The two-sided Dyck Language is not in MCFL(1)

Paul Gallot, Makoto Kanazawa
paul.gallot@ens-cachan.fr, kanazawa@nii.ac.jp

Dyck languages

The classical Dyck language, noted ${D'_1}^*$, is known to be the set of well-parenthesized words over one set of parentheses. This work focuses on the commutative closure of ${D'_1}^*$, called two-sided Dyck language and noted L.

It is simpler to see it as the set of words over $\Sigma = \{a, b\}$ that have the same number of a and b :

$$u \in L \iff |u|_a - |u|_b = 0$$

This language is interesting because it is heavily linked with MIX languages.

Previous work on ${D'_1}^*$

Brigitte Rozoy showed in [1] that :

$$L \notin MCFL(1)$$

The proof consists in assuming *ad absurdum* that there is an MCFG(1) whose language is L, deducing properties for that grammar, and finally exhibiting a counter-example word in L which, according to those properties, can't be constructed using the grammar.

The paper then argues that, since MCFG(1) and EDT0L of finite index are equivalent :

$${D'_1}^* \notin k\text{-order } EDT0L$$

Finally, relying on the paper [2] which proves :

$${D'_1}^* \notin k\text{-order } EDT0L \Rightarrow {D'_1}^* \notin EDT0L$$

Rozoy deduces :

$${D'_1}^* \notin EDT0L$$

Our work focused on replicating that same result for the two-sided Dyck language.

References

[1] Brigitte Rozoy. The dyck language d is not generated by any matrix grammar of finite index. 74:64–89, 07 1987.

[2] Michel Latteux. Substitutions dans les edtol systèmes ultralinéaires. 42:194–260, 08 1979.

Non-branching Multiple Context-Free Grammars

MCFL(1) stands for non-branching Multiple Context-Free Languages, it is the set of languages of non-branching Multiple Context-Free Grammars, or MCFG(1). A MCFG(1) is a tuple $G = (N, \Sigma, P, S)$ where N is the set of non-terminals, Σ is the terminal alphabet, $S \in N$ is the initial non-terminal and P is a set of rules of the form :

$$A(u_1, \dots, u_n) \leftarrow B(x_1, \dots, x_m)$$

where $x_1, \dots, x_m$ are free variables and the word $u_1 u_2 \dots u_n$ is in $\Sigma^* x_1 \Sigma^* x_2 \dots x_m \Sigma^*$; or :

$$A(u_1, \dots, u_n) \leftarrow$$

with $u_1, \dots, u_n \in \Sigma^*$. Non-terminals work as predicates which depend on a fixed number of words $u_1, \dots, u_n$, and a predicate is true is it can be deduced by a sequence of rules. The language $L(G)$ of the grammar G is :

$$L(G) = \{u \mid S(u)\}$$

Main result

The main result of this work is the proof that the two-sided Dyck language L is not an MCFL(1). Similarly to the proof of ${D'_1}^* \notin MCFL(1)$ by Brigitte Rozoy, it relies on an *ad absurdum* hypothesis that there exists a grammar G whose language $L(G)$ is L, and a counter-example word $u \in L \setminus L(G)$.

The main notion used in this proof is that of divergence. The divergence of a word $u \in \Sigma^*$ is the difference between its number of occurrences of a and b :

$$d(u) = |u|_a - |u|_b$$

It gives a simpler definition for the two-sided Dyck language :

$$u \in L \Leftrightarrow d(u) = 0$$

Theorem :
All non-terminal $A \in N$ of the grammar G has a divergence $d(A)$:

$$A(u_1, \dots, u_n) \Rightarrow d(u_1) + \dots + d(u_n) = d(A)$$

We then use the divergence to prove that a counter-example word in L is not in $L(G)$.

Implications

First, since MCFG(1) and EDT0L of finite index are equivalent :

$$L \notin k\text{-order } EDT0L$$

If it can be proven, as showed in [2] for ${D'_1}^*$, that :

$$L \notin k\text{-order } EDT0L \Rightarrow L \notin EDT0L$$

then we have :

$$L \notin EDT0L$$

which would imply that :

The language MIX-4 is not indexed