $GR_{\underline{a}}$, $\overline{GR}$ = $GR_{\underline{a}}$.

Ad 6.B: a. C = A($\overline{GR}$,$\underline{i}$,g,0) .
b. Conditions 6.A.a - 6.A.b hold.
c. g ∈ Q.

Ad 6.C: a. C = A(GR,$\underline{i}^{\underline{e}}$,g,$\underline{cn}$) .
b. g ∈ Q.
c. The set of grammatemes of the lexical unit $\underline{a}$ is $GR_{\underline{a}}$, $\overline{GR}$ = $GR_{\underline{a}}$.
d. Either $\overline{GR} \subseteq GR$,
or A = N, and the following conditions hold:
d1. $\underline{pl}$ ∈ GR .
d2. $\overline{GR}$ \ {$\underline{pl}$,$\underline{sg}$} ⊆ GR \ {$\underline{pl}$,$\underline{sg}$} .
($\overline{GR}$ is consistent with the word class A, as for the grammatemes appropriate for a given word class; $\underline{pl}$ and $\underline{sg}$ denote the grammatemes of plural and singular, respectively, a coordinated group of nouns is generated from the coordination group symbol marked as plural, because such a group has the syntactical distributional properties - agreement, etc. - of a plural noun.)

Ad 6.D: a. $\bar{C}$ = A($\overline{GR}$,$\underline{i}$,$\bar{g}$,0) ,
C = A(GR,$\underline{i}^{\underline{e}}$,g,$\underline{cn}$) .
b. g,$\bar{g}$ ∈ Q .
c. Condition 6.C.d holds.

Ad 7: a. U = $D_{\underline{n}}$(GR',$\underline{h}$'$^{\underline{e}'}$) .
b. If $\underline{h}$ = $\underline{t}$ then $\underline{e}$ = 1.

Ad 8.A: a. Either a1. U = $A_{\underline{n}}$(GR,$\underline{h}^{\underline{e}}$) ,
U^1 = A^1(GR,$\underline{h}^{\underline{y}}$);
or a2. U = W,
U^1 = W^1($0^{\underline{y}}$) .
b. $\underline{y}$ = 1 iff either $\bar{\underline{h}}$ = $\underline{f}$ or $\bar{\underline{e}}$ = 1;
otherwise $\underline{y}$ = 0.

Ad 8.B: a. C = D(GR,$\underline{h}^{\underline{e}}$,g,$\underline{cn}$) ,
C' = D(GR,$\underline{h}^{\underline{e}'}$,g,$\underline{cn}$') .
b. Either b1. $\underline{cn}$ = 0, $\underline{cn}$' = 1, g ∈ Q;
or b2. $\underline{cn}$' = $\underline{cn}$ = 1, g ∈ $Q_{\underline{it}}$.
c. If $\bar{\underline{e}}$ = 0 then $\underline{e}$' = 0; otherwise $\underline{e}$' = $\underline{e}$.

Ad 8.C: a. C = A(GR,$\underline{h}^{\underline{e}}$,g,$\underline{cn}$) .
b. $\underline{r}_{\underline{a}}$ is a case frame for a CA represented by the complex symbol C, i.e. an ascending sequence of elements of $\bar{H}$.
c. Condition 8.B.c holds.
d. $\underline{cn}$ = 1.

REFERENCES

HAJIČOVÁ, E. (1980), A Dependency Based Specification of Topic and Focus I: Background and Motivation. SMIL 1/2, pp. 93-109.

PLÁTEK, M. and SGALL, J. and SGALL, P. (1984), A Dependency Base for a Linguistic Description. pp. 63-97 in: Sgall, P. (Ed.), Contributions to Functional Syntax, Semantics, and Language Comprehension. Praha, Academia.

SGALL, P. and NEBESKÝ, L. and GORALČÍKOVÁ, A. and HAJIČOVÁ, E. (1969), A Functional Approach to Syntax in Generative Description of Language. New York.

SGALL, P. (1980), A Dependency Based Specification of Topic and Focus II: Formal Account. SMIL 1/2, pp. 110-140.

SGALL, P. and HAJIČOVÁ, E. and PANEVOVÁ, J. (1986), The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Prague, Academia, pp. 100-266.