KYBERNETIKA ČÍSLO 2, ROČNÍK 4/1968

Some Theorems on Labelled Bracketings Used in Transformational Grammars

KAREL ČULÍK

There are proved several lemmas and theorems concerning certain types of decompositions of well-formed labelled bracketings which are nothing else than a linear expression of phrase-markers used in context-free grammars.

The following theorems concern the notions introduced in [1] in order to formalize the theory of transformational grammar presented in [2]. Thus primarily the labelled bracketings have their meaning in linguistics or in the mathematical theory of languages because they are sequences of symbols expressing uniquely the phrase-markers of context-free grammars. There is a correspondence between the well formed labelled bracketings and the phrase-markes and markers defined as the special graphs in [4] and [5]. On the other hand some pure abstract results have more general mathematical character and are connected with the bracketing mentioned in [3].

Finite, disjoint sets V_T and V_N are said to be *terminal* and *nonterminal vocabularies* resp. The pair $([, A)$ or $(], A)$ is said to be a *left* or *right labelled bracket* resp. where $A \in V_N$ and instead of $([, A)$ or $(], A)$ one writes $\underset{A}{[}$ or $\underset{A}{]}$ resp. Then $L = \{\underset{A}{[}; A \in V_N\}$ and $R = \{\underset{A}{]}; A \in V_N\}$ and a *terminal labelled bracketing* (lb) is a finite string of symbols from $V_T \cup L \cup R$. The free semigroup of all strings the generators of which belong to the set M is denoted by M^∞ and $M^{\infty_0} = M^\infty \cup \{e\}$ where e is the *identity element* of the semigroup M^∞, i.e. e is the empty string the length of which $l(e) = 0$. Many other special definitions and notations are introduced in [1] and here accepted without any change. First of all in the definition 1.1 of [1] a *well formed labelled bracketing* (wflb) is introduced as follows: a lb ψ is a wflb if either (i) $\psi \in V_T \cup V_N$, or (ii) $\psi = \psi_1\psi_2$ where ψ_1, ψ_2 are wflb or (iii) $\psi = \underset{A}{[}\psi'\underset{A}{]}$ where $\underset{A}{[} \in L, \underset{A}{]} \in R$ and ψ' is a wflb.

A lb ψ is said to be *in the basic form* if $\psi = \lambda_1 X_1 \varrho_1 \lambda_2 X_2 \varrho_2 \dots \lambda_n X_n \varrho_n$ where $n \geqq 1$, $X_i \in V_T$, $\lambda_i \in L^{\infty_0}$ and $\varrho_i \in R^{\infty_0}$ for each $i = 1, 2, \dots, n$.

Let ψ be a lb and let $\psi = \alpha a \beta \bar{a} \gamma$, where $a \in L$ and $\bar{a} \in R$. The occurrence shown of $\bar{a}$ is said to be a *corresponding occurrence* to the shown occurrence of a (and conversely) if it is the first occurrence of $\bar{a}$ in ψ on the right of a which satisfies the following condtions: a and $\bar{a}$ are labelled by the same nonterminal symbol and the number of occurrences of the left brackets in β is the same as the number of the right ones.

A lb ψ satisfies the *bracket condition* if to each occurrence of a left bracket in ψ there exists the corresponding occurrence of a right bracket in ψ and if the number of occurrences of right brackets in ψ is not greater than of left ones.

Lemma 1. *Let ψ be a* lb *satisfying the bracket condition and let $\psi = \delta a \varphi \bar{a} \gamma$ where $\varphi = \alpha b \beta$; $a, b \in L$, $\bar{a} \in R$ and a and $\bar{a}$ are the corresponding brackets. If $\bar{b}$ is the corresponding bracket to b, then $\bar{b}$ can occur neither in δ nor in γ but always in β. Therefore φ and $\delta\gamma$ satisfy the bracket condition too.*

Proof. Let us assume that $\bar{b}$ do no occurs in φ. Then according to the bracket condition the number of left brackets in φ is the same as the number of the right ones and therefore there must be a right bracket $\bar{c} \in R$ occuring in φ the corresponding left bracket c of which does not belong to φ. This means that c must occur either in γ what is a contradiction because the corresponding right bracket $\bar{c}$ is on the left and not on the right of the left bracket c, or c occurs in δ. In this case we repeat the previous considerations for the pair c, $\bar{c}$ instead of a, $\bar{a}$ and for the left bracket a instead of b. This leads to a regress ad infinitum what is a contradiction to the finiteness of $l(\varphi)$.

Thus $\bar{b}$ must occur in φ and this is true for each left bracket b in φ. Therefore – as ψ satisfies the bracket condition – φ satisfies it as well and in a similar way one proves the same for $\delta\gamma$.

Theorem 1. *A* lb *ψ is a terminal* wflb *if and only if ψ is in the basic form and if ψ satisfies the bracket condition.*

Proof. Let ψ be a terminal wflb. If $l(\psi) = 1$, then $\psi \in V_T$ and therefore ψ is in the basic form. The condition concerning the brackets is satisfied trivially (there is no bracket in ψ). If $l(\psi) = k > 1$, then either $\psi = \psi'\psi''$ or $\psi = a\psi'\bar{a}$, where ψ and ψ'' are the terminal wflb's such that $l(\psi') < k$, $l(\psi'') < k$ and $a \in L$, $\bar{a} \in R$ and $\bar{a}$ is the corresponding occurrence to a. In the first case according to the inductive assumption $\psi' = \lambda_1' X_1 \varrho_1' \dots \lambda_{n'}' X_{n'} \varrho_{n'}$, and $\psi'' = \chi_1 X_1 \varrho_1'' \dots \lambda_{n''}'' X_{n''}'' \varrho_{n''}''$ and therefore $\psi'\psi''$ is in the basic form too. Further ψ' and ψ'' satisfy our condition concerning their brackets and therefore obviously this condition is satisfied by $\psi'\psi''$ too.

In the second case by the inductive assumption it follows that ψ' is in the basic form and that ψ' satisfies the bracket condition. It is quite clear that than $a\psi'\bar{a}$ satisfies both these conditions too.

Now on the contrary let $\psi = \lambda_1 X_1 \varrho_1 \dots \lambda_n X_n \varrho_n$ and let ψ satisfy the bracket condition. If $l(\psi) = 1$, then $\psi \in V_T$ and ψ is a terminal wflb. If $l(\psi) = k > 1$, then we shall distinguish two possibilities $\lambda_1 = e$ and $\lambda_1 \neq e$.

In the first case from the bracket condition if follows $\varrho_1 = e$ and therefore it is clear that $\varphi = \lambda_2 X_2 \varrho_2 \ldots \lambda_n X_n \varrho_n$ satisfies the bracket condition. Thus by the inductive assumption – because $l(\varphi) < k$ – φ is a terminal wflb and therefore $\psi = X_1 \varphi$ a terminal wflb too.

In the second case one can write $\lambda_1 = a\lambda_1'$ where $a \in L$. From the bracket condition follows the existence of φ and γ such that $\psi = a\varphi\bar{a}\gamma$, where $\bar{a}$ is the corresponding right bracket to a. By Lemma 1, φ and γ (because $\delta = e$) must satisfy the bracket condition and therefore they must have the basic forms. Thus by the inductive assumption – because $l(\varphi) < k$ and $l(\gamma) < k$ – φ and γ are the terminal wflb's and therefore $\psi = a\varphi\bar{a}\gamma$ must be a terminal wflb too.

According to the definition 1.2 of [1] one can assigne the *debracketization* $d(\varphi)$ to the lb φ as follows: if $\varphi = X_1 X_2 \ldots X_n$ where $X_i \in V_T \cup L \cup R$ for each $i = 1, 2, \ldots, n$ then $d(\varphi) = x_{k_1} x_{k_2} \ldots x_{k_p}$ where $1 \leq k_1 < k_2 < \ldots < k_p \leq n$ and $x_{k_i} \in V_T$ for each $i = 1, 2, \ldots, p$ but $x_j \in L \cup R$ for each j such that $1 \leq j \leq n$ and $j \neq k_i$ for each $i = 1, 2, \ldots, p$.

The further important notion is the *standard factorization*. A sequence of lb's $(\psi_1, \psi_2, \ldots, \psi_k)$ is said to be the *standard factorization* of lb ψ if (i) $\psi = \psi_1 \psi_2 \ldots \psi_k$, (ii) either $\psi_i = e$ or $d(\psi_i) \neq e$ and (iii) the leftmost or rightmost symbol of ψ_i is not a right or left bracket resp.

In the definition 1.4 of [1] it is inconvenient to allow $\psi_i = e$ and to prescribe the number k characterizing the sequence $(\psi_1, \psi_2, \ldots, \psi_k)$. Therefore we shall call a standard factorization $(\psi_1, \psi_2, \ldots, \psi_k)$ *right* if $d(\psi_i) \neq e$ for each $i = 1, 2, \ldots, k$. Further the *maximal right standard factorization* of a wflb has the maximal length k.

It is clear that it is sufficient to study only the right standard factorizations because each not right standard factorization can be obtained from a right one by adding some elements e between some neighbooring strings in the sequence.

Theorem 2. *Let $\lambda_1 X_1 \varrho_1 \lambda_2 X_2 \lambda_2 \ldots \lambda_n X_n \varrho_n$ be the basic form of a terminal* wflb ψ *and let us denote $w_i = \lambda_i X_i \varrho_i$ for each $i = 1, 2, \ldots, n$. Then $(w_1, w_2, \ldots, w_n)$ is the maximal standard factorization of ψ. Further a sequence of strings $(\psi_1, \psi_2, \ldots, \psi_k)$ is a right standard factorization of ψ if and only if there are integers $1 \leqq p_1 < p_2 < \ldots p_k = n$ such that $\psi_1 = w_1 w_2 \ldots w_{p_1}$ and $\psi_j = w_{p_{j-1}+1} w_{p_{j-1}+2} \ldots w_{p_j}$ for each $j = 2, 3, \ldots, k$.*

Proof. It is clear that really $(w_1, w_2, \ldots, w_n)$, is the maximal right standard factorization of ψ. Further let us assume that $(\psi_1, \psi_2, \ldots, \psi_k)$ is a right standard factorization of ψ, i.e. $\psi_1 \psi_2 \ldots \psi_k = \psi$ and $d(\psi_i) \neq e$ and the leftmost or rightmost symbol of ψ_i does not belong to R or to L resp. for each $i = 1, 2, \ldots, k$. Then $\psi_1 \psi_2 \ldots \psi_k = \lambda_1 X_1 \varrho_1 \lambda_2 X_2 \varrho_2 \ldots \lambda_n X_n \varrho_n$ and between X_i and X_{i+1} there can be at most one cut and if it is the case this cut must be between ϱ_i and λ_{i+1} what means that there are the required integers p_i. On the other side, if there are the required integers p_i such that $\psi_1 = w_1 w_2 \ldots w_{p_1}$ and $\psi_j = w_{p_{j-1}+1} w_{p_{j-1}+2} \ldots w_{p_j}$ for $j = 2, 3, \ldots, k$, then it is obvious that $(\psi_1, \psi_2, \ldots, \psi_k)$ is a right standard factorization.

A *deconcatenation* of a string φ is a sequence of strings $(\varphi_1, \varphi_2, \ldots, \varphi_n)$ such that $\varphi_1\varphi_2 \ldots \varphi_n = \varphi$ and $\varphi_i \neq e$ for each $i = 1, 2, \ldots, n$. The number n is said to be the *length* of the deconcatenation $(\varphi_1, \varphi_2, \ldots, \varphi_n)$. If $l(\varphi) = k$, then by the induction one easy proves that there are 2^{k-1} deconcatenations of the string φ. In fact, the right standard factorization is a special case of the deconcatenation.

Theorem 3. *If $(\psi_1, \psi_2, \ldots, \psi_k)$ is a right standard factorization of a terminal* wflb *ψ, then $(d(\psi_1), d(\psi_2), \ldots, d(\psi_k))$ is a deconcatenation of the debracketization $d(\psi)$ of ψ. The mapping assigning in this way deconcatenations to the factorizations is a one-to-one mapping of the set of all right standard factorizations of ψ into the set of all deconcatenations of $d(\psi)$.*

Proof. Using Theorem 2 one can express explicitly the corresponding elements in the considered mapping as follows:

$(\lambda_1 X_1 \varrho_1 \ldots \lambda_{p_1} X_{p_1} \varrho_{p_1}, \lambda_{p_1+1} X_{p_1+1} \varrho_{p_1+1}, \ldots \lambda_{p_2} X_{p_2} \varrho_{p_2}, \ldots$
$\ldots, \lambda_{p_{k-1}+1} X_{p_{k-1}+1} \varrho_{p_{k-1}+1} \lambda_{p_{k-1}+2} X_{p_{k-1}+2} \varrho_{k-1+2} \ldots \lambda_{p_k} X_{p_k} \varrho_{p_k})$ and
$(X_1 X_2 \ldots X_{p_1}, X_{p_1+1} \ldots X_{p_2} \ldots X_{p_{k-1}+1} \ldots X_{p_k})$. Now Theorem 3 is obvious.

Lemma 2. *If $\alpha X \beta'$ and $\alpha'' X \beta$ are the* wflb's *such that $X \in V_T$, $\alpha \in L^\infty$, $\alpha = \alpha'\alpha''$ and $\beta = \beta'\beta''$, then $\alpha' = e$ and β'' is a wflb also.*

Proof. By the definition 1.1 of [1] it is clear what is the pair of the corresponding brackets and that in a wflb are contained either both of the corresponding brackets or none of them. Now, if $a \in L$ is an arbitrary bracket contained in α and if $\bar{a}$ is its corresponding bracket, then $\bar{a}$ must be contained in α and thus in β' also. By the same reasoning a must be contained in α'' and therefore $\alpha'' = \alpha$, i.e. $\alpha' = e$.

Now $\alpha X \beta'$ and $\alpha X \beta'\beta''$ are the wflb's and therefore by Theorem 1 both of them satisfy the bracket condition and are im the basic form. From this it follows that β'' satisfies the bracket condition too and then that β'' is in the basic form. Thus by Theorem 1, β'' is a wflb.

Finally the following definition 1.3 of [1] will be used. The *interior* of a terminal lb φ – written $I(\varphi)$ is the longest wflb ψ such that (i) $d(\varphi) = d(\psi)$, and (ii) there are lb's σ, τ such that $\varphi = \sigma\psi\tau$, if such ψ exists. We shall call σ the *left exterior* of φ (written $E_l(\varphi)$) and τ the *right exterior* of φ ($E_r(\varphi)$). If there is no such ψ we leave $I(\varphi)$, $E_l(\varphi)$ and $E_r(\varphi)$ undefined. We also leave the interior (and exteriors) of labelled bracketing φ undefined if φ is not terminal.

Theorem 4. *Let $\varphi = \psi_i$ for some i, where $(\psi_1, \psi_2, \ldots, \psi_k)$ is a right standard factorization of a terminal* wflb *ψ and let the interior $I(\varphi)$ exist. If $\lambda_1 X_1 \varrho_1 \lambda_2 X_2 \varrho_2 \ldots \ldots \lambda_n X_n \varrho_n$ is the basic form of φ, the following three possibilities can appear : either* $E_l(\varphi) = E(\varphi) = e$ and $I(\varphi) = \varphi$; *in this case φ is a* wflb *itself, but in the remaining two cases it is not; or $E_l(\varphi) = e$, $I(\varphi) = \lambda_1 X_1 \varrho_1 \ldots \lambda_n X_n \varrho_n'$ and $E_r(\varphi) = \varrho_n'' \neq e$ where $\varrho_n = \varrho_n'\varrho_n''$ or $E_r(\varphi) = e$, $I(\varphi) = \lambda_1' X_1 \varrho_1 \ldots \lambda_n X_n \varrho_n$ and $E_l(\varphi) = \lambda_1'' \neq e$ where $\lambda_1 = \lambda_1''\lambda_1'$, i.e. there can never be $E_l(\varphi) \neq e \neq E_r(\varphi)$.*

Proof. If φ is not wflb, then $E_l(\varphi)\, E_r(\varphi) \neq e$ because of $\varphi = E_l(\varphi)\, I(\varphi)\, E_r(\varphi)$. Further it is clear that either $E_l(\varphi) = e$ or there exists λ_1' such that $E_l(\varphi)\, \lambda_1' = \lambda_1$ and similarly either $E_r(\varphi) = e$ or there exists ϱ_1 such that $\varrho_1'\, E_r(\varphi) = \varrho_1$ (obviously it is allowed $\lambda_1' = e$ and $\varrho_1' = e$). Now it is sufficient to exclude the possibility of $E_l(\varphi) \neq e \neq E_r(\varphi)$.

Therefore let us assume $E_l(\varphi) \neq e \neq E_r(\varphi)$. Under this condition $\lambda_1 \neq e \neq \varrho_n$ and we can write $\lambda_1 = a\lambda_1'$ where $a \in L$ and $\varrho_n = \varrho_n'\bar{b}$ where $\bar{b} \in R$.

Now, let $\bar{a}$ denote the bracket corresponding in ψ to the a and let us ask whether $\bar{a}$ belongs to φ or not. If the answer is yes, then there is an integer j such that $1 \leqq j \leqq n$, $\varrho_j = c_1 c_2 \ldots c_p$ where $p \geqq 1$ and $c_h \in R$ for each $h = 1, 2, \ldots, p$ and $\bar{a} = c_m$ for some $1 \leqq m \leqq p$. Thus $\varphi' = \lambda_1 X_1 \varrho_1 \ldots \lambda_j X_j c_1 c_2 \ldots c_m$ is in the basic form and by Lemma 1 it satisfies the bracket condition too. Therefore by Theorem 1 φ' is a terminal wflb. On the other hand, $I(\varphi)$ is also a terminal wflb and $\alpha'\, I(\varphi) = \varphi'\varphi''$ where $\alpha'\alpha'' = \lambda_1$. Therefore by Lemma 2 $\alpha' = e$, i.e. $E_l(\varphi) = e$ what is a contradiciton.

If the answer is no, i.e. $\bar{a}$ does not belong to φ, then the corresponding left bracket b to $\bar{b}$ must belong to φ and by a quite similar reasoning one obtains $E_r(\varphi) = e$, i.e. a contradiction again.

Lemma 3. *Let* $(\lambda_1 X_1 \varrho_1, \lambda_2 X_2 \varrho_2, \ldots, \lambda_n X_n \varrho_n)$ *be the maximal right standard factorization of a terminal* wflb ψ. *Then* $\lambda_i X_i \varrho_i$ *has its interior and if* $\lambda_i = a_p a_{p-1} \ldots \ldots a_1 \neq e$ *where* $a_j \in L$ *for each* $1 \leqq j \leqq p$ *and* $\varrho_i = b_1 b_2 \ldots b_q \neq e$ *where* $b_j \in R$ *for each* $1 \leqq j \leqq q$, *then* $I(\lambda_i X_i \varrho_i) = a_s a_{s-1} \ldots a_1 X_i b_1 b_2 \ldots b_s$ *where* $s = \min(p, q)$. *If either* $\lambda_i = e$ or $\varrho_i = e$, *then* $I(\lambda_i X_i \varrho_i) = X_i$.

Proof. It is clear that $d(\lambda_i X_i \varrho_i) = d(a_s a_{s-1} \ldots a_1 X_i b_1 b_2 \ldots b_s) = d(X_i)$ and therefore one needs to prove that the considered strings are wflb and have the maximal length. It is obvious in the latter case. In the former case when $\lambda_i \neq e\ \varrho_i$ one can ask whether the corresponding bracket $\bar{a}_p$ to a_p belongs to $\lambda_i X_i \varrho_i$ or not.

If the answer is yes, then $\bar{a}_p = b_j$ for some j, $1 \leqq j \leqq q$, and therefore by the definition 1.1 of [1] $a_p a_{p-1} \ldots a_1 X_i b_1 b_2 \ldots b_j$ must be wflb what means $j = p$. In this case evidently $p = \min(p, q) = s$ and also one can easy see that there is no wflb containing $a_s a_{s-1} \ldots a_1 X_i b_1 b_2 \ldots b_s$ and being contained in $\lambda_i X_i \varrho_i$, i.e. $a_s a_{s-1} \ldots a_1 X_i b_1 b_2 \ldots b_s = I(\lambda_i X_i \varrho_i)$.

If the answer is no, then one can ask a similar question whether the corresponding bracket $\bar{b}_q$ to the b_q belongs to $\lambda_i X_i \varrho_i$ or not. One easy sees that the answer must be yes. Then by a similar reasoning one proves the required result again.

(Received September 11th, 1967.)

REFERENCES

[1] S. Peters, R. W. Ritchie: On the generative power of transformational grammars (mimeographed).

[2] N. Chomsky: Aspects of the Theory of Syntax. MIT 1965.

[3] S. C. Kleene: Introduction to Metamatematics, N. Y. 1965.

[4] K. Čulík: On some transformations in context-free grammars and languages. Czech. Math. Journ. *17* (92), (1967), 278–311.

[5] K. Čulík: On the ordered rooted trees used in theory of languages. Théorie des graphes-journées internationales d'étude, Rome, juillet 1966. Dunod, Paris—Gordon and Breach, New York 1967, 69–76.

VÝTAH

Některé věty o závorkování pro transformační gramatiky

KAREL ČULÍK

Je dokázána řada vět týkajících se lineárních zápisů (a jistých jejich rozkladů) frázových ukazatelů užívaných v bezkontextových gramatikách.

Doc. Dr. Karel Čulík, DrSc., Matematický ústav ČSAV, Žitná 25, Praha 1.