Kybernetika

Miroslav Novotný

Complete characterization of context-sensitive languages

Kybernetika, Vol. 10 (1974), No. 2, (73)--79

Persistent URL: http://dml.cz/dmlcz/124862

KYBERNETIKA — VOLUME 10 (1974), NUMBER 2

Complete Characterization of Context-Sensitive Languages

MIROSLAV NOVOTNÝ

Intrinsic complete characterizations of constructive, context-free and regular languages have been formulated by means of configurations of languages. The definition of a semiconfiguration is given here by generalizing the definition of a configuration. By means of semiconfigurations, an intrinsic complete characterization of context-sensitive languages is formulated.

1. Languages and generalized grammars. If V is a set we denote by V^* the free monoid over V, i.e. the set of all finite sequences of elements of the set V including the empty sequence Λ this set being provided by the binary operation of concatenation. We identify one-member-sequences with elements of V; it follows $V \subseteq V^*$. If $x = x_1 x_2, \dots x_n \in V^*$ where n is a natural number and $x_i \in V$ for $i = 1,2, \dots, n$ we put $|x| = n$; further, we put $|\Lambda| = 0$.

An ordered pair (V, L) where V is a set and $L \subseteq V^*$ is called a *language*. The elements of V^* are called *strings*. If (V, L), (U, M) are languages then we define the *intersection* $(V, L) \cap (U, M)$ of these languages by the formula $(V, L) \cap (U, M) = (V \cap U, L \cap M)$.

Let V be a set, suppose $R \subseteq V^* \times V^*$. Let us have $x, y \in V^*$. We writte $x \to y(R)$ if $(x, y) \in R$. Further, we put $x \Rightarrow y(R)$ if there exist such strings $u, v, t, z \in V^*$ that $x = utv$, $uzv = y$, $t \to z(R)$. Finally, we write $x \Rightarrow^* y(R)$ if there exist an integer $p \geqq 0$ and some strings $x = t_0, t_1, \dots, t_p = y$ in V^* that $t_{i-1} \Rightarrow t_i(R)$ for $i = 1, 2, \dots \dots, p$. Then the sequence of strings $(t_i)_{i=0}^p$ is called an *x-derivation of y of length p in R*.

Let V be a set, $V_T \subseteq V$, $S \subseteq V^*$, $R \subseteq V^* \times V^*$. Then the quadruple $G = \langle V, V_T, S, R \rangle$ is called a *generalized grammar*. We put $\mathscr{L}(G) = \{x;\ x \in V_T^*$, there exists an $s \in S$ with $s \Rightarrow^* x(R)\}$. Then $(V_T, \mathscr{L}(G))$ is called the *language generated by the generalized grammar* G. A generalized grammar $G = \langle V, V_T, S, R \rangle$ is called *special* if $V_T = V$; then we write $\langle V, S, R \rangle$ instead of $\langle V, V, S, R \rangle$. A generalized grammar $G = \langle V, V_T, S, R \rangle$ is called a *grammar* if the sets V, S, R are finite.

2. Phrase structure grammars. Let $G = \langle V, V_T, S, R\rangle$ be a grammar. This grammar is said to satisfy the condition

(A) if $(x, y) \in R$ implies $\Lambda \neq x$;
(B) if $(x, y) \in R$ implies $x \in (V - V_T)^*$;
(C) if there exists and element $\sigma \in V - V_T$ with the property $S = \{\sigma\}$;
(D) if $(x, y) \in R$ implies $|x| \leqq |y|$;
(E) if $(x, y) \in R$ implies $|x| = 1$;
(F) if $(x, y) \in R$ implies $1 = |x| \leqq |y|$.

A grammar with the properties (A), (B), (C) is called a *phrase structure grammar.* A phrase structure grammar with the property (D) is called *context sensitive.* A phrase structure grammar with the property (E) is called *context free.* A phrase structure grammar with the property (F) is called *context free Λ-free.*

A language is called *constructive* [*context sensitive, context free, context free Λ-free*] if it is generated by a phrase structure grammar [by a context-sensitive, by a context-free, by a context-free Λ-free grammar] (cf. [1]). Clearly, each context-free Λ-free grammar is context sensitive. Thus, each context-free Λ-free language is context sensitive.

3. Theorem. (A) *To each grammar* $G = \langle V, V_T, S, R\rangle$ *there exists a phrase structure grammar* $H = \langle U, V_T, \{\sigma\}, P\rangle$ *such that* $\mathscr{L}(H) = \mathscr{L}(G)$.

(B) *To each grammar* $G = \langle V, V_T, S, R\rangle$ *with the property* (D) *there exists a context-sensitive grammar* $H = \langle U, V_T, \{\sigma\}, P\rangle$ *such that* $\mathscr{L}(H) = \mathscr{L}(G) - \{\Lambda\}$.

(C) *To each grammar* $G = \langle V, V_T, S, R\rangle$ *with the property* (E) *there exists a context-free grammar* $H = \langle U, V_T, \{\sigma\}, P\rangle$ *such that* $\mathscr{L}(H) = \mathscr{L}(G)$.

(D) *To each grammar* $G = \langle V, V_T, S, R\rangle$ *with the property* (F) *there exists a context-free Λ-free grammar* $H = \langle U, V_T, \{\sigma\}, P\rangle$ *such that* $\mathscr{L}(H) = \mathscr{L}(G) - \{\Lambda.\}$

The assertions (A), (B) can be found in [2] Theorem 4.4, the proofs can be found in [3] p. 51–52. The assertion (C) coincides with 1.16 of [4]. The assertion (D) follows from (C) by Theorem 1.8.1 of [1].

4. Conditions for grammars. Let $G = \langle V, V_T, \{\sigma\}, R\rangle$ be a phrase structure [context-sensitive, context-free, context-free Λ-free] grammar. Then, we can suppose, without loss of generality, that G has the following two properties: (M) $(x, y) \in R$ implies $x \neq y$; (N) $(x, y) \in R$ implies the existence of such $z \in V_T^*$, $u, v \in V^*$ that $\sigma \Rightarrow^* uxv(R)$, $uyv \Rightarrow^* z(R)$.

Clearly, each $(x, y) \in R$ for which the condition contained in (M) is not fulfilled can be cancelled and the language generated by the grammar obtained in this way is $(V_T, \mathscr{L}(G))$. Thus, we can suppose that G has the property (M). Similarly, a pair $(x, y) \in R$ which does not fulfil the condition contained in (N) does not appear in any σ-derivations of strings of $\mathscr{L}(G)$ in R. Thus, each such pair can be cancelled and the language generated by the grammar obtained in this way is $(V_T, \mathscr{L}(G))$.

5. Topics of paper. The definitions of constructive, context-sensitive, context-free and regular languages (cf. [1], Chapter II, 2. 1) are formulated by means of grammars with certain properties. A complete characterization of regular languages which does not use explitcitly the concept of a grammar is well known ([1] Theorem 2.1.5). The author found complete characterizations of constructive languages [5], of context-free languages [4] and of regular languages [6] in the terms of the theory of configurations.

The aim of this paper is to give an intrinsic complete characterization of context-sensitive languages, i.e. a complete characterization which does not use explicitly the concept of a grammar. It was necessary to generalize the notion of a configuration to this aim. A modification of this generalized notion gives a new intrinsic complete characterization of context-free languages.

6. Definitions. Let (V, L) be a language.

For $x \in V^*$ we put $x\ v\ (V, L)$ if there exist such strings $u, v \in V^*$ that $uxv \in L$.

For $x, y \in V^*$ we put $x > y(V, L)$ if, for all $u, v \in V^*$, $uxv \in L$ implies $uyv \in L$.

For $x, y \in V^*$ we put $(y, x) \in E(V, L)$ if the following conditions are satisfied: $y\ v\ (V, L)$, $y > x(V, L)$, $y \neq x$, $|y| \leqq |x|$. Then x is called a *semiconfiguration with the resultant* y *in the language* (V, L).

7. Remark. If (V, L) is a language, $t, z \in V^*$ such strings that $t \Rightarrow^* z\ (E(V, L))$ then $|t| \leqq |z|$ which follows from the fact that $(y, x) \in E(V, L)$ implies $|y| \leqq |x|$.

8. Definition. Let (V, L) be a language. Then, for $x \in L$, we put $x \in B(V, L)$ if, for each $t \in L$, $t \Rightarrow^* x(E(V, L))$ implies $|t| = |x|$.

9. Remark. Let (V, L) be a language. Then for each $x \in L$ there exists a string $s \in B(V, L)$ that $s \Rightarrow^* x(E(V, L))$. – Indeed, there exists at least one string $s \in L$ with the property $s \Rightarrow^* x(E(V, L))$; e.g. we can put $s = x$. If we take such an s of minimal length then, clearly, $s \in B(V, L)$.

10. Definitions. Let (V, L) be a language. If $s, t \in V^*$ are such strings that $s \Rightarrow t(E(V, L))$ then we put $|(s, t)| = \min\{|q|; (p, q) \in E(V, L), s \Rightarrow t(\{(p, q)\})\}$. If $s, t \in V^*$ are strings and $(t_i)_{i=0}^p$ and s-derivation of t in $E(V, L)$ then we put $\|(t_i)_{i=0}^p\| = 0$ if $p = 0$ and $\|(t_i)_{i=0}^p\| = \max\{|(t_{i-1}, t_i)|; i = 1, 2, \ldots, p\}$ otherwise. The integer $\|(t_i)_{i=0}^p\|$ is called the *norm of the s-derivation* $(t_i)_{i=0}^p$ *of* t *in* $E(V, L)$. If $s, t \in V^*$ are such strings that $s \Rightarrow^* t(E(V, L))$ then we define the *norm* $\|(s, t)\|$ *of the ordered pair* (s, t) to be the minimum of norms of all s-derivations of t in $E(V, L)$. If $t \in L$ then we put $\|t\| = \min\{\|(s, t)\|; s \in B(V, L), s \Rightarrow^* t(E(V, L))\}$; the integer $\|t\|$ is called the *norm of* t.

11. Lemma. *Let* (V, L) *be a language. Then, for each* $t \in L$, *there exists a string* $s \in B(V, L)$ *and an s-derivation of* t *in* $E(V, L)$ *such that the norm of this s-derivation is equal to* $\|t\|$.

Indeed, there exists such an element $s \in B(V, L)$ that $\|(s, t)\| = \|t\|$. It means the existence of such an s-derivation of t in $E(V, L)$ that its norm is equal to $\|t\|$.

12. Definition. Let (V, L) be a language. Then we put $X(V, L) = \{(y, x);\ (y, x) \in E(V, L),\ |x| > \|t\| \text{ for each } t \in L\}$, $Z(V, L) = E(V, L) - X(V, L)$.

13. Corollary. *Let (V, L) be a language. Then, for each $t \in L$, there exists at least one element $s \in B(V, L)$ such that $s \Rightarrow^* t(Z(V, L))$.*

Proof. According to 11, there exists a string $s \in B(V, L)$ and an s-derivation $(t_i)_{i=0}^{p}$ of t in $E(V, L)$ such that $\|(t_i)_{i=0}^{p}\| = \|t\|$. It follows from 10 that $|(t_{i-1}, t_i)| \leqq \|t\|$ for $i = 1, 2, \ldots, p$. Thus, for each $i = 1, 2, \ldots, p$, there exists an element $(p_i, q_i) \in E(V, L)$ such that $t_{i-1} \Rightarrow t_i(\{(p_i, q_i)\})$ and $|q_i| = |(t_{i-1}, t_i)| \leqq \|t\|$. It follows $(p_i, q_i) \in Z(V, L)$ for $i = 1, 2, \ldots, p$ and $s \Rightarrow^* t(Z(V, L))$.

14. Definitions. Let (V, L) be a language. We put $K(V, L) = \langle V, B(V, L), Z(V, L)\rangle$.

15. Theorem. *Let (V, L) be a language. Then $\mathscr{L}(K(V, L)) = L$.*

Proof. According to 13, $L \subseteq \mathscr{L}(K(V, L))$.

Let $V(n)$ denote the following assertion: If $t \in \mathscr{L}(K(V, L))$ and there exists an element $s \in B(V, L)$ and an s-derivation of t of length n in $Z(V, L)$ then $t \in L$.

If $t \in \mathscr{L}(K(V, L))$ and there exists an element $s \in B(V, L)$ and an s-derivation of t of length 0 in $Z(V, L)$ then $t = s \in B(V, L) \subseteq L$. Thus $V(0)$ holds true.

Let $m \geqq 0$ be an integer and suppose that $V(m)$ holds true. Let us have $t \in \mathscr{L}(K(V, L))$, $s \in B(V, L)$ and an s-derivation $(t_i)_{i=0}^{m+1}$ of t of length $m + 1$ in $Z(V, L)$. Then $t_m \in L$ according to $V(m)$. Further, $t_m \Rightarrow t(Z(V, L))$ which means the existence of strings $u, v, x, y \in V^*$ such that $t_m = uyv$, $uxv = t$, $(y, x) \in Z(V, L) \subseteq E(V, L)$. It implies $y > x(V, L)$, thus, $t \in L$. We have proved that $V(m)$ implies $V(m + 1)$.

It follows that $V(n)$ holds true for $n = 0, 1, 2, \ldots$. It means $\mathscr{L}(K(V, L)) \subseteq L$.

16. Definition. Let (V, L) be a language. Then it is called *finitely semigenerated* if the sets V, $B(V, L)$, $Z(V, L)$ are finite.

17. Lemma. *Let (V, L) be a finitely semigenerated language such that $\Lambda \notin L$, U an arbitrary finite set. Then $(V, L) \cap (U, U^*)$ is a context-sensitive language.*

Proof. If (V, L) is a finitely semigenerated language then $L = \mathscr{L}(K(V, L))$ according to 15 and $K(V, L) = \langle V, B(V, L), Z(V, L)\rangle$ is a special grammar according to 16. We put $H = \langle V, V \cap U, B(V, L), Z(V, L)\rangle$. Then H is a grammar with the following properties: $(y, x) \in Z(V, L)$ implies $|y| \leqq |x|$ and $\mathscr{L}(H) = \mathscr{L}(K(V, L)) \cap U^* = L \cap U^*$. According to 3 (B) there exists a context-sensitive grammar $G = \langle W, V \cap U, \{\sigma\}, R\rangle$ such that $\mathscr{L}(G) = \mathscr{L}(H) - \{\Lambda\} = L \cap U^* - \{\Lambda\} = L \cap U^*$.

Thus, $(V, L) \cap (U, U^*) = (V \cap U, L \cap U^*)$ is the language generated by the context-sensitive grammar G, i.e. it is a context-sensitive language.

18. Lemma. *Let (U, M) be a context-sensitive language. Then there exists a finitely semigenerated language (V, L) with the property $\Lambda \notin L$ such that $(V, L) \cap (U, U^*) = (U, M)$.*

Proof. A) There exists a context-sensitive grammar $G = \langle W, U, \{\sigma\}, R\rangle$ such that $\mathscr{L}(G) = M$. According to 4, we can suppose that $(y, x) \in R$ implies $y \neq x$ and the existence of strings $z \in U^*$, $u, v \in W^*$ such that $\sigma \Rightarrow^* uyv(R)$, $uxv \Rightarrow^* z(R)$. We put $H = \langle W, \{\sigma\}, R\rangle$. Then $\mathscr{L}(G) = \mathscr{L}(H) \cap U^*$. We prove that $(W, \mathscr{L}(H))$ is a finitely semigenerated language. Clearly, $\Lambda \notin \mathscr{L}(H)$.

B) First of all, as $(y, x) \in R$ implies the existence of $u, v \in W^*$ with the property $\sigma \Rightarrow^* uyv(R)$, we have $uyv \in \mathscr{L}(H)$ and $yv(W, \mathscr{L}(H))$.

Further, $(y, x) \in R$ implies $y > x$ $(W, \mathscr{L}(H))$ and $y \neq x$ follows from our hypothesis. The fact $|y| \leqq |x|$ follows from the supposition that G is context sensitive.

Thus, $(y, x) \in R$ implies $(y, x) \in E(W, \mathscr{L}(H))$ and $R \subseteq E(W, \mathscr{L}(H))$.

C) Let us have $z \in \mathscr{L}(H)$, $|z| > 1$. Then $\sigma \Rightarrow^* z(R)$ which implies $\sigma \Rightarrow^* \Rightarrow^* z(E(W, \mathscr{L}(H))$ according to B. As $|\sigma| = 1$, we have $z \notin B(W, \mathscr{L}(H))$ according to 8. Thus, $z \in B(W, \mathscr{L}(H))$ implies $|z| \leqq 1$ and $B(W, \mathscr{L}(H))$ is finite. Clearly, $\sigma \in B(W, \mathscr{L}(H))$.

D) We put $N = \max\{|x|;\ (y, x) \in R\}$. Since $z \in \mathscr{L}(H)$ implies $\sigma \Rightarrow^* z(R)$ and $R \subseteq E(W, \mathscr{L}(H))$ according to B, we have $\|z\| \leqq N$ for each $z \in \mathscr{L}(H)$. According to 12, $(y, x) \in Z(W, \mathscr{L}(H))$ implies $(y, x) \in E(W, \mathscr{L}(H))$ and the existence of a $z \in L(H)$ such that $|x| \leqq \|z\|$ which implies $|y| \leqq |x| \leqq N$. It implies the finiteness of $Z(W, \mathscr{L}(H))$.

E) It follows from C and D that $(W, \mathscr{L}(H))$ is finitely semigenerated language and that $(U, M) = (U, \mathscr{L}(G)) = (W \cap U, \mathscr{L}(H) \cap U^*) = (W, \mathscr{L}(H)) \cap (U, U^*)$.

19. Theorem. *Let U be a finite set, (U, M) a language. Then the following two assertions are equivalent:*

(A) *(U, M) is a context-sensitive language.*

(B) *There exists a finitely semigenerated language (V, L) with the property $\Lambda \notin L$ such that $(V, L) \cap (U, U^*) = (U, M)$.*

It is a consequence of 17 and 18.

20. Remarks, definitions. We can modify the concept of a semiconfiguration in the following way: Let (V, L) be a language. For $x, y \in V^*$ we put $(y, x) \in \bar{E}(V, L)$ if the following conditions are satisfied: $yv(V, L)$, $y > x(V, L)$, $y \neq x$, $1 = |y| \leqq |x|$. Then x is called a *strong semiconfiguration with the resultant y in the language (V, L)*. For $x \in L$ we put $x \in \bar{B}(V, L)$ if, for each $t \in L$, $t \Rightarrow^* x(\bar{E}(V, L))$ implies $|t| = |x|$. Further, for $s, t \in V^*$ such that $s \Rightarrow t(\bar{E}(V, L))$, we put $[(s, t)] = \min\{|q|;\ (p, q) \in$

$\in \bar{E}(V, L)$, $s \Rightarrow t(\{(p, q)\})\}$. If $s, t \in V^*$ are strings and $(t_i)_{i=0}^p$ is an s-derivation of t in $\bar{E}(V, L)$ then we put $[\![(t_i)_{i=0}^p]\!] = 0$ if $p = 0$ and $[\![(t_i)_{i=0}^p]\!] = \max\{[(t_{i-1}, t_i)]; i = 1, 2, \ldots, p\}$ otherwise. The integer $[\![(t_i)_{i=0}^p]\!]$ is called the *strong norm of the s-derivation* $(t_i)_{i=0}^p$ *of* t *in* $\bar{E}(V, L)$. If $s, t \in V^*$ are such strings that $s \Rightarrow^* t(\bar{E}(V, L))$ then we define the *strong norm* $[\![(s, t)]\!]$ *of the ordered pair* (s, t) to be the minimum of strong noms of all s-derivations of t in $\bar{E}(V, L)$. If $t \in L$ then we put $[\![t]\!] = \min\{[\![(s, t)]\!]; s \in \bar{B}(V, L), s \Rightarrow^* t(\bar{E}(V, L))\}$; the integer $[\![t]\!]$ is called the *strong norm of t.*

Further, we put $\bar{X}(V, L) = \{(y, x); (y, x) \in \bar{E}(V, L), |x| > [\![t]\!]$ for each $t \in L\}$, $\bar{Z}(V, L) = \bar{E}(V, L) - \bar{X}(V, L)$. Finally, we define $\bar{K}(V, L) = \langle V, \bar{B}(V, L), \bar{Z}(V, L)\rangle$. Similarly as in 15 we prove

21. Theorem. *Let* (V, L) *be a language. Then* $\mathscr{L}(\bar{K}(V, L)) = L$.

22. Definition. Let (V, L) be a language. Then (V, L) is called *strongly finitely semigenerated* if the sets V, $\bar{B}(V, L)$, $\bar{Z}(V, L)$ are finite.

Similarly as in 19 we prove

23. Theorem. *Let* U *be a finite set,* (U, M) *a language. Then the following two assertions are equivalent:*

(A) (U, M) *is a context-free* Λ*-free language.*

(B) *There exists a strongly finitely semigenerated language* (V, L) *with the property* $\Lambda \notin L$ *such that* $(V, L) \cap (U, U^*) = (U, M)$.

If we take into account the connection between context-free Λ-free grammars and context-free grammars described in the Theorem 1.8.1 of [1] then we obtain

24. Theorem. *Let* U *be a finite set,* (U, M) *a language. Then the following two assertions are equivalent:*

(A) (U, M) *is a context-free language.*

(B) *There exists a strongly finitely semigenerated language* (V, L) *such that* $(V, L) \cap (U, U^*) = (U, M)$.

(Received December 4, 1972.)

REFERENCES

[1] S. Ginsburg; The mathematical theory of context-free languages. McGraw-Hill Book Company, 1966.

[2] M. Novotný: Algebraic structures of mathematical linguistics. Bull. Math. de la Soc. Sci. Math. de la R. S. de Roumanie *12 (60)* (1969), 87—101.

[3] M. Novotný: Einführung in die algebraische Linguistik. Rheinisch-Westfällisches Institut für Instrumentelle Mathematik an der Universität Bonn, 1967 (Skriptum).

[4] M. Novotný: On a class languages. Archivum Mathematicum Brno, *6* (1970), 155—170.
[5] M. Novotný: On the role of configurations in the theory of grammars. Archivum Mathematicum Brno, *6* (1970), 171—184.
[6] M. Novotný: Über endlich charakterisierbare Sprachen. Publ. Fac. Sci. Univ. J. E. Purkyně, Brno, No 468 (1965), 495—502.

RNDr. Miroslav Novotný, DrSc., Matematický ústav ČSAV — pobočka Brno (Mathematical Institute of the Czechoslovak Academy of Sciences — Branch Brno), Janáčkovo nám. 2a, 662 95 Brno. Czechoslovakia.