Blažej Kříž Generalized grammatical categories in the sense of Kunze

Archivum Mathematicum, Vol. 17 (1981), No. 3, 151--158

Persistent URL: http://dml.cz/dmlcz/107105

Terms of use:

© Masaryk University, 1981

Institute of Mathematics of the Academy of Sciences of the Czech Republic provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these *Terms of use*.



This paper has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library* http://project.dml.cz

ARCH. MATH. 3, SCRIPTA FAC. SCI. NAT. UJEP BRUNENSIS XVII: 151-158, 1981

GENERALIZED GRAMMATICAL CATEGORIES IN THE SENSE OF KUNZE

BLAŽEJ KŘÍŽ, Brno (Received March 10, 1980)

0. INTRODUCTION

General linguistics studies the so-called morphological categories. Let us take as an example from the Czech language the category "neuter noun in genitive of singular". This category consists of word-forms whose examples are: města, moře, stavení.

In addition to morphological categories, general linguistics deals with so-called syntactical categories, too. An example taken from English is "noun phrase". The syntactical categories are node descriptions in the phrase marker of the sentence. Analysis of these notions has been done in algebraic linguistics. To be able to explain it let us mention some basic notions of algebraic linguistics.

A formal language is an ordered pair (V, L), where V is a finite set and L is a set containing some finite sequences or, by another name, strings consisting of elements of the set V. We have $L \subseteq V^*$ if we denote by V^* the set of all such strings. The associative operation of concatenation is defined on the set V^* . We denote by xy the result of concatenation of the strings x, y.

A context-free grammar is an ordered quadruple $G = (V_T, V_N, x_0, F)$, where V_T, V_N are disjoint finite sets of terminals and nonterminals, respectively, $x_0 \in V_N$, and F is a finite set of ordered pairs (p, q) - called rules - such that $p \in V_N$ and $q \in (V_T \cup V_N)^*$. We write $x \Rightarrow y$ for $x, y \in (V_T \cup V_N)^*$ if there are $(u, v) \in V^* \times V^*$ and $(p, q) \in F$ such that x = upv, uqv = y. Let $\stackrel{*}{\Rightarrow}$ be the reflexive transitive closure of the relation \Rightarrow on $(V_T \cup V_N)^*$. Then the set $\mathscr{L}(G) = \{w \in V_T^* \mid x_0 \stackrel{*}{\Rightarrow} w\}$ is the language generated by the grammar G. A regular grammar is a special case of a context-free grammar. A context-free grammar (V_T, V_N, x_0, F) is called regular, if for every $(p, q) \in F$ either $q \in V_T V_N$ or $q = \lambda$ holds, where λ denotes the empty string.

Grammatical categories were introduced for formal languages in several ways. The simplest way is the following: Let (V, L) be a formal language. For every $X \subseteq V$, we define $\psi(X)$ as the set of all $x \in V$ such that $uxv \in L$ for every $(u, v) \in V^* \times V^*$ with the property $utv \in L$ for every $t \in X$. The set $\psi(X)$ is called the grammatical category generated by the set X. In a special case, when we take V as the set of all word-forms in a fragment of a natural language and L as the set of all correct sentences of this fragment, sets of the form $\psi(X)$ for $X \subseteq V$ correspond to morphological categories of this fragment, if the set X is chosen suitably. See Kunze [4], [5], Novotný [6].

In general, the assumption is accepted that a fragment of a natural language can be generated by a context-free grammar in such a way that nonterminals correspond to syntactical categories. If a context-free grammar (V_T, V_N, x_0, F) is given, we can assign the set $\{w \in V_T^* \mid n \stackrel{*}{\Rightarrow} w\}$ to each nonterminal $n \in V_N$. It is quite natural to precise syntactical category n as the corresponding set of strings. If nstands for "noun phrase", then the mentioned set consists of all noun phrases that are substrings of correct sentences in our fragment. We see that it is possible to transfer the notion of syntactical category to every language generated by a contextfree grammar. Thus, we shall use rather the notion of syntactical category of a contextfree grammar than the notion of syntactical category of the generated language.

We now are in the situation that we have defined grammatical categories for every formal language and also syntactical categories for every formal language generated by a context-free grammar. The substantial difference is the fact that a grammatical category is a set of symbols and a syntactical category is a set of strings of these symbols. To avoid this difference, we generalize the notion of grammatical category admitting sets of strings as grammatical categories. The complete characterization of regular languages is formulated here by means of the generalized grammatical categories. A natural problem arises here: Which languages can be generated by context-free grammars in such a way that their syntactical categories are grammatical ones? We prove that all regular languages satisfy this condition. A regular grammar is constructed to every regular language in such a way that its syntactical categories are the generalized grammatical categories of the language and that the grammar generates the given language.

1. GENERALIZED GRAMMATICAL CATEGORIES IN THE SENSE OF KUNZE

1.1. Definition: Let W be a set. A mapping ψ from 2^{W} into 2^{W} is called a *closure* operator on 2^{W} , if the following three conditions are satisfied for arbitrary X, $Y \subseteq W$: 1. $\psi(X) \supseteq X$.

 $1. \ \psi(X) \equiv X.$ $2. \ \psi(\psi(X)) = \psi(X).$

3. $X \subseteq Y$ implies $\psi(X) \subseteq \psi(Y)$.

A set X with the property $X = \psi(X)$ is called ψ -closed or shortly closed.

To construct grammatical categories in the sense of Kunze, we shall need the notion of Calois connection:

1.2. Definition: Let S and T be sets, σ a mapping from 2^s into 2^T and τ a mapping from 2^T into 2^s . We say that the ordered pair of mappings (σ , τ) establishes a Galois connection between the sets 2^s and 2^T , if σ and τ satisfy the following conditions.

(a) $X_1 \subseteq X_2$ implies $\sigma(X_1) \supseteq \sigma(X_2)$ for arbitrary $X_1, X_2 \subseteq S$.

(b) $Y_1 \subseteq Y_2$ implies $\tau(Y_1) \supseteq \tau(Y_2)$ for arbitrary $Y_1, Y_2 \subseteq T$.

(c) $X \subseteq \tau \sigma(X)$ for every set $X \subseteq S$.

(d) $Y \subseteq \sigma\tau(Y)$ for every set $Y \subseteq T$.

For Galois connections the following theorems hold.

1.3. Theorem: If the ordered pair of mappings (σ, τ) establishes a Galois connection between the sets 2^s and 2^τ , then $\tau\sigma$ is a closure operator on the set 2^s and $\sigma\tau$ is a closure operator on the set 2^τ .

Proof: See [8], Theorem 16.

1.4. Theorem: Let the ordered pair of mappings (σ, τ) establish a Galois connection between the sets 2^{S} and 2^{T} . Then $\sigma(X) = \sigma(\tau\sigma(X))$ holds for any $X \subseteq S$, and $\tau(Y) = \tau(\sigma\tau(Y))$ holds for any $Y \subseteq T$.

Proof: See [1], page 89.

We now establish the notions of the grammatical category in the sense of Kunze and the generalized grammatical category in the sense of Kunze.

1.5. Definition: Let (V, L) be a language. We say that an ordered pair (u, v) from the Cartesian product $V^* \times V^*$ is a context of an element $a \in V$ in the language (V, L) and write $(a, (u, v)) \in \varrho$ if and only if $uav \in L$ holds. We define a pair of mappings σ from 2^{V} into $2^{V^* \times V^*}$ and τ from $2^{V^* \times V^*}$ into 2^{V} in the following way.

$$\sigma(X) = \{(u, v) \in V^* \times V^* \mid (x, (u, v)) \in \varrho \text{ for every } x \in X\},\$$

where $X \subseteq V$.

 $\tau(Y) = \{x \in V \mid (x, (u, v)) \in \varrho \text{ for every } (u, v) \in Y\},\$

where $Y \subseteq V^* \times V^*$.

It is easy to see that the ordered pair of mappings (σ, τ) is a Galois connection between the sets 2^{ν} and $2^{\nu^* \times \nu^*}$.

1.6. Definition: We put $\psi(X) = \tau \sigma(X)$ for every $X \subseteq V$. Then ψ is a closure operator on 2^{V} . The set $\psi(X)$ is called the grammatical category in the sense of Kunze generated by the set X.

We generalize the definition of grammatical category.

1.7. Definition: Let (V, L) be a language. We say that an ordered pair $(u, v) \in V^* \times V^*$ is a context of the string $x \in V^*$ in the language (V, L) and write $(x, (u, v)) \in e_*$ if and only if $uxv \in L$ holds. We define a pair of mappings σ_* from 2^{V^*} into $2^{V^* \times V^*}$ and τ_* from $2^{V^* \times V^*}$ into 2^{V^*} as follows:

 $\sigma_*(X) = \{(u, v) \in V^* \times V^* \mid (x, (u, v)) \in \varrho_* \text{ for every } x \in X\},\$

153

where $X \subseteq V^*$.

 $\tau_{*}(Y) = \{ x \in V^{*} \mid (x, (u, v)) \in \varrho_{*} \text{ for every } (u, v) \in Y \},\$

where $Y \subseteq V^* \times V^*$.

1.8. Note. The ordered pair of mappings (σ_*, τ_*) is a Galois connection between the sets 2^{V^*} and $2^{V^* \times V^*}$. It follows from the definitions of the relations ρ and ρ_* that $(a, (u, v)) \in \rho$ holds if and only if $(a, (u, v)) \in \rho_*$ for every symbol a of the alphabet V.

1.9. Definition. We put $\psi_*(X) = \tau_* \sigma_*(X)$, where X is an arbitrary subset of the set V^* . Then ψ_* is a closure operator on 2^{V^*} . The set $\psi_*(X)$ is said to be the generalized grammatical category in the sense of Kunze generated by the set X or shortly the generalized category of the set X in the sense of Kunze.

In what follows we say "grammatical category" and "generalized grammatical category" if meaning "grammatical category in the sense of Kunze" and "generalized grammatical category in the sense of Kunze", respectively. We now show that grammatical categories represent a special case of generalized grammatical categories.

1.10. Lemma. Let (V, L) be a language, let X be a subset of the alphabet V and let Y be a subset of the set of all contexts $V^* \times V^*$. Then the following assertions hold.

(a) $\sigma(X) = \sigma_*(X)$.

(b) $\tau(Y) = \tau_*(Y) \cap V$.

(c) $\psi(X) = \psi_*(X) \cap V$.

Proof: (a) If $X \subseteq V$, we obtain, by 1.8., $\sigma_*(X) = \{(u, v) \in V^* \times V^* \mid (x, (u, v)) \in \varrho_*$ for every $x \in X\} = \{(u, v) \in V^* \times V^* \mid (x, (u, v)) \in \varrho \text{ for every } x \in X\} = \sigma(X).$

(b) We have $\tau_*(Y) \cap V = \{x \in V^* \mid (x, (u, v)) \in \varrho_* \text{ for every } (u, v) \in Y\} \cap V = \{x \in V \mid (x, (u, v)) \in \varrho_* \text{ for every } (u, v) \in Y\} = \{x \in V \mid (x, (u, v)) \in \varrho \text{ for every } (u, v) \in Y\} = \tau(Y).$

(c) Let $X \subseteq V$, then $\psi_*(X) \cap V = \tau_*(\sigma_*(X)) \cap V = \tau_*(\sigma(X)) \cap V = \tau(\sigma(X)) = \psi(X)$, Q.E.D.

1.11. Note. Clearly, we have $\psi_*(X) = \{v \in V^* \mid \psi_*(X) \supseteq \psi_*(\{v\})\}$, which is the consequence of 1.1.

2. CHARACTERIZATION OF REGULAR LANGUAGES BY MEANS OF GENERALIZED CATEGORIES

An important characterization of the class of regular languages is described in [2]. Definitions 2.1., 2.2. and Theorem 2.4. can be found there.

2.1. Definition. Let \equiv be an equivalence on V^* . If $x \equiv y$ implies $uxv \equiv uyv$ for every $u, v \in V^*$, then \equiv is called a *congruence*.

2.2. Definition. The *index of an equivalence* is ∞ , if the number of equivalence classes is infinite; it equals the number of equivalence classes, if this number is finite.

2.3. Definition. Let (V, L) be a language. We put $(x, y) \in \equiv_L$ for $x, y \in V^*$ if and only if $\sigma_*(\{x\}) = \sigma_*(\{y\})$ holds true. Obviously, \equiv_L is a congruence on V^* . It is called a *principal congruence*.

2.4. Theorem. For every language (V, L) the following three statements are equivalent:

(a) (V, L) is regular.

(b) There is a congruence \equiv on V^* of finite index such that L is a union of some of its classes.

(c) The relation \equiv_L is of finite index.

2.5. Note. If E is a class of the principal congruence, then $\sigma_*(\{e\}) = \sigma_*(E)$ and, hence, $\psi_*(\{e\}) = \psi_*(E)$ holds for all strings $e \in E$.

2.6. Lemma. Let (V, L) be a regular language. Then every nonempty generalized grammatical category is the union of some classes of the principal congruence corresponding to the language.

2.7. Lemma. A language having a finite number of generalized grammatical categories is regular.

Proof: Let us put $x \approx y$ for strings x, y if $\psi_*(\{x\}) = \psi_*(\{y\})$ holds. Then $\psi_*(\{x\}) = \psi_*(\{y\})$ is equivalent with $\sigma_*(\{x\}) = \sigma_*(\{y\})$ by Theorem 1.4. and, therefore, $x \equiv_L y$. Hence, $\approx = \equiv_L$. If a language possesses a finite number of generalized grammatical categories, the relation \approx is of finite index and the language is regular, Q.E.D.

2.8. Theorem. A language is regular if and only if it has a finite number of generalized grammatical categories.

Proof: Due to Lemma 2.6. and statement (c) of Theorem 2.4., a regular language possesses a finite number of generalized grammatical categories. The converse implication is included in Lemma 2.7., Q.E.D.

3. SYNTACTICAL CATEGORIES

We define a syntactical category by modifying slightly the definition in [3].

3.1. Definition. Let $G = (V_T, V_N, x_0, F)$ be a context-free grammar and let $x \in V_N$. We put

$$\mathscr{L}(x,G) = \{z \in V_T^* \mid x \stackrel{*}{\Rightarrow} z\}$$

and this set is called the syntactical category of the nonterminal symbol x in the grammar G. If $x = x_0$, then we omit the nonterminal symbol x and we put $\mathcal{L}(G) = \mathcal{L}(x_0, G)$.

There exists a regular language having a regular grammar such that some syntactical categories are not generalized grammatical categories, as it follows from the next example.

3.2. Example. Let $V_T = \{a, b, c\}$, $V_N = \{S, A, B, C, D\}$, $F = \{(S, aA), (A, bB), (B, \lambda), (S, aD), (D, cC), (C, \lambda)\}$. Then $G = (V_T, V_N, S, F)$ is a regular grammar generating the language $\{ab, ac\}$. Obviously, $A \stackrel{*}{\Rightarrow} b$, $D \stackrel{*}{\Rightarrow} c$. Thus the sets $\{b\}, \{c\}$ are syntactical categories of this grammar. On the other hand, $b \equiv_L c$ holds and, consequently, every generalized grammatical category either contains both symbols b, c or none of them. This implies that $\{b\}, \{c\}$ are not generalized grammatical categories.

On the other hand, to every regular language, there exists a regular grammar generating the language such that the syntactical categories of this grammar are generalized grammatical categories.

3.3. Theorem. Let (V, L) be a regular language. We define $\partial_p L = \tau_*(\{(p, \lambda)\})$ for every $p \in V^*$. Further we put

 $V_T = V,$ $V_N = \{\partial_p L \mid \lambda \neq p \in V^*, \ \partial_p L \neq \emptyset\} \cup \{\partial_\lambda L\},$ $F = \{(\partial_n L, a \ \partial_{nn} L) \mid \partial_{nn} L \in V_N, \ a \in V, \ p \in V^*\} \cup \{(\partial_n L, \lambda) \mid \partial_n L \in V_N, \ \lambda \in \partial_n L\}.$

Then the ordered quadruple $G = (V_T, V_N, \partial_\lambda L, F)$ is a regular grammar generating the language (V, L) and its syntactical categories are generalized grammatical categories of (V, L).

Proof: We prove that $\partial_p L = \mathscr{L}(\partial_p L, G)$ holds for every $\partial_p L \in V_N$.

1. Let us take $x \in \partial_p L$. If $x = \lambda$, then F contains the rule $(\partial_p L, \lambda)$ and $\partial_p L \stackrel{*}{\Rightarrow} \lambda$. Let $x = a_1, \ldots, a_n$, where $n \ge 1$ and $a_1, \ldots, a_n \in V_T$. This implies that F contains the rules $(\partial_p L, a_1 \partial_{pa_1} L)$, $(\partial_{pa_1} L, a_2 \partial_{pa_1 a_2} L)$, $\ldots, (\partial_{pa_1 \ldots a_{n-1}} L, a_n \partial_{pa_1 \ldots a_n} L)$ and ' $(\partial_{pa_1 \ldots a_n} L, \lambda)$. Hence, $\partial_p L \stackrel{*}{\Rightarrow} a_1 \ldots a_n = x$ and $\partial_p L \subseteq \mathcal{L}(\partial_p L, G)$ holds.

2. Suppose $\partial_p L \stackrel{*}{\Rightarrow} x$. Thus, we obtain $\partial_p L \Rightarrow a_1 \partial_{pa_1} L \Rightarrow a_1 a_2 \partial_{pa_1a_2} L \Rightarrow ... \Rightarrow \Rightarrow a_1 \dots a_n \partial_{pa_1 \dots a_n} L \Rightarrow a_1 \dots a_n$. Hence, $\lambda \in \partial_{pa_1 \dots a_n} L$ and this entails $pa_1 \dots a_n \in L$ and, hence, $a_1 \dots a_n = x \in \partial_p L$. We have proved $\partial_p L \supseteq \mathcal{L}(\partial_p L, G)$. If we put $p = \lambda$,

156

we obtain $\partial_{\lambda}L = \mathscr{L}(\partial_{\lambda}L, G) = \mathscr{L}(G)$. The assertion follows from the fact that $\partial_{p}L = \tau_{*}(\{(p, \lambda)\}) = \tau_{*}(\sigma_{*}(\tau_{*}(\{(p, \lambda)\}))) = \psi_{*}(\tau_{*}(\{(p, \lambda)\})))$ is a generalized grammatical category, Q.E.D.

4. EXAMPLE

4.1. Example. We consider the regular language (V, L), where $V = \{a, b\}$ and $L = \{b, ba\} \cup \{ab^m \mid m \ge 0\}$. We shall construct the principal congruence $=_L$.

Let C be the set of all strings in V^* that are not substrings of any string contained in L. We have $\sigma_*(C) = \emptyset$. We obtain also the other sets of contexts

$$\begin{aligned} \sigma_*(\{\lambda\}) &= \{(\lambda, b), (\lambda, ba), (b, \lambda), (b, a), (ba, \lambda)\} \cup \\ &\cup \{(\lambda, ab^i) \mid i \ge 0\} \cup \{(ab^i, b^j) \mid i \ge 0, j \ge 0\}, \\ \sigma_*(\{a\}) &= \{(b, \lambda)\} \cup \{(\lambda, b^i) \mid i \ge 0\}, \\ \sigma_*(\{b\}) &= \{(\lambda, \lambda), (\lambda, a)\} \cup \{(ab^i, b^j) \mid i \ge 0, j \ge 0\}, \\ \sigma_*(\{ab^i \mid i \ge 1\}) &= \{(\lambda, b^i) \mid i \ge 0\}, \\ \sigma_*(\{ba\}) &= \{(\lambda, \lambda)\}, \\ \sigma_*(\{b^i \mid i \ge 2\}) &= \{(ab^i, b^j) \mid i \ge 0, j \ge 0\}. \end{aligned}$$

We form the nonempty intersections of the sets $\sigma^*(X)$

$$\sigma_1 = \sigma_*(\{\lambda\}) \cap \sigma_*(\{a\}) = \{(b,\lambda),(\lambda,b)\},$$

$$\sigma_2 = \sigma_*(\{\lambda\}) \cap \sigma_*(\{b\}) = \{(\lambda,a)\} \cup \{(ab^i,b^j) \mid i \ge 0, j \ge 0\},$$

$$\sigma_3 = \sigma_*(\{\lambda\}) \cap \sigma_*(\{ab^i \mid i \ge 1\}) = \{(\lambda,b)\}.$$

Then the semilattice, possessing the above mentioned sets as elements and the intersection as operation, is given by the following Hasse diagram (it is constructed by the method described in [8]):



Let us put $A = \{\lambda\}$, $A = \{a\}$, $B = \{b\}$, $AB = \{ab^i \mid i \ge 1\}$, $BA = \{ba\}$, $BB = \{b^i \mid i \ge 2\}$. These sets together with C form a partition on V^* . In accordance with Definition 2.3., these sets are just all classes of the principal congruence \equiv_L .

Let $X \subseteq V^*$ be an arbitrary set. Then Definition 1.7. implies $\sigma_*(X) = \bigcap_{x \in X} \sigma_*(\{x\})$ and, therefore, $\sigma_*(X) = \bigcap_{T \in V^* \equiv L} \sigma_*(T)$ by Note 2.5.

We shall find all classes T that have nonempty intersections with X and then the infimum of the sets $\sigma_*(T)$. This infimum equals $\sigma_*(X)$. By Note 1.11. $\psi_*(X)$ consists of all $x \in V^*$ satisfying $\psi_*(X) \supseteq \psi_*(\{x\})$ which means $\sigma_*(X) \subseteq \sigma_*(\{x\})$ by Theorem 1.4. These elements x can be found by means of the diagram. In this way, we find all generalized grammatical categories: Λ , A, B, $A \cup AB$, $A \cup B \cup AB \cup U$ and U = U.

To construct a grammar by Theorem 3.3., we find all nonempty sets of the form $\tau_*(\{(x, \lambda)\})$, where $x \in V^*$. For the sake of brevity, we write $\tau_*(\{(x, \lambda)\}) = \partial_x L$ as in Theorem 3.3. It is easy to see that $\partial_x L$ is a union of the classes $T \in V^*/\equiv_L$ with the property $(x, \lambda) \in \sigma_*(T)$. Hence, we obtain

$$\partial_{\lambda}L = A \cup B \cup AB \cup BA,$$

$$\partial_{a}L = A \cup B \cup BB = \partial_{ab}L,$$

$$\partial_{b}L = A \cup A,$$

$$\partial_{b}L = A.$$

Thus $V_N = \{\partial_{\lambda}L, \partial_a L, \partial_b L, \partial_{ba}L\}$. The empty word is contained in the sets $\partial_a L$, $\partial_b L$, $\partial_{ba}L$. Therefore, $F = \{(\partial_{\lambda}L, a \partial_a L), (\partial_{\lambda}L, b \partial_b L), (\partial_a L, b \partial_a L), (\partial_b L, a \partial_{ba}L), (\partial_a L, \lambda), (\partial_b L, \lambda), (\partial_b L, \lambda)\}$. The resulting grammar is the ordered quadruple:

$$G = (V, V_N, \partial_{\lambda} L, F).$$

LITERATURE

- [1] Birkhoff, G.: Teorija struktur, Izdateľstvo inostrannoj literatury, Moscow 1952.
- [2] Ginsburg, S.: The Mathematical Theory of Context-free Languages, Mc Graw-Hill Book Company, Inc., 1966.
- [3] Hopcroft, J. E. and Ullman, J. D.: Formal Languages and Their Relation to Automata, Adison-Wesley Publishing Company, Inc., 1969.
- [4] Kunze, J.: Versuch eines objektivierten Grammatikmodells I, II, Z. Phonetik Sprachwiss. Kommunikat., 20, 1967, 415-448 and 21, 1968, 421-466.
- [5] Kunze, J. and Priess, W.: Versuch eines objektivierten Grammatikmodells III, Z. Phonetik Sprachwiss. Kommunikat., 23, 1970, 345–378.
- [6] Novotný, M.: Abstract Grammatical Categories, in print (Prague Studies in Mathematical Linguistics).
- [7] Salomaa, A.: Formal Languages, Academic Press, Inc., 1973.
- [8] Szász, G.: Introduction to Lattice Theory, Publishing House of the Hungarian Academy of Sciences, Budapest, 1962.

B. Kříž 664 72 Jinačovice' 76 Czechoslovakia