Maciej Grochowski

NATURAL LANGUAGE DICTIONARY AND THE ARTIFICIAL METALANGUAGE IN GENERATIVE DESCRIPTION

Originally published as "Słownik języka naturalnego a sztuczny metajęzyk w opisie generatywnym," Studia Semiotyczne 12 (1982), 71–78. Translated by Agnieszka Ostaszewska.

1. 1. Practical value of each theory of language is decided above all by the capability of such theory to be used for description of particular natural languages. It is impossible to expect that a theory of language will be developed, which would have a universal explanatory power with respect to all facts, concerning all languages. Although numerous theories of language developed over the last twenty five years, belonging to the broadly understood generative trend[1] make it possible to explain a considerably greater number of linguistic facts than the structural theories, still it is impossible to speak of practical significance of the generative models, since no natural language has been described in a comprehensive and detailed manner with the use of a model of such kind.

What is especially convincing are the ideas of generative linguistics (shared by most of its representatives, more general, however, than the language description models represented by each of them)[2], and in particular the postulates that generative grammar is to model the language competence of an ideal sender-recipient, that it should be formulated explicitly and that it should be of prognostic (predictive) character. One of the initial theses of generativism, that there exists an analogy between the description of

[1]Cf. e.g. Chomsky, 1957; 1965; Katz, J.A. Fodor, 1963; Chafe, 1971; Katz, 1972a; Bartsch, Vennemann, 1972; Melčuk; 1974; J.D. Fodor, 1977.

[2]Cf. e.g. Chomsky, 1965; Katz, 1972a.

language and the description of a finite set of organized operations aimed at production of a certain type of products seems to be fully correct. It is also impossible to reject the hypothetico-deductive method approved of by the generativism. Even if one embraces the mere doctrine of generativism, it does not mean that one is forced to acknowledge the adequacy of any of the proposed generative description models, even less so that one is to accept the prevalent conceptual apparatus or the "generating" technique.

Generative linguistics assumes, as it follows from the analysis of but a few models, that it is possible to describe a natural language with the help of an artificial one[3]. Therefore, one assumes as given a system of nonterminal symbols (i.a. consisting of such variables as: S, NP, VP, N, A, Adv), which in accordance with the rules of substitution, represent various subsets of terminal symbols, i.e. natural language expressions. Sets of terminal symbols and the collection of generative rules are also considered to be given. Linguistic competence modelling is aimed at determining how to apply rules to a set of symbols, so as to make it possible to generate an infinite number of sentences of a given natural language[4].

One of the most important assumptions adopted by construction of generative models, namely that the natural language dictionary is a set of elements given a priori, is difficult to accept. If one rejects such hypothesis as being too strong, it is impossible to determine in advance, what set of artificial language symbols is necessary and at the same time sufficient, to represent the full set of expressions of a given natural language.

1.2. One of the objectives of this paper is to justify the thesis that a natural language dictionary is not a fully a priori given set of elements. This issue will be discussed in the light of syntax and semantic facts of the contemporary Polish language.

A thesis formulated in such manner immediately results in another issue, i.e. whether it is possible to assume a priori a specific artificial language dictionary, which would be adequate for description of a given natural language, and above all, to what extent such dictionary of artificial metalanguage is at all necessary for determination of the rules of generation of sentences in the natural language. An attempt at answering this question constitutes the principal objective of this article.

[3]Cf. footnote 1

[4]Pazuchin (1979) seems to be right to observe that the hypothesis that generative rules are adequate for an indefinite number of sentences is too strong. In his opinion it is impossible to prove such hypothesis, since it is possible to verify it only on the example of a finite number of sentences.

In accordance with the reductionism principle[5], any and all entities which are not necessary in science, are at the same time dispensable. Reduction of the number of auxiliary notions by formulation of a hypothesis is a course of action exposed to a smaller degree of falsity than multiplying the notions without any particular need[6].

It is also assumed that every natural language contains such signs, which are fully comprehensible for all native speakers of this language, and therefore do not need to be explained. One rejects however the assumption that there are commonly comprehensible artificial language signs, whose explication would be absolutely redundant.

It is assumed that sentence generation rules constitute an idealisation of certain actual human actions on objects in the form of language units. The generating rules should be of the character of ordinary practical rules, which would make it possible for a human being to master a language as a foreign language. If one assumed that generative grammar has nothing to do with the practical conduct of a language user, as it is often claimed by the theoreticians of generativism[7], then the answer to the question whether it is possible to present a description of a natural language with the use of a generative model would have to be "no". A description of a language competence obviously does not include a description of such empirical phenomena, as streaks of occlusal movements of the speaker, or the psychological processes occurring in their brain, nonetheless competence modelling is based on the samples of language performance[8].

It is also assumed that description of the rules of sentence generation should meet the requirements of maximum precision and maximum simplicity. Generative models, in particular those which are an attempt at description of semantic relations between expressions[9], have the flaw of being too general,

[5]Cf. Okham, 1971.

[6]One should also adopt Popper's (1977: 219) methodological postulate of reduction of the number of axioms in science to a system of axioms of maximum universality.

[7]Cf. e.g. Chomsky, 1965; J. D. Fodor, 1977.

[8]Bartsch and Vennemann (1972: 9) accuse Chomsky's theory of being internally inconsistent, claiming that since his model is to be a language competence model, and the competence is in his opinion an inborn feature of the human mind, then it should also have some sort of psychological reality. Chomsky however rejects the feasibility of the competence, maintaining at the same time that it is inborn. Also Wierzbicka (1977) sees Chomsky's refusal to acknowledge the feasibility of competence as resignation from his initial intentions, namely, that a formal theory of language would make it possible to reveal how the human mind operates.

[9]Cf. e.g. the articles in the collection edited by Nawrocka-Fisiak (1976), as well as the works of Dowty (1972) and Ross (1972).

and therefore, the description is anything but precise.

2. In order to determine the semaxntic and syntactic rules of joining lexical units, one needs to dispose of a given set of such units. Items in the dictionaries of natural languages (e.g. in The Dictionary of Polish Language ed. W Doroszyński), only for practical and technical reasons (i.e. for one to be able to find a given item quickly) are identified with graphical words, i.e. streaks of diacritic elements occurring between two subsequent pauses. One cannot however assume a priori that shapes distinguished solely on the basic of graphic criteria have some semantic features. In other words, graphical boundaries between the words do not correspond in any regular way to the boundaries between meaningful units. Therefore, it is usually necessary to distinguish streaks of diacritic elements longer than those resulting from orthographical conventions, and first such streaks – proper lexical units – should be ascribed meaning. Lexical units are therefore such semantically indivisible sets of diacritic elements, which can be automatically reproduced as ready formulas by the speakers in the course of generating a text.

If a hypothetical lexical unit consists of at least two graphical words (segments), then one may find whether such unit is actually indivisible into two units, by making attempt at a general semantic characterisation of the substitutive classes, to which the given words belong. Such an attempt is bound to fail, if the graphical words constitute a closed class, i.e. if they are possible to enumerate only[10].

In the Polish language it is impossible, for example, to determine any such regularities, that verbs from semantic class A imply co-occurrence of preposition x, or that verbs from semantic class B imply co-occurrence of preposition y. Therefore, it is only possible to enumerate the verbs requiring a given preposition. The verbs together with the prepositions constitute inseparable lexical units, cf. the examples of the verb units containing the preposition do [to; up to; as far as; at, etc.][11]: ktoś celuje do kogoś [someone is aiming at someone], ktoś dodzwonił się do kogoś [somebody has reach someone over the phone], ktoś przyzwyczaja się do czegoś [someone gets used to something], ktoś skłania się do czegoś [someone feels inclined to do something], ktoś szykuje się do czegoś [someone is preparing for something],

[10]The presented hypothesis on semantic indivisibility of multi-segment lexical units is based on the theory of language units formulated by Bogusławski (1976a, 1978a, 1978b)

[11]Translator's note: as the entire paper is based on the analysis of the phenomena of the Polish language, the translator did not attempt to provide a similar discussion of the English language, limiting herself to providing a translation of the examples in Polish [in square brackets].

ktoś tęskni do kogoś [someone misses someone], ktoś zapisał się do czegoś [someone signed up to something], ktoś zaprowadził kogoś do kogoś [someone has shown the way to someone to someone][12].

Similarly, adpositional phrases containing more than one preposition should also be considered to be indivisible lexical units, e.g.: bez względu na (coś) [irrespective of (something)], na domiar (czegoś) [in addition to (something)], na mocy (czegoś) [by virtue of (something)], na przekór (czemuś) [in defiance of (something)], na skutek (czegoś) [as a result of something], w miarę (czegoś) [in the course of (something)], w razie (czegoś) [in case of (something)], w związku z (czymś) [in connection with (something)], z uwagi na coś [in view of (something)], as well as conjunctions, e.g.: chyba, że [unless], chyba, żeby [unless], dlatego, że [for this reason], dopóty..., dopóki... [as long as], ilekroć..., tylekroć... [whenever], im..., tym... [the more... the more], mimo, że [despite], podczas gdy [whereas], w miarę jak [while], wprawdzie..., ale [although], zarówno..., jak [both..., as well as...].

An attempt at determination of the rules according to which one would be able to add negation in the Polish language (and in particular the segment nie [no, not]) to the lexical units from various classes, would require that one first distinguishes all such units, in which nie is one of the graphical segments. The following expressions belong to the class of such units: bynajmniej nie [by no means], ktoś nie ma za grosz czegoś [someone has no something], nie tylko..., ale także [not only..., but also...], o mało nie [nearly], omal nie [nearly], zgoła nie [no whatsoever]. The fact that negation is a part of the abovementioned units does not follow from the application of any rules; such units cannot be juxtaposed with sequences without negation. Cf. the abovementioned expressions with the following oppositions: czyta – nie czyta [he is reading – he is not reading], ładny – niełady [pretty – not pretty], wczoraj – nie wczoraj (ale przedwczoraj) [yesterday – not yesterday (but two days ago)], pies – nie pies (lecz kot) [a dog – not a dog (but a cat)].

There is no material difference between the presented examples of multisegment lexical units and the expressions described in Polish philology linguistic literature as phrasemes (idiomatic phrases). Cf. the above examples with such phrasemes as: ktoś dał komuś kosza [someone gave someone the mitten], ktoś rzuca słowa na wiatr [someone speaks idly], ktoś trzyma język za zębami [someone keeps his/her mouth shut], ktoś zmieszał kogoś z błotem [someone hauls someone over the coals], ktoś zrobił kogoś w konia [someone

[12]Expressions ktoś, do kogoś (someone, to someone), are characterised by the valence properties of the listed units. These features are an inseparable part of the description of the units.

fooled someone]. The only difference between one-segment lexical units (graphical words) and the phrasemes is merely of external character. The difference is irrelevant both from the semantic, as well as from the syntactic point of view, and pertains to the continuity of those units: a phraseme is a discontinuous unit, i.e. a sequence of diacritic elements containing pauses.

If one accepts the provided justification of the thesis that lexical units in the Polish language in a prevailing majority of cases do not correspond to the graphical words of this language, then one also needs to assume that a set of lexical units of the Polish language is not fully given. It is necessary to successively discover new units and to register the units already known, although the latter are not always fully realised by the scholars. Without having this task fulfilled, it will not be possible to determine a full set of rules of generating sentences.

3. Classes of lexical units may be described, in particular from the semantic point of view, only in a very approximate manner. All the more that the rules of joining units from different classes may be formulated only with the use of a considerable simplification. With respect to classes of units one may obviously use the symbols of an artificial meta-language, however as long as no analysis of the classes is carried out, there will be no guarantee, that indeed a given class constitutes a not closed set of units, and therefore, that it is subject to general semantic description. If it turned out that description of a class may only consists in enumeration of the units, then consequently, one should ascribe different symbols to particular units. Such enterprise, boiling down to multiplying artificial symbols, would be inconsistent with the explanatory function of science. Therefore, use of artificial meta-language symbols is possible only in instances where there is a considerable probability of a generalised description of the class of the units.

Below are examples of two general rules of generating purpose-communicating sentences in the Polish language:

1. xnv1 + żeby [so that] vinf // ynvpraet;

2. xnv2 + v1inf.

Particular symbols have the following meaning: x – personal noun, y – noun, n – nominativus, v – verb, v1 – action-naming verb, v2 motion-naming verb, inf – infinitivus, praet – praeteritum.

The set introduced by żeby is limited by the following formal limitations ((a) rule): if the expression represented by v refers at the same time to the

expression represented by xn, i.e. if the v and v1 reference are identical, then v has the form of infinitivus and the noun in nominativus is used once only (it is not repeated). If v and v1 do not have the same reference, i.e. if v refers to an expression represented by yn, then v is in the past tense form. If v is used in the first or the second person (both singular or plural), then the morphemes -m, -ś, -śmy, -ście, are added to the conjunction żeby. The schemata provided are written down in a simplified form, i.e. for one-argument predicates. On the basis of the same rules it is possible to construct sentences with more than one-argument predicates.

Below please find sentences exemplifying the application of these rules[13]:

1. Jan wybiegł z mieszkania, żeby zdążyć na pociąg. (John run out of the apartment to make it for the train.) Matka przyniosła kożuch, żeby Piotr nie zmarzł. (Mother brought a coat for Peter not to get cold.) Prokurator wezwał cię, żebyś złożył wyjaśnienia. (The prosecutor summoned you so that you would provide explanations.)

2. Jan skoczył do wody ratować tonącego. (John jumped into the water to save the drowning man.) Jan idzie do biblioteki wypożyczyć książkę. (John is going to the library to borrow a book.)

Although it is justified to assume that the classes of verbs, both the action-naming and motion-naming verbs, can be described generally and not by enumeration of particular units, nonetheless, without an explanation of the notions of action and motion, it will be impossible to provide such a description. All the more that the symbols used for recording the proposed rules are only of auxiliary character, but without an explanation of their meanings they would be nothing but empty shapes.

4. The proposed postulate to minimise the dictionary of the artificial meta-language should be considered in the light of certain conventions of syntactic description, and in particular of the semantic description, popularised in the generative linguistics.

Non-terminal symbols representing syntactic construction classes (e.g. NP, VP) and classes of parts of speech (e.g. N, V) are useful for designation of connection of units or one-segment units[14]. It is impossible to describe multi-segment lexical units in detail, either in terms of the construction, or

[13]These rules are discussed in more detail in: Grochowski, 1980.

[14]Uselessness of the initial generative description symbol of a sentence (S) was pointed out by Chafe (1971: 47). A sentence may be represented with the use of a predicative element together with all of its implied positions.

in the terms of the categories of the parts of speech, cf. e.g. ktoś dał nura [someone dived], ktoś plecie trzy po trzy [someone is talking nonsense], ktoś nie ma za grosz czegoś [someone has no something]. A claim that a given unit may be described with the use of the functional logic terminology, and then assuming that this unit is an n-argument predicate, is an admissible generalisation for all units with the exception of index expressions.

One of the basic hypotheses formulated by the generative semantics, namely that surface structures are generated from deep structures [15], is impossible to accept. The idea of lexical decomposition resulting from this hypothesis, according to which syntactic structures with a given lexeme should be described with the use of abstract atom predicates (written down with the use of capital letters), result in solutions characterised by arbitrariness, imprecision and lack of simplicity. E.g. the common semantic interpretation of the unit kill as derived from CAUSE TO DIE (or CAUSE BECOME NOT ALIVE) is erroneous, since there is no equivalence between the derivative base and this unit. The meaning of the base, unlike the meaning of the unit, does not imply that the persons (the killer and the killed) meet in time and space, and the meaning of the unit does not imply the component of "cause" or the awareness of the perpetrator of the action, which on the other hand follows from the derivative base[16]. Similarly e.g. a sentence On otwiera drzwi [He opens the door] cannot be derived from the base On robi coś z drzwiami w tym celu, żeby były otwarte [He does something to the door so that it would be open], since only the base implies the will and the awareness of the acting person[17].

The mere idea of generating sentences from a semantic base cannot be confirmed by empirical facts[18]. Hypotheses which surface structures have the same deep structure can be justified only on the basis of the semantic analysis of the surface structure. There is no possibility to know the semantic structures deprived of any shape, and shaping such as sequences of artificial language symbols[19] brings an inevitable question as to the meaning of these symbols. Answering this question one cannot escape from

[15]Cf. footnote 9, as well as e.g. Chomsky, 1997: 393

[16]A critical analysis of the descriptions of the expression kill (in the works of generative semantics) is presented i.a. in: Bartsch, Vennemann, 1972, Wierzbicka, 1975; J. D. Fodor 1977.

[17]Cf. more on the topic in: Bogusławski, 1974: 48.

[18]Cf. elaboration on the topic i.a. in: Bogusławski, 1976b: 11.

[19]Artificial semantic meta-language expressions are e.g. Katz's and Fodor's marks and distinguishers, abstract atom predicates of generative semanticists, Melčuk's semantic graphs, Melčuk's and Apresjan's lexical functions.

the use of natural language expressions. Katz (1972: 162) stated that the elements of the artificial language are unambiguous, and the ambiguity of the natural language expressions makes it impossible to use them in semantic interpretation. If one finds correct the view that the artificial language expressions are understandable first after they have been translated into natural language expressions, then one needs to reject Katz's view on the non-ambiguity of artificial symbols. Moreover, there is no basis to ascribe the feature of ambiguity to all natural languages expressions. Separation of multi-segment lexical units makes it possible to eliminate the ambiguity of graphic words to a considerable extent[20].

5. In the light of the above deliberations, the basic question of this paper concerning the usefulness of the artificial meta-language for generative description of natural language should be answered in the following manner:

There is no need to use an artificial meta-language in the process of determination of the rules of generating sentences. It is admissible to use the artificial meta-language symbols only in such cases, when the existence of generalisation can be fully justified. Reaching authentic (and not apparent and proximate) generalisations of semantic facts of a given natural language is indeed a very difficult path, since it requires a full register and full semantic analysis of the lexical units of such natural language.

Bibliography

1. Bartsch R., Vennemann T. (1972), Semantic Structures. A Study in the Relation Between Semantics and Syntax, Frankfurt/Main: Athenäum Verlag.

2. Bogusławski A. (1974), "Preliminaries for Semantic-Syntactic Description of Basic Predicate Expressions With Special Reference to Polish Verbs", O predykacji, ed. A. Orzechowska, R. Laskowski, Wrocław, 39-57.

3. Bogusławski A. (1976a), "O zasadach rejestracji jednostek języka", Poradnik językowy, 8: 356-364.

4. Bogusławski A. (1976b), "Segmenty, operacje, kategorie a morfologia imienia polskiego", Kategorie gramatyczne grup immiennych w języku polskim, ed. R. Laskowski, Wrocław, 7-42.

[20]Cf. more on the topic in: Grochowski, 1981.

5. Bogusławski A. (1978a), "Towards an Operational Grammar", Studia semiotyczne, VIII: 29-90.

6. Bogusławski A. (1978b), "Jednostki języka, a produkty językowe. Problem tzw. orzeczeń peryfrastycznych", Z zagadnień słownictwa współczesnego języka polskiego, ed. M. Szymvzak, Wrocław, 17-30.

7. Chafe W. L. (1971), Meaning and the Structure of Language, Chicago-London.

8. Chomsky N. (1957), Syntactic Structures, The Hague: Mouton.

9. Chomsky N. (1965), Aspects of the Theory of Syntax, MIT Press, Cambridge Mass.

10. Chomsky N. (1977), "Lingwistyka a filozofia" transl. U. Niklas, in Lingwistyka a filozofia. Współczensny spór o filozoficzne założenia teorii języka, ed. B. Stanosz, Warszawa, 391-437.

11. Dowty D. R. (1972), "On the Syntax and Semantics of the Atomic Predicate CAUSE", in Papers from the Eighth Regional Meeting of the Chicago Linguistic Society, Chicago, 62-74.

12. Fodor J. D. (1977), Semantics: Theories of Meaning in Generative Grammar, The Harvester Press, Hassocks, Sussex.

13. Grochowski M. (1980), Pojęcie celu. Studia semantyczne, Wrocław.

14. Grochowski M. (1981), "Problem rozróżnienia znaczeń czasowników polisemicznych", Studia z filologii rosyjskiej i słowiańskiej (in print).

15. Katz J. J. (1972a), Semantic Theory, New York.

16. Katz J. J. (1972b), Linguistic Philosophy. The Underlying Reality of Language and its Philosophical Import, London.

17. Katz J. J., Fodor J. A. (1963), "The Structure of a Semantic Theory", Language, 39: 170-210.

18. Melčuk I. A. (1974), "Opyt tieorii lingvističeskich modelej "smysł ↔tekst"". Semantika, sintaksis, Moskva.

19. Nawrocka-Fisiak J. (ed.) (1976), Readings in Generative Semantics, Poznań.

20. Ockham W. (1971), Suma logiczna, Warszawa.

21. Pazuchin R. (1979), "O uzasadnieniu teorii lingwistycznych za pomocą modeli", Studia semiotyczne, IX: 145-168.

22. Popper K. R. (1977), Logika odkrycia naukowego, Warszawa.

23. Ross J. R. (1972), "Act", Semantics of Natural Language, ed. D. Davidson, G. Harman, Dordrecht-Boston, 70-126.

24. Wierzbicka A. (1975), "Why "Kill" Does Not Mean "Cause to Die": The Semantics of Action Sentences", Foundations of Language, XIII: 491-528.

25. Wierzbicka A. (1977), Syntax vs. Semantics (typescript).