University of Pennsylvania Working Papers in Linguistics

Volume 10
Issue 1 *Proceedings of the 27th Annual Penn Linguistics Colloquium*

Article 10

1-1-2004

Too many languages satisfy Ogden's Lemma

Marcus Kracht

Too many languages satisfy Ogden's Lemma

Too Many Languages Satisfy Ogden's Lemma

Marcus Kracht

1 Introduction

There are various pumping lemmata for context free languages, the strongest of which is Ogden's Lemma. It is known that it does not fully characterize context free languages. In an attempt to remedy the situation, Manaster-Ramer, Moshier and Zeitman have strengthened this lemma. As we shall show here, there exist non-semilinear languages that satisfy this stronger lemma and also the lesser known interchange lemma, also due to Ogden.

2 Preliminaries

Let A be a finite set, $\mathbb{N}^A$ the set of functions from A to $\mathbb{N}$. Denote by $\overline{0}$ the function that sends every element from A to 0. Further, define $f + g$ by

$$(f+g)(\mathsf{a}) := f(\mathsf{a}) + g(\mathsf{a})$$

Denote the structure $\langle \mathbb{N}^A, \overline{0}, + \rangle$ by $\Omega(A)$. We define nv inductively as follows. $0v := \overline{0}$, $(n+1)v := nv + v$. We write $\mathbb{N}v$ for the set $\{nv : n \in \mathbb{N}\}$. Finally, for $v \in \mathbb{N}^A$ and subsets $V, W \subseteq \mathbb{N}^A$ write $v + W := \{v + w : w \in W\}$ and $V + W := \{v + w : v \in V, w \in W\}$. A subset S of $\mathbb{N}^A$ is called *linear* if it can be written as

$$S = v_0 + \mathbb{N}v_1 + \mathbb{N}v_2 + \cdots + \mathbb{N}v_n \tag{1}$$

for some n (which may be zero, in which case we get the singleton $\{v_0\}$) and some $v_i \in \mathbb{N}^A$ $(i < n)$. A set is called *semilinear* if it is the finite union of semilinear sets. The *Parikh map* from the set A^* to $\Omega(A)$ is defined as follows. If $a \in A$, let e_a be the function that sends a to 1, and every other letter to 0.

$$\pi(\varepsilon) := \overline{0} \tag{2}$$

$$\pi(\vec{x}a) := \pi(\vec{x}) + e_a \tag{3}$$

In what is to follow below, we shall actually write a in place of e_{a}. For a set $L \subseteq A^*$, $\pi[L] := \{\pi(\vec{x}) : \vec{x} \in L\}$. L is called *linear* if $\pi[L]$ is linear, and *semilinear* if $\pi[L]$ is semilinear. A useful theorem is this. Call a subset S of $\mathbb{N}$ *almost periodical* if there are numbers n_0, k, such that for every number $n \geq n_0$: $n \in S$ iff $n + k \in S$.

Theorem 2.1 *Let $A = \{\mathrm{a}\}$. Then the map $\alpha : f \mapsto f(\mathrm{a})$ is an isomorphism from $\Omega(A)$ onto $\langle \mathbb{N}, 0, + \rangle$. Moreover, a subset of $\Omega(A)$ is semilinear iff its image under α is almost periodical.*

Proof. We tacitly identify $\Omega(A)$ with $\mathbb{N}$. A linear subset has the form $n_0 + \mathbb{N}n_1 + \cdots \mathbb{N}n_k$ for some k. Now

$$\begin{aligned} n_0 + \mathbb{N}n_1 p = (n_0 + \mathbb{N}n_1 p) \cup ((n_0 + n_1) + \mathbb{N}n_1 p) + \\ \cdots + ((n_0 + (p-1)n_1) + \mathbb{N}n_1 p) \end{aligned} \tag{4}$$

Hence we can represent the linear set as a union of sets of the form $(n_0 + v) + \mathbb{N}n_1 n_2 \cdots n_k$. Hence, a linear set is almost periodical. It is not hard to see that also a finite union of linear sets is almost periodical, by extending the modulus to the least common multiple of all cyclic vectors involved. Conversely, an almost periodical set S is semilinear. For let n_0 and k be given. Let P the set of numbers $h < k$ such that there is a $n \geq n_0$ with $n \equiv h \pmod{k}$. For convenience we may assume that n_0 is a multiple of k. Then S is the union of the set of members $< n_0$ (which is finite, hence semilinear) and sets of the form $n_0 + h + \mathbb{N}k$, which are linear. □

Corollary 2.2 *There are countably many semilinear languages over a one-letter alphabet.*

We also remark that an intersection of two semilinear subsets of $\Omega(A)$ is again semilinear (Ginsburg and Spanier (1964)). In fact, seen as subsets of $\mathbb{N}^n$, semilinear sets are exactly the ones definable by elementary formulae in Presburger arithmetic (see Ginsburg and Spanier (1966) for a proof). This does not hold for semilinear *languages*, though.

Let $\vec{x}$ be a string. An *occurrence* of a string $\vec{y}$ in $\vec{x}$ is a pair $C = \langle \vec{v}, \vec{w} \rangle$ such that $\vec{x} = \vec{v}\vec{y}\vec{w}$. We say for two occurrences $C = \langle \vec{v}_1, \vec{w}_1 \rangle$ and $D = \langle \vec{v}_2, \vec{w}_2 \rangle$ of strings $\vec{u}_1$ and $\vec{u}_2$ in $\vec{x}$ that C *precedes* D — in symbols $C < D$ — if $\vec{v}_1\vec{u}_1$ is a prefix of $\vec{v}_2$. C *contains* D if $\vec{u}_1$ is a prefix of $\vec{u}_2$ and $\vec{v}_1$ a suffix of $\vec{v}_2$.

If L is a language and $\vec{z} \in L$, a *pumping pair for $\vec{z}$ in L* is a pair $\langle C, D \rangle$ of occurrences of strings $\vec{x}$, $\vec{y}$ such that $C = \langle \vec{u}_1, \vec{v}_1 \rangle$, $D = \langle \vec{u}_2, \vec{v}_2 \rangle$ and

$$\vec{z} = \vec{u}_1 \vec{x} \vec{w} \vec{y} \vec{v}_2 \tag{5}$$

for a certain $\vec{w}$ (so that $\vec{v}_1 = \vec{w}\vec{y}\vec{v}_2$ and $\vec{u}_2 = \vec{u}_1\vec{x}\vec{w}$) and

$$\{\vec{u}_1 \vec{x}^i \vec{w} \vec{y}^i \vec{v}_2 : i \in \omega\} \subseteq L \tag{6}$$

3 Ogden's Lemmata

The following is from Ogden (1968).

Lemma 3.1 (Ogden's Lemma) *Let L be a context free language. Then there exists a number n_L such that for every string $\vec{x} \in L$: if P is a set of at least n_L occurrences of letters in $\vec{x}$ then there exists a pumping pair containing at least one member of P and at most n_L of them.*

If L is a language, let L_n denote the set of strings that are in L and have length n. The following is from Ogden, Ross, and Winkelmann (1985).

Lemma 3.2 (Interchange Lemma) *Let L be a context free language. Then there exists a real number c_L such that for every natural number n and every set $Q \subseteq L_n$ there is $k \geq \ulcorner |Q|/(c_L n^2) \urcorner$, and strings $\vec{x}_i$, $\vec{y}_i$, $\vec{z}_i$, $i < k$, such that*

1. *for all $i < k$: $\vec{x}_i \vec{y}_i \vec{z}_i \in Q$,*
2. *for all $i < j < n$: $\vec{x}_i \vec{y}_i \vec{z}_i \neq \vec{x}_j \vec{y}_j \vec{z}_j$,*
3. *for all $i < i < k$: $|\vec{x}_i| = |\vec{x}_j|$, $|\vec{y}_i| = |\vec{y}_j|$, and $|\vec{z}_i| = |\vec{z}_j|$,*
4. *for all $i < k$: $n > |\vec{x}_i \vec{z}_i| > 0$, and*
5. *for all $i, j < k$: $\vec{x}_i \vec{y}_j \vec{z}_i \in L_n$.*

Note that if the sequence of numbers L_n/n^2 is bounded, then the language satisfies the Interchange Lemma. For assume that for n_0 we have $L_{n_0}/n_0^2 \leq c$. Then set $c_L := \max\{c|L_m|m^2 : m \leq n_0\}$. Then for every subset Q of L_n, $\ulcorner |Q|/(c_L n^2) \urcorner \leq 1$. However, with $k = 1$ the conditions above become empty.

Theorem 3.3 *Every language L where $\lim_{n \to \infty} |L_n|/n^2$ is bounded satisfies the Interchange Lemma. In particular, every one-letter language satisfies the Interchange Lemma.*

4 A Family of Languages that Satisfy Ogden's Lemmata

Let Ω be a subset of ω. Now define

$$L_\Omega = \{\mathtt{a}^m\mathtt{b}^n : m \neq n\} \cup \{\mathtt{a}^n\mathtt{b}^n : n \in \Omega\} \tag{7}$$

Lemma 4.1 *L_Ω is semilinear iff Ω is.*

Proof. Notice that $\pi[L_\Omega]$ has the following decomposition

$$(8) \qquad \begin{aligned} \pi[L_\Omega] = \quad & \mathtt{a} + \mathbb{N}\mathtt{a} + \mathbb{N}(\mathtt{a} + \mathtt{b}) \\ & \cup \mathtt{b} + \mathbb{N}\mathtt{b} + \mathbb{N}(\mathtt{a} + \mathtt{b}) \\ & \cup \{n(\mathtt{a} + \mathtt{b}) : n \in \Omega\} \end{aligned}$$

The first two sets are linear. Suppose now that Ω is semilinear. Then the mapping $n \mapsto n(\mathtt{a} + \mathtt{b})$ actually translates semilinear subsets of $\mathbb{N}$ into semilinear subsets of $\mathbb{N}^{\{\mathtt{a},\mathtt{b}\}}$ and non-semilinear subsets into non-semilinear subsets. So, if Ω is semilinear, then so is L_Ω. Conversely, suppose that L_Ω is semilinear. Then so is its intersection with $\mathbb{N}(\mathtt{a} + \mathtt{b})$. This is $\{n(\mathtt{a} + \mathtt{b}) : n \in \Omega\}$. This set is semilinear iff Ω is. So, Ω is semilinear. □

Corollary 4.2 *There are only countably many Ω for which L_Ω is semilinear.*

Theorem 4.3 *For every Ω, L_Ω satisfies the Interchange Lemma.*

Proof. Notice that for each n the number of strings of L_Ω of length n is $\leq n + 1$. So, $\lim_{n \to \infty} |L_n|/n^2 = 0$. Whence by Theorem 3.3 L satisfies the Interchange Lemma. □

Lemma 4.4 *Suppose that $\vec{x} \in L_\Omega$ contains an unequal number of $\mathtt{a}$'s and $\mathtt{b}$'s. Further, let C be an occurrence of $\mathtt{a}^k$ and D an occurrence of $\mathtt{b}^k$ in $\vec{x}$ for some $k > 0$. Then $\langle C, D\rangle$ is a pumping pair for $\vec{x}$ in L_Ω.*

Proof. For suitable numbers q_0, q_1, q_2 and q_3 we have

$$(9) \qquad \vec{x} = \mathtt{a}^{q_0}\mathtt{a}^k\mathtt{a}^{q_1}\mathtt{b}^{q_2}\mathtt{b}^k\mathtt{b}^{q_3}$$

and

$$(10) \qquad C = \langle \mathtt{a}^{q_0}, \mathtt{a}^{q_1}\mathtt{b}^{q_2}\mathtt{b}^k\mathtt{b}^{q_3}\rangle, \qquad D = \langle \mathtt{a}^{q_0}\mathtt{a}^k\mathtt{a}^{q_1}\mathtt{b}^{q_2}, \mathtt{b}^{q_3}\rangle$$

By assumption, $q_0 + k + q_1 \neq q_2 + k + q_3$. It follows that $q_0 + ik + q_1 \neq q_2 + ik + q_3$ for every $i \in \mathbb{N}$. Now suppose we pump the pair i times. Then we get the string

$$(11) \qquad \vec{y} = \mathtt{a}^{q_0}\mathtt{a}^{ik}\mathtt{a}^{q_1}\mathtt{b}^{q_2}\mathtt{b}^{ik}\mathtt{b}^{q_3}$$

Then $\vec{y} \in L_\Omega$ as well. □

Lemma 4.5 *Suppose that $\vec{x} = \mathtt{a}^m\mathtt{b}^n \in L_\Omega$. If $m > n$, then any occurrence of $\mathtt{a}$ together with any occurrence of the empty string is a pumping pair for $\vec{x}$ in L_Ω. If $m < n$, then any occurrence of $\mathtt{b}$ together with any occurrence of the empty string is a pumping pair for $\vec{x}$ in L_Ω. If $m = n$, then any occurrence of a single letter together with any occurrence of the empty string is a pumping pair for $\vec{x}$ in L_Ω.*

Proof. The proof is as straightforward as the previous. Let us just prove the last case, $m = n$. $\vec{x} = \mathtt{a}^n\mathtt{b}^n$. Take an occurrence C of a letter, say $C = \langle \mathtt{a}^p, \mathtt{a}^{n-p-1}\mathtt{b}^n\rangle$, which is an occurrence of a. Then let $D = \langle \mathtt{a}^n\mathtt{b}^q, \mathtt{b}^{n-q}\rangle$ or $D' = \langle \mathtt{a}^q, \mathtt{a}^{n-q}, \mathtt{b}^n\rangle$ be an occurrence of the empty string. Then $C < D$ and $C < D'$, unless $q \leq p$, in which case $D' < C$. If we iterate zero times, we get the string $\mathtt{a}^{n-1}\mathtt{b}^n$; and if we iterate $i > 1$ times we get $\mathtt{a}^{n+i-1}\mathtt{b}^n$, all of which are in L_Ω. Similarly for occurrences of b. □

Theorem 4.6 *For every Ω, L_Ω satisfies Ogden's Lemma.*

Proof. We show that we can choose $n_L := 2$. Let $\vec{x} \in L_\Omega$. Fix a set P of two occurrences of letters in $\vec{x}$. We assume first that $\vec{x}$ has an unequal number of a's and b's. Case 1. P contains one occurrence of a and one occurrence of b. Then these two occurrences form a pumping pair by Lemma 4.4. Case 2. The occurrences are all occurrences of a. Case 2a. $\vec{x}$ contains a b. We match one of the a with that b. This forms a pumping pair, by Lemma 4.4. Case 2b. $\vec{x}$ contains no b. Then any occurrence of a together with any occurrence of the empty string is a pumping pair for $\vec{x}$ in L_Ω, by Lemma 4.5. So, we now have to look at the case where the string has an equal number of a and b. Then, pick a member of P. Again by Lemma 4.5, that occurrence of a letter together with any occurrence of the empty string is a pumping pair. □

In an unpublished paper Manaster-Ramer, Moshier, and Zeitman (1992) have proposed the following strengthening of Ogden's Lemma. Call a set of pumping pairs $\{\langle C_i, D_i\rangle : i < p\}$ *independent* if for all $i < j < p$ either (1a) $C_i < D_i < C_j < D_j$ or (1b) $C_j < D_j < C_i < D_i$ or (1c) $C_i < C_j < D_j < D_i$ and (2) all pairs can be pumped independently of each other. (If either of the occurrences is an occurrence of the empty string, it is ignored in the condition, as the empty string can be placed anywhere.)

Theorem 4.7 (Multiple Pumping Lemma) *Suppose that L is context free. Then there exists a number p_L such that for any string $\vec{x}$ and a set P of kp_L occurrences of letters in $\vec{x}$ there exist k independent pumping pairs each containing at least one and at most k members of P.*

Theorem 4.8 *For every Ω, L_Ω satisfies the Multiple Pumping Lemma.*

Proof. The pair is not unlike the first one, except that we need to be careful with the selection of pumping strings. We shall show that the claim holds for $p_{L_\Omega} := 2$. Assume that $\vec{z} \in L_\Omega$. Select a set P of occurrences of letters in $\vec{z}$. P is the disjunction of the subset P_A of occurrences of a and the subset P_B of occurrences of b. We may assume that $P_A = \{C_i : i < p\}$ and $P_B = \{F_j : j < q\}$, where $C_i < C_j$ iff $i < j$ and $F_i < F_j$ iff $i < j$. Assume that $|P| = p + q = 2k$. We need to establish at least k independent pumping pairs. Case A. $\vec{z} = \mathtt{a}^m\mathtt{b}^n$ with $m < n$. Case Aa. $|P_A| \geq |P_B|$. Then put $D_i := \langle \mathtt{a}^m\mathtt{b}^{m-i-1}, \mathtt{b}^i\rangle$. $\langle C_i, D_i\rangle$ is a pumping pair, and is independent from $\langle C_j, D_j\rangle$. Namely, it is verified that all occurrences satisfy (1c): while the occurrences of $\mathtt{a}$ are aligned in ascending order, the occurrences of the $\mathtt{b}$ are aligned in descending order. Moreover, the occurrences can be independently pumped. A pair consisting of an occurrence of $\mathtt{a}$ plus an occurrence of $\mathtt{b}$ can be pumped or taken away without affecting the difference between the number of $\mathtt{a}$'s and the number of $\mathtt{b}$'s. Case Ab. $|P_A| < |P_B|$. Here we put $E_i := \langle \mathtt{a}^i, \mathtt{a}^{m-i-1}, \mathtt{b}^n\rangle$, if $i < m$, $E_i := \langle \mathtt{a}^i, \mathtt{a}^{m-i}\mathtt{b}^n\rangle$ otherwise. $\mathcal{P} := \{\langle E_i, F_i\rangle : i < q\}$ is a set of independent pumping pairs. There is just one case that needs attention. That is the case where P_B contains all occurrences of $\mathtt{b}$. In that case, depumping all strings leaves us with the empty string, which is not in L_Ω if $0 \not\in \Omega$. In that case, we put $\mathcal{P} := \{\langle E_i, F_i\rangle : 0 < i < p\}$. (For connoisseurs: we might have to make sure to match at least one of P_A with a P_B in order to keep $p_{L_\Omega} = 2$, but that is a matter of detail.) Case B. $m > n$: Similarly. Case C. $\vec{z} = \mathtt{a}^n\mathtt{b}^n$. Assume that $|P_A| \geq |P_B|$. Then $|P_A| \geq |P|/2$. Put $\mathcal{P} := \{\langle C_i, D\rangle : i < p\}$, where $D = \langle \mathtt{a}^n, \mathtt{b}^n\rangle$ is an occurrence of the empty string. This is a set of independent pumping pairs. If on the other hand $|P_A| < |P_B|$ then $\mathcal{Q} := \{\langle D, F_i\rangle : i < q\}$ is a set of independent pumping pairs. □

It is immediate that there are continuously many languages that satisfy all three conditions above simultaneously, and are semilinear.

Corollary 4.9 *1. There exist continuously many non-semilinear languages that satisfy the Multiple Pumping Lemma and the Interchange Lemma.*

2. There exist continuously many undecidable languages that satisfy the Multiple Pumping Lemma and the Interchange Lemma.

5 Conclusion

This paper shows that to require independent pumping strings of whatever number will not do to characterize CFLs. Assuming that a pumping pair indicates a pair of subconstituents that have the same category, pumping lemmata reveal part of the context-free structure of a string. If the goal is to characterize CFLs exactly, one would have to guarantee that there are plenty of pumping pairs, part of which will be dependent. Such a characterization, though in principle available, might not be easy to use in practice. So far,a practical characterization of CFLs in terms of pumping properties has not been found.

References

Ginsburg, Seymour, and Edwin H. Spanier. 1964. Bounded ALGOL-Like Languages. *Transactions of the American Mathematical Society* 113:333–368.

Ginsburg, Seymour, and Edwin H. Spanier. 1966. Semigroups, Presburger Formulas, and Languages. *Pacific Journal of Mathematics* 16:285–296.

Manaster-Ramer, Alexis, M. Andrew Moshier, and R. Suzanne Zeitman. 1992. An extension of Ogden's Lemma. Manuscript, Wayne State University.

Ogden, R., R. J. Ross, and K. Winkelmann. 1985. An "Interchange Lemma" for Context Free Languages. *SIAM Journal of Computing* 14:410–415.

Ogden, R. W. 1968. A helpful result for proving inherent ambiguity. *Mathematical Systems Sciences* 2:191–194.

Department of Linguistics, UCLA
405 Hilgard Avenue
PO Box 951543
Los Angeles, CA 90095-1543
kracht@humnet.ucla.edu