The language of Lyndon words is not context-free

Jean Berstel
Institut Gaspard Monge (IGM)
Université Marne-la-Vallée
2, rue de la Butte Verte, 93166 Noisy-le-Grand Cédex

Luc Boasson
Laboratoire d'informatique algorithmique: fondements et applications (LIAFA)
Université Denis-Diderot
2, place Jussieu, 75251 Paris Cédex 05

1 Introduction

A word is *primitive* if it is not a proper power of a shorter word. A *Lyndon* word is a primitive word which is minimal under cyclic permutation (for properties of these words, see [the new printing of] Lothaire's book [4]). The status of the languages Q of primitive words and L of Lyndon words with respect to the Chomsky hierarchy appears still to be open (see Petersen [5] for a discussion). It has been shown in [5] that these languages cannot be unambiguous context-free languages (another proof, based on automatic sequences, is given by Allouche [1]). A proof that Q is not context-free would also give a proof that L is not context-free, because Q is the cyclic closure of L and context-free languages are closed under cyclic permutation.

We prove here that the language L of Lyndon words over a two alphabet $\{a, b\}$ is not context-free. This is an easy consequence of Ogden's iteration lemma, and may constitute a good exercise in a course on Formal Languages.

2 Proof

Recall that Ogden's iteration lemma (see e.g. [3]) states that, for every context-free language L there exists an integer N such that, for any word $w \in L$ and for any choice of at least N distiguished positions in w, there exists a factorization

$$w = x\,u\,y\,v\,z$$

such that

(1) either x, u, y each contain at least one distiguished position, or y, v, z each contain at least one distiguished position.

(2) for any $n \geq 0$, the word xu^nyv^nz is in L.

Now, assume that the language L of Lyndon words over $\{a,b\}$ (with $a < b$) is context-free, and consider the word

$$w = a^{N+1}ba^N ba^N$$

where N is the constant of Ogden's lemma. Distinguish the central group of N letters a. Then either the factor u of Ogden's factorization or the factor v (or both) are contained in the central group. Three cases arise:

(i) If both u and v are in the central group (this includes the case where u or v is the empty word), then pumping *up* twice, one gets a word of the form $a^{N+1}ba^m ba^N$ with $m > N+1$ which is not Lyndon.

(ii) If u is in the first group and v is in the second group of a's, then, pumping *down*, on gets a word $w' = a^k ba^m ba^N$ with $k \leq N$ and $m < N$. This word is not Lyndon because it is greater than its conjugate $a^N a^k ba^m b$.

(iii) If u is in the central group and v is in the third group of a's, then pumping *up* twice, one gets a word $w' = a^{N+1}ba^m ba^k$ with $m, k \geq N+2$ which again is not Lyndon.

3 Final remark

There seems not to exist such an easy proof for the set Q of primitive words. Indeed, it has been shown in[2] that the set Q satisfies strong iteration lemmas.

References

[1] J.-P. Allouche. Note on the transcendence of a generation function. In A. Laurinčikas and E. Manstavičius, editors, *Proc. of the Palanga Conference for the* 75*th birthday of Prof. Kubilius*, New Trends in Probab. and Statist., 1997. to appear.

[2] P. Dömösi, S. Horváth, M. Ito, L. Kászonyi, and M. Katsura. Formal languages consisting of primitive words. In *Proc. FCT'93*, number 710 in Lect. Notes Comp. Sci., pages 194–203. Springer-Verlag, 1993.

[3] M. Harrison. *Introduction to Formal Language Theory.* Addison-Wesley, 1978.

[4] M. Lothaire. *Combinatoirics on Words.* Cambridge University Press, 1997. (new printing, first edition by Addison-Wesley, 1983).

[5] H. Petersen. On the language of primitive words. *Theoret. Comput. Sci.*, 161:141–156, 1996.