COLING 82, J. Horecký (ed.)
North-Holland Publishing Company

RANDOM GENERATION OF CZECH SENTENCES

Jarmila Panevová

Department of Applied Mathematics
Faculty of Mathematics and Physics
Charles University
Prague
Czechoslovakia

The experiments testing the theoretical adequacy and the practical usefulness of the Functional Generative Description (FGD) are described. The FGD consists of a generative component, which, in the experimental version, has the shape of a context-free grammar combined with elements of dependency approach, and the other components having the form of pushdown store automata. The latter components have a transductive role, transducing the semantic (tectogrammatical) representations of sentences to the lower levels of the language system. The transduction is articulated into several steps corresponding more or less to the levels of language system (surface syntax, morphemics, morphophonemics, phonemics, or, as the case may be, graphemics) postulated in European structural linguistics. The theoretical and practical qualities of the system are evaluated.

1. The model of generative description called Functional Generative Description (FGD) was proposed early in the sixties (Sgall, 1964), and it is worked out from that time by the group of algebraic linguistics, Charles University, Prague. This description is being enriched and completed from the empirical point of view and from the point of view of its theoretical adequacy. To allow for a systematic elaboration of both of these aspects, FGD is tested on computers in the form of random generation of semantic representations of Czech sentences and of transducing them to their outer shape. The computer testing fulfils also another aim: FGD, beside being an appropriate framework for empirical study and theoretical description of language, can also be applied as a background for practical projects, such as synthesis for machine translation into Czech, synthesis for answers in question-answering systems etc.

2. The main features distinguishing FGD from most of other linguistic frameworks are: (i) in FGD there is no place where transformationrules are needed; (ii) the generative power is concentrated in its first component, generating underlying representations on the level of linguistic meaning representing a specific patterning of extralinguistic, ontological content , the set of generated strings surpassing only moderately the set of context-free languages (Plátek and Sgall, 1978); (iii) FGD is based on a dependency approach to syntax; (iv) a stratificational approach is used here, articulating the generation of sentences into several steps corresponding to particular language levels ordered from meaning to the outer shape

(level of tectogrammatics, of surface syntax, of morphemics, morphophonemics and at the end graphemics in our case, or the phonemic level in a theoretical description).

Describing the theoretical features of FGD it must be stressed that the description intself is being developed much more quickly than the system of programmes can reflect. The first component of FGD, generating tectogrammatical representations (TR´s), was reformulated from the shape of formalism corresponding to context-free phrase structure grammar with elements of dependency features (Sgall, 1967) into a pure dependency formalism using pushdown store automata and including also description of the topic/focus articulation of the sentence (see Hajičová and Sgall, 1980).

3. The component generating TR´s was implemented in the older form, as a context-free grammar. The first experiments with random generation are restricted to a relatively small lexicon: something about 300 "deep" lexical units; the number of units increases on the surface level, where also function words are present, as well as the units gained by means of "syntactic" derivation in Kuryłowicz (1936) sense (suffixation and prefixation serving for nominalizons, etc.). The enriched output lexicon consists of more than 1500 units. As for the grammatical phenomena concerning the different linguistic levels, we tried to make the system relatively complete even in the first stage of the experiments; the coordinated constructions, the pronouns of the 1st and 2nd person and some of the possible word order variants were omitted in this stage.

4. We concentrate our attention, first of all, on the transductive components. All the linguistically relevant semantic information is included in the TR´s, where we have to do with disambiguated representations, identical for all synonymous surface variants. This means that the transductive components describe the asymmetric dualism between a function and its forms (in the sense of Karcevskij, 1929, and the Prague School of Linguistics, cf. Vachek, 1964). The relation between form and function may be illustrated by the following examples, concerning the relations between a TR and the corresponding surface syntactic representations and between these and the morphemic ones: The participant actor may be expressed either by surface subject, or by an adverbial of actor (in passive constructions), by a possessive adjective, or a noun in genitive or instrumental (with nominalizations); the functor (Fillmore´s 'case')Instrument may be expressed by the morphemic case of instrumental, or by prepositional constructions na + locative, pomocí + genitive.

The mathematical apparatus used for the transduction components of FGD is a sequence of pushdown store automata, transducing the TR into the surface representations (dependency trees) and the latter into morphemic ones (strings); then follows a finite automaton transducing the representation into the graphemic output form (there are some differences in this respect between the theoretical description of language and the procedure serving for applied projects). On both levels of the structure of the sentence the dependency tree representing the structural order has as its root the predicate of the main clause, which is the only one node not dependent on any other node. Every node is labelled by a representation of a single (autosemantic) word form, having the shape of a complex symbol containing its syntactic, morphological and lexical parts, corresponding to the character of the particular level. The dependency tree preserves the condition of projectivity.

Each transduction of the representation of the structure of the sentence to the adjacent level needs a pair of automata. The condit-

ions constraining the transduction to the next level can be characterized as follows:
(a) In a given step only a single dependency syntagm (the governing word and its modification) is processed; one of the two steps adjusts the morphological features (called grammatemes, cf. Panevová, 1979; 1980) in accordance with certain properties of the other member of the syntagm; mostly the modifier changes according to the character of its governing word: e.g. the actor of a passive verb comes over into an adverbial, that of a nominalized verb (a surface noun) into an attribute: hosté přišli - příchod hostů (the guests arrived - the arrival of the guests).
(b) A single pass through the sentence (in the text-to-rule order) is sufficient for every transducer.
(c) The process of transduction is based on the governing unit being handled by every pushdown automaton earlier that its modifications (dependent words). We work with the new characteristics of the governing word when its modifications are being processed.

The main programme (the defining function) of every automaton is based on the fact that the root of every dependency tree is processed as first. (It should be noted that the linearized dependency tree is converted into the sequence "regens post rectum"; nevertheless, not only this structural order, but also the linear order, more or less directly corresponding to the surface (morphemic) word order, are preserved in it). Then the first member from the right depending on the root is read by the automaton I and compared with the root, i.e. modified according to its properties. If the last word form read has no modifiers, it may be printed on the output from the given automaton; if there is some modifier present, the governing word is placed into the pushdown store and this pair of word forms connected by the dependency relation is then compared and evaluated. This means that the matrices or tables described in detail in Panevová (1979) are involved, where the changes obligatory or optional for the given pair of word forms (syntagm) are determined. These tables form an inner part of the automaton, which cannot be separated from the work of the whole procedure; the empirical data determining the choice of the means (forms) for functions (meanings) are involved here. The word form processed in a given time point can be printed in the output only in such a point when all the subtrees dependent on it have already been printed. On the output of every pushdown transducer but the last (i.e. with the exception of the morphemic representation) we again receive an order of word forms adapted to the further processing of the root of the tree as the first node; the modifying (dependent) nodes are read from the right side.

5. We want to present here a short survey of the linguistic problems solved by the individual automata: The first pair of automata chooses the active and passive constructions, nominalization, infinitive construction or subordinate clause. Due to the optionality of some rules (synonymy between e.g. Snažil se , aby přišel včas- subordinate clause, Snažil se přijít včas - infinitive construction, Snažil se o včasný příchod - nominalized form of He attempted to come in time), in the cases where we have to do with the choice between several equivalent constructions a probability for particular possibilities has been added (determined in a quite preliminary way, the results of which are being checked in the course of the experiments).

The rules of choosing simple or prepositional cases, subordinating conjunctions, etc. are a matter of the transduction from surface syntax to morphemics. There also the morphemic units of number, verbal aspect, tense etc. are assigned; e.g. with the word forms characterized (by an index) as a "plurale tantum", any grammateme of

number is obligatorily converted to plural as a morphemic unit: nůžky (scissors), kalhoty (trousers). The rules of grammatical congruence are also applied here (congruence between adjectival adjuncts and their head noun, between subject and predicate, etc.).

For the formulation of such rules, of course, detailed empirical studies about contextual conditions influencing the choice of the particular expression for instance of such underlying units as Actor, Instrument, etc. are needed (Instrumental case - psát perem, prepositional phrase na + Loc - psát na psacím stroji, pomocí + Gen - přeložit pomocí slovníku, etc., i.e. write with a pen, write on a typewriter, translate by means of a dictionary, respectively, all correspond to Instrument).

The next step consists in the procedure of morphemic synthesis (see Weisheitelová, 1979).This procedure is adapted to the purposes of practical projects, so that a direct transition to graphemics is attempted at. Here the structural order is no more needed and the representation of word forms provided with information on the morphemes included can be submitted to the procedure of the combination of lexical stem (and, as the case may be, its alternations) with endings to create a correct sequence of Czech word forms corresponding to the meaning represented in the TR with which the whole procedure of transduction started.

6. Some tectogrammatical representations of Czech sentences that were already gained as a result of the functioning of the procedure of random generation at the computer EC 1040 can serve as illustrations. Most of them are correct from the grammatical point of view, though their meaningfulness can often be doubted; however, the constraints on the semantic compatibility of lexical units are - in our opinion - a matter reaching beyond the linguistic competence as such. The questions of the boundary between these semantic selection restrictions and the grammatical conditions on strict subcategorization are by far not clear; this can be illustrated by the following examples the underlying trees corresponding to which were derived by our system during the first experiments:

Něco chtělo mít Kladno.- Something wanted to have (the town of) Kladno.

Právě něco těšilo každého muže.- Exactly something pleased every man.

Máme rozvit svátek. - We have expanded a holiday.

Co byla paměť spodem? - What was the memory from the bottom?

Each of these sentences seems to be connected with specific empirical problems concerning the mentioned boundary, and thus also the boundary between the system of language and the domain of cognition. Needless to say, clearly acceptable sentences were derived, too, by our system, such as:

Každá žena vyráběla nůžky. - Every woman manufactured scissors.

Panovník měl být co nejkratší. - The sovereign should have been as short as possible.

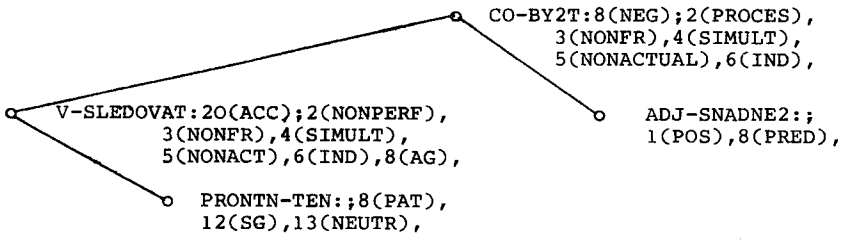
Musíme seznámit s kvalitou paměť. - We must make memory acquainted with quality.

Pán vyrobil listí. - The gentleman manufactured leaves.

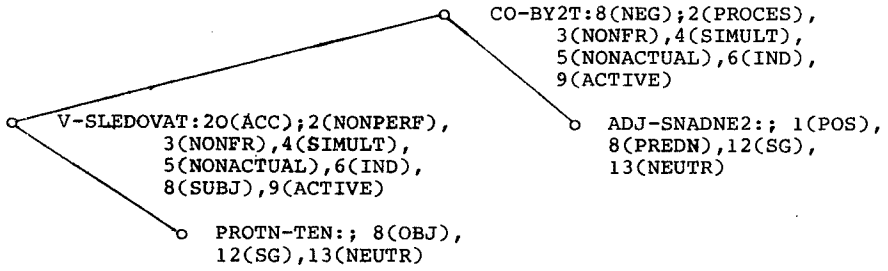
The system of programmes for the transductive components is extremely complex and due to this fact the linking of its partial programmes (procedures and subprocedures) for individual automata is a difficult task from the point of view of computer storage and of the human work concerning the debugging of the programmes. A recent sample of computer outputs will be demonstrated at the conference.

APPENDIX

The TR of the sentence "Sledovat to není snadné" (To keep track of it is not easy):



The syntactic surface representation of the same sentence:



The morphemic representation:

V-SLEDUJ:;2(NONPERF),3(NONFR),6(IND),9(ACTIVE),1O(INF),PRONTN-TEN:;
8(ACC),12(SG),13(NEUTR),CO-BY2T:;2(NONPERF),3(NONFR),4(PRES),5(NEG),
6(IND),9(ACTIVE),11(3PERS),ADJ-SNADNE2:;1(POS),8(NOM),12(SG),13(NEUTR)

In the Appendix we present a slightly simplified representation of a Czech sentence on the different levels. The sequence of data in the complex symbols functioning as labels of a single node is as follows: part of speech, lexical item, indices (between the signs ":" and ";"), grammatemes (after the ";" sign). The correspondence between function and its expression (form) on the adjacent level may be interpreted on the base of our example. The rules from which the correspondences between a function and its form(s) may be obtained on the basis of the main programme of a pushdown transducer, or of its subprocedures having the form of tables, were characterized in Panevová (1979; 1980).

REFERENCES:

[1] Hajičová, E. and Sgall, P., A dependency-based specification of topic and focus, SMIL - Journal of Linguistic Calculus 1-2 (1980) 93-109.

[2] Karcevskij, S., Du dualisme asymmétrique du signe linguistique, in: Travaux du Cercle linguistique de Prague 1(1929)88ff.

[3] Kuryłowicz, J., Dérivation lexicale et dérivation syntaxique, Bulletin de la Soc. ling. de Paris 37(1936) 79-92.

[4] Panevová, J., From tectogrammatics to morphemics, in: Explizite Beschreibung der Sprache und automatische Textbearbeitung 4(1979) 3-166.

[5] Panevová, J., Formy a funkce ve stavbě české věty [Forms and functions in the structure of Czech sentence] (Academia, Praha, 1980).

[6] Plátek, M. and Sgall, P., A scale of context-sensitive languages: Applications to natural language, Information and Control 38 (1978) 1-20.

[7] Sgall, P., Zur Frage der Ebenen im Sprachsystem, in: Travaux linguistiquesde Prague 1 (1964) 95-106.

[8] Sgall, P., Generativní popis jazyka a česká deklinace (Academia, Praha, 1967).

[9] Vachek, J., A Prague School reader in linguistics(Indiana University Press, Bloomington, 1964).

[10] Weisheitelová, J., Transducing components of functional generative description 2, in: Explizite Beschreibung der Sprache und automatische Textbearbeitung 5 (1979) 3-67.