Ogden's Lemma, Multiple Context-Free Grammars, and the Control Language Hierarchy

Makoto Kanazawa
National Institute of Informatics and SOKENDAI
Japan

Multiple Context-Free Grammars

- Introduced by Seki, Matsumura, Fujii, and Kasami (1987–1991)
- Independently by Vijay-Shanker,Weir, and Joshi (1987)
- **Many** equivalent models
- Often thought to be an adequate formalization of **mildly context-sensitive** grammars (Joshi 1985)

Arising from concerns in computational linguistics.
CFGs are almost good enough for NL grammars, but not quite; a mild extension of CFGs is needed.
Several criteria were put forward as to what constitutes a "mild" extension.

Which properties of CFGs are shared by/ generalize to MCFGs?

Multiple Context-Free Grammars

$A(\alpha_1,\ldots,\alpha_q) \leftarrow B_1(\mathbf{x}_{1,1},\ldots,\mathbf{x}_{1,q_1}),\ldots,B_n(\mathbf{x}_{n,1},\ldots,\mathbf{x}_{n,q_n})$

$n \geq 0,\ q,\ q_i \geq 1,$
$\alpha_k \in (\Sigma \cup \{\ \mathbf{x}_{i,j} \mid i \in [1,n],\ j \in [1,q_i]\ \})^*$
each $\mathbf{x}_{i,j}$ occurs exactly once in $(\alpha_1,\ldots,\alpha_q)$

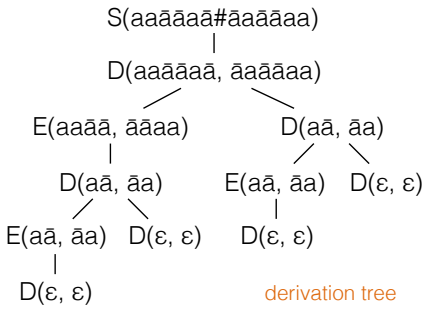
- $q = \dim(A)$ (**dimension** of A)
- $\dim(S) = 1$
- $L(G) = \{\ w \in \Sigma^* \mid G \vdash S(w)\ \}$

It's best to think of an MCFG as a kind of logic program.
Each rule is a definite clause.
Nonterminals are predicates on strings.

$S(\mathbf{x}_1\#\mathbf{x}_2) \leftarrow D(\mathbf{x}_1, \mathbf{x}_2)$
$D(\varepsilon, \varepsilon) \leftarrow$
$D(\mathbf{x}_1\mathbf{y}_1, \mathbf{y}_2\mathbf{x}_2) \leftarrow E(\mathbf{x}_1,\mathbf{x}_2), D(\mathbf{y}_1,\mathbf{y}_2)$
$E(a\mathbf{x}_1\bar{a}, \bar{a}\mathbf{x}_2 a) \leftarrow D(\mathbf{x}_1,\mathbf{x}_2)$

2-MCFG
2-ary branching

$\{\ w\#w^R \mid w \in D_1^*\ \}$

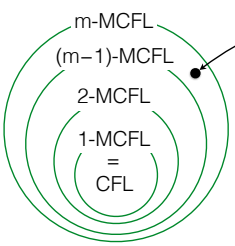


derivation tree

m-MCFG = MCFG with nonterminal dimension not exceeding m
1-MCFG = CFG
Derivation tree for w = proof of S(w)

$S(\mathbf{x}_1\ldots\mathbf{x}_m) \leftarrow A(\mathbf{x}_1,\ldots,\mathbf{x}_m)$
$A(\varepsilon,\ldots,\varepsilon) \leftarrow$
$A(a_1\,\mathbf{x}_1\,a_2,\ldots,a_{2m-1}\,\mathbf{x}_m\,a_{2m}) \leftarrow A(\mathbf{x}_1,\ldots,\mathbf{x}_m)$

non-branching m-MCFG

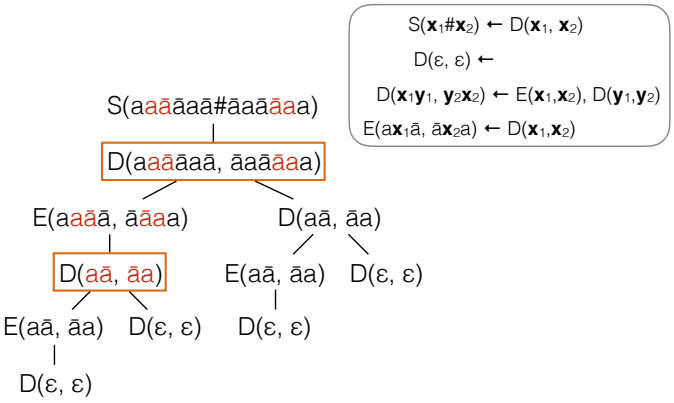


$\{\ a_1^n a_2^n \ldots a_{2m-1}^n a_{2m}^n \mid n \geq 0\ \}$

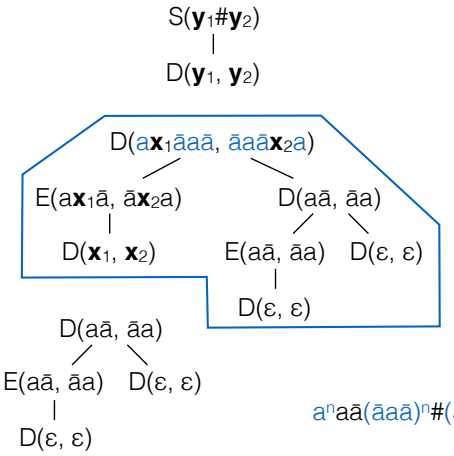
Seki et al. 1991

The languages of MCFGs form an infinite hierarchy.

Pumping



Derivation trees of MCFGs are similar to those of CFGs.
When the same nonterminal occurs twice on the same path of a derivation tree,…



$a^n aā(āaā)^n\#(āaā)^n āaa^n \in L(G)$

You can decompose the derivation tree into three parts, and the middle part can be iterated any number of times, including zero times.
In the overall derivation tree, the variables $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{y}_1$, $\mathbf{y}_2$ are instantiated by …
The number of iterated substrings (factors) larger than two.

Iterative Properties

L is **k-iterative** iff $\exists p\ \forall z \in L(|z| \geq p \Rightarrow$

$\exists u_1 \ldots u_{k+1}\ v_1 \ldots v_k($

$z = u_1 v_1 \ldots u_k v_k u_{k+1} \wedge$

$v_1 \ldots v_k \neq \varepsilon \wedge$

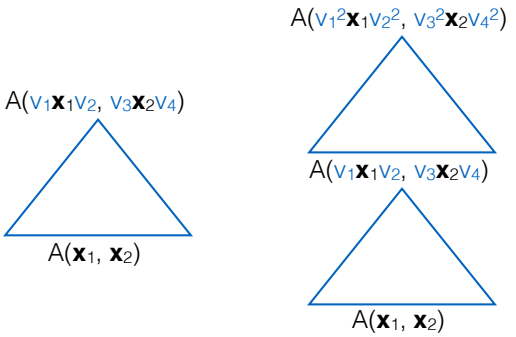
$\forall n \geq 0 (u_1 v_1^n \ldots u_k v_k^n u_{k+1} \in L))$

$L \in CFL \Rightarrow L$ is 2-iterative

$L \in$ m-MCFL $\Rightarrow L$ is 2m-iterative **?**

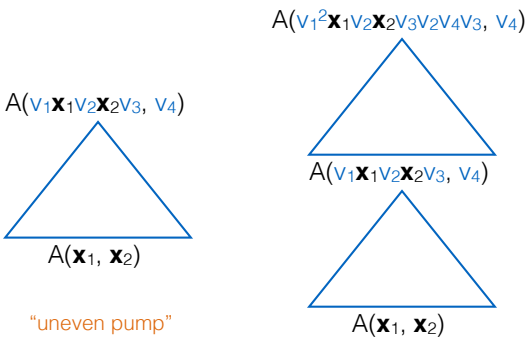
For MCFGs, need to consider a generalized form of the condition of the puming lemma.
Not straightforward; open question for a long time.

Difficulty with Pumping



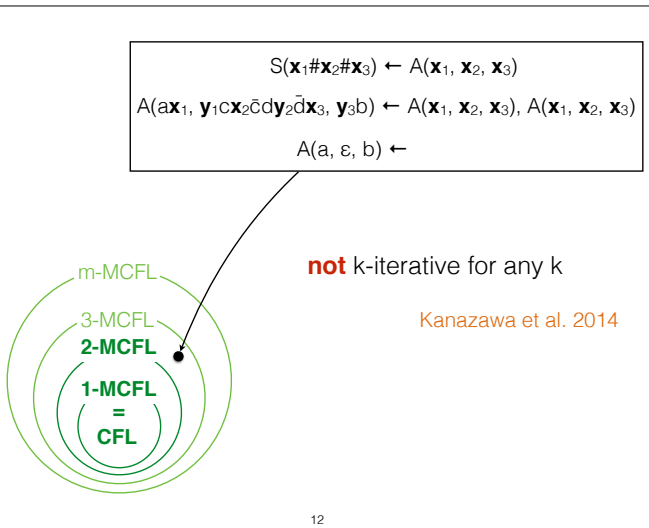
The middle part of the derivation tree may look like this.

Difficulty with Pumping



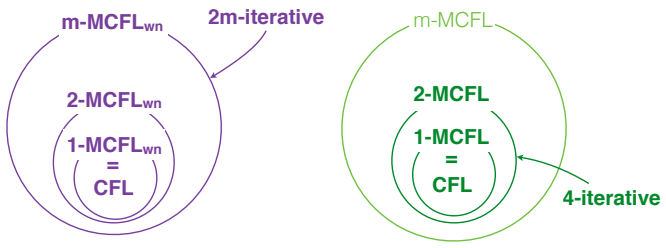
"uneven pump"

Or like this.



The pumping lemma fails for 3-MCFGs.

Pumping Lemma for Subclasses



well-nested MCFGs

Kanazawa 2009

Pumping possible for special cases. Well-nested MCFGs.

Well-Nestedness

$\{ w\#w^R \mid w \in D_1^* \}$ $\{ w\#w \mid w \in D_1^* \}$

$S(\mathbf{x}_1\#\mathbf{x}_2) \leftarrow D(\mathbf{x}_1, \mathbf{x}_2)$

$D(\varepsilon, \varepsilon) \leftarrow$

$D(\mathbf{x}_1\mathbf{y}_1, \mathbf{y}_2\mathbf{x}_2) \leftarrow E(\mathbf{x}_1,\mathbf{x}_2), D(\mathbf{y}_1,\mathbf{y}_2)$

$E(a\mathbf{x}_1\bar{a}, \bar{a}\mathbf{x}_2a) \leftarrow D(\mathbf{x}_1,\mathbf{x}_2)$

$S(\mathbf{x}_1\#\mathbf{x}_2) \leftarrow D(\mathbf{x}_1, \mathbf{x}_2)$

$D(\varepsilon, \varepsilon) \leftarrow$

$D(\mathbf{x}_1\mathbf{y}_1, \mathbf{x}_2\mathbf{y}_2) \leftarrow E(\mathbf{x}_1,\mathbf{x}_2), D(\mathbf{y}_1,\mathbf{y}_2)$

$E(a\mathbf{x}_1\bar{a}, a\mathbf{x}_2\bar{a}) \leftarrow D(\mathbf{x}_1,\mathbf{x}_2)$

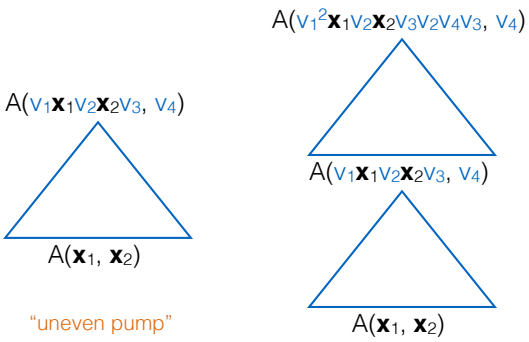
well-nested **non-well-nested**

$\{ w\#w \mid w \in D_1^* \} \notin \text{MCFL}_{wn}$

Kanazawa and Salvati 2010

Has a natural equivalent characterization: yCFT$_{sp}$

Difficulty with Pumping



"uneven pump"

Pumping not easy to prove even form well-nested MCFGs: this situation can still arise.

$S(\mathbf{x}_1\mathbf{x}_2) \leftarrow A(\mathbf{x}_1, \mathbf{x}_2)$
$A(a\mathbf{x}_1b\mathbf{x}_2c, d) \leftarrow A(\mathbf{x}_1, \mathbf{x}_2)$
$A(\varepsilon, \varepsilon) \leftarrow$

non-branching ⊆ well-nested

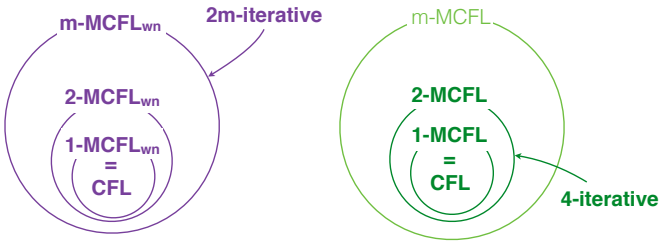
i=0	i=1	i=2	i=3
S(ε)	S(abcd)	S(aabcbdcd)	S(aaabcbdcbdc,d)
A(ε, ε)	A(abc, d)	A(aabcbdc,d)	A(aaabcbdcbdc,d)
	A(ε, ε)	A(abc, d)	A(aabcbdc,d)
		A(ε, ε)	A(abc, d)
			A(ε, ε)

$\{ \varepsilon \} \cup \{ a^{i-1}abc(bdc)^{i-1}d \mid i \geq 1 \}$

16

A very simple example.
The only choice you can make is the number of times you use the second rule.
Actually 2-iterative, but no straightforward connection between the iterated substrings and parts of derivation trees.

Pumping Lemma for Subclasses



Kanazawa 2009, by grammar splitting and transformation

What about **Ogden's Lemma**?

17

My proof of the pumping lemma for m-MCFL$_{wn}$ and 2-MCFL is not straightforward.

Ogden's Lemma for CFL

$L \in CFL \Rightarrow$

$\exists p\ \forall z \in L$(at least p positions of z are marked $\Rightarrow$

$\exists u_1u_2u_3\ v_1v_2$(

$z = u_1v_1u_2v_2u_3 \wedge$

(u_1, v_1, u_2 each contain a marked position $\vee$

u_2, v_2, u_3 each contain a marked position) $\wedge$

$v_1u_2v_2$ contains no more than p marked positions $\wedge$

$\forall n \geq 0(u_1v_1^nu_2v_2^nu_3 \in L)$)

Ogden 1968

18

L has the **weak Ogden property** iff

$\exists p\ \forall z \in L$(at least p positions of z are marked $\Rightarrow$

$\exists k \geq 1\ \exists u_1 \ldots u_{k+1}\ v_1 \ldots v_k$(

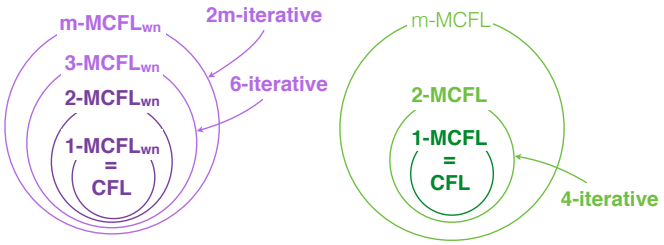
$z = u_1 v_1 \ldots u_k v_k u_{k+1} \wedge$

$\exists i$(v_i contains a marked position) $\wedge$

$\forall n \geq 0 (u_1 v_1^n \ldots u_k v_k^n u_{k+1} \in L)$)

There are various ways of generalizing Ogden's lemma suitable for MCFGs. At least this much should be implied.

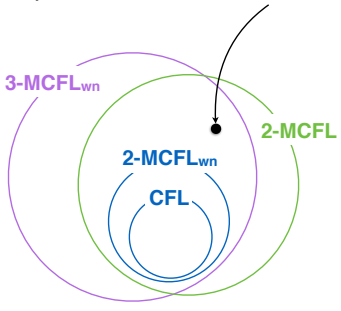
The Failure of Ogden's Lemma



The weak Ogden property **fails** for 3-MCFL$_{wn}$ and 2-MCFL.

This is the first new result in this talk.

$\{ a^{i_1} b^{i_0} \$ a^{i_2} b^{i_1} \$ a^{i_3} b^{i_2} \$ \ldots \$ a^{i_n} b^{i_{n-1}} \mid n \geq 3,\ i_0, \ldots, i_n \geq 0 \}$



A language for which the weak Ogden property fails.

$A(\varepsilon) \leftarrow$
$A(b\mathbf{x}_1) \leftarrow A(\mathbf{x}_1)$
$B(\mathbf{x}_1, \varepsilon) \leftarrow A(\mathbf{x}_1)$
$B(a\mathbf{x}_1, b\mathbf{x}_2) \leftarrow B(\mathbf{x}_1, \mathbf{x}_2)$
$C(\mathbf{x}_1, \mathbf{x}_2, \varepsilon) \leftarrow B(\mathbf{x}_1, \mathbf{x}_2)$
$C(\mathbf{x}_1, a\mathbf{x}_2, b\mathbf{x}_3) \leftarrow C(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$
$C(\mathbf{x}_1\$\mathbf{x}_2, \mathbf{x}_3, \varepsilon) \leftarrow C(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$
$D(\mathbf{x}_1\$\mathbf{x}_2, \mathbf{x}_3) \leftarrow C(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$
$D(\mathbf{x}_1, a\mathbf{x}_2) \leftarrow D(\mathbf{x}_1, \mathbf{x}_2)$
$S(\mathbf{x}_1\$\mathbf{x}_2) \leftarrow D(\mathbf{x}_1, \mathbf{x}_2)$

non-branching 3-MCFG

$A(\varepsilon) \leftarrow$
$A(b\mathbf{x}_1) \leftarrow A(\mathbf{x}_1)$
$B(\mathbf{x}_1, \varepsilon) \leftarrow A(\mathbf{x}_1)$
$B(a\mathbf{x}_1, b\mathbf{x}_2) \leftarrow B(\mathbf{x}_1, \mathbf{x}_2)$
$C(\varepsilon, \varepsilon) \leftarrow$
$C(a\mathbf{x}_1, b\mathbf{x}_2) \leftarrow C(\mathbf{x}_1, \mathbf{x}_2)$
$D(\mathbf{x}_1\$\mathbf{y}_1\mathbf{x}_2, \mathbf{y}_2) \leftarrow B(\mathbf{x}_1, \mathbf{x}_2), C(\mathbf{y}_1, \mathbf{y}_2)$
$D(\mathbf{x}_1\$\mathbf{y}_1\mathbf{x}_2, \mathbf{y}_2) \leftarrow D(\mathbf{x}_1, \mathbf{x}_2), C(\mathbf{y}_1, \mathbf{y}_2)$
$E(\mathbf{x}_1, \mathbf{x}_2) \leftarrow D(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$
$E(\mathbf{x}_1, a\mathbf{x}_2) \leftarrow E(\mathbf{x}_1, \mathbf{x}_2)$
$S(\mathbf{x}_1\$\mathbf{x}_2) \leftarrow E(\mathbf{x}_1, \mathbf{x}_2)$

2-MCFG

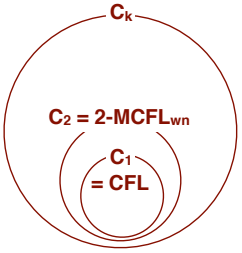
$\{ a^{i_1}b^{i_0}\$a^{i_2}b^{i_1}\$a^{i_3}b^{i_2}\$\ldots\$a^{i_n}b^{i_{n-1}} \mid n \geq 3, i_0,\ldots,i_n \geq 0 \}$

$\{ a^{i_1}b^{i_0}\$a^{i_2}b^{i_1}\$a^{i_3}b^{i_2}\$\ldots\$a^{i_n}b^{i_{n-1}} \mid n \geq 3, i_0,\ldots,i_n \geq 0 \}$
is **2-iterative**

$a\$a^2b\$a^3b^2\$\ldots\$a^{p+1}b^p$

Mark the positions of $.

Weir's (1992) Control Language Hierarchy



$C_k \subseteq 2^{k-1}$-MCFL

Kanazawa and Salvati 2007

Generalization of Ogden's lemma

Palis and Shende 1995

Subclasses of MCFL that are known to have an Ogden property.

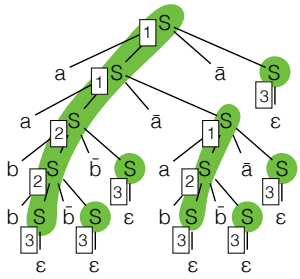
Control Grammars

G:
1: S → aSāS
2: S → bSb̄S
3: S → ε

CFG with child selection

$K = \{ 1^n 2^n 3 \mid n \geq 0 \}$

control set



$L(G, K) = D_2^* \cap (\{ a^n b^n \mid n \geq 1 \}\ \bar{b}\ \{\bar{a},\bar{b}\}^*)^*$

Languages in each level of the control language hierarchy are given by "control grammars".

Control Language Hierarchy

$C_1 = \text{CFL}$

$C_{k+1} = \{ L(G, K) \mid K \in C_k \}$

Ogden's Lemma for C_k

$L \in C_k \Rightarrow$

$\exists p\ \forall z \in L$(at least p positions of z are marked $\Rightarrow$

$\exists u_1 \ldots u_{2k+1}\ v_1 \ldots v_{2k}$(

$z = u_1 v_1 \ldots u_{2k} v_{2k} u_{2k+1} \wedge$

$\exists i$(u_i, v_i, u_{i+1} each contain a marked position) $\wedge$

$v_{2k-1} u_{2k-1} v_{2k-1+1}$ contains no more than p marked positions $\wedge$

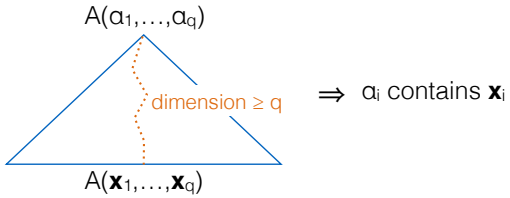
$\forall n \geq 0 (u_1 v_1^n u_2 v_2^n \ldots u_{2k} v_{2k}^n u_{2k+1} \in L)$)

Palis and Shende 1995

- k = 1 gives Ogden's (1968) original lemma.

It is quite straightforward to prove an Ogden's lemma for C_k.

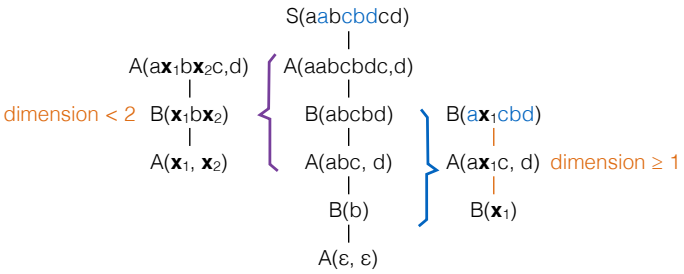
Proper MCFGs



Approach this existing result from the MCFG formalism.
Sufficient condition for an Ogden property.

$S(\mathbf{x}_1\mathbf{x}_2) \leftarrow A(\mathbf{x}_1, \mathbf{x}_2)$
$A(a\mathbf{x}_1c, d) \leftarrow B(\mathbf{x}_1)$
$B(\mathbf{x}_1b\mathbf{x}_2) \leftarrow A(\mathbf{x}_1, \mathbf{x}_2)$
$A(\varepsilon, \varepsilon) \leftarrow$

proper 2-MCFG



Slight variation of an earlier example.

Ogden's Lemma for m-MCFL$_{prop}$

L ∈ m-MCFLprop ⇒

∃p ∀z ∈ L(at least p positions of z are marked ⇒

∃$u_1 \ldots u_{2m+1}$ $v_1 \ldots v_{2m}$(
$z = u_1v_1 \ldots u_{2m}v_{2m}u_{2m+1}$ ∧
∃i(u_i, v_i, u_{i+1} each contain a marked position) ∧
$v_1u_2v_2, v_3u_4v_4, \ldots, v_{2m-1}u_{2m}v_{2m}$ together contain no more than p marked positions ∧
∀n ≥ 0($u_1v_1^n u_2v_2^n \ldots u_{2m}v_{2m}^n u_{2m+1}$ ∈ L))

- m = 1 gives Ogden's (1968) original lemma.

Constrain all of $v_1,\ldots,v_{2m}$

$C_k \subseteq 2^{k-1}\text{-MCFL}_{\text{prop}}$

G: 1: S → aSāS
2: S → bSb̄S
3: S → ε

l(1) = a, r(1) = āS
l(2) = b, r(2) = b̄S homomorphisms
l(3) = ε, r(3) = ε

$K = \{ 1^n 2^n 3 \mid n \geq 0 \}$

$\langle (1,R), l, r \rangle(K) = \{ l(w)r(w^R) \mid w \in K \}$
homomorphic replication

K

$K(\mathbf{x}_1 3) \leftarrow E(\mathbf{x}_1)$
$E(1\mathbf{x}_1 2) \leftarrow E(\mathbf{x}_1)$
$E(\varepsilon) \leftarrow$

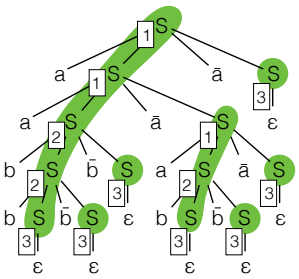
m-MCFG

$K'(\mathbf{x}_1\mathbf{x}_2) \leftarrow K(\mathbf{x}_1, \mathbf{x}_2)$
$K(\mathbf{x}_1 l(3), r(3)\mathbf{x}_2) \leftarrow E(\mathbf{x}_1, \mathbf{x}_2)$
$E(l(1)\mathbf{x}_1 l(2), r(2)\mathbf{x}_2 r(1)) \leftarrow E(\mathbf{x}_1, \mathbf{x}_2)$
$E(\varepsilon, \varepsilon) \leftarrow$

2m-MCFG

31

The earlier result is subsumed by the present result.



l(11223)r(32211) = aabbb̄Sb̄Sāb̄Sā
l(3)r(3) = ε
l(123)r(321) = abb̄Sā

$\langle (1,R), l, r \rangle(K) = \{ l(w)r(w^R) \mid w \in K \}$

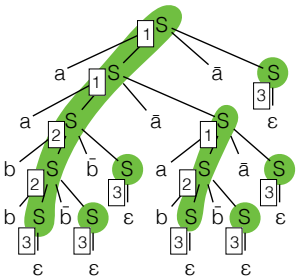
$K'(\mathbf{x}_1\mathbf{x}_2) \leftarrow K(\mathbf{x}_1, \mathbf{x}_2)$
$K(\mathbf{x}_1 l(3), r(3)\mathbf{x}_2) \leftarrow E(\mathbf{x}_1, \mathbf{x}_2)$
$E(l(1)\mathbf{x}_1 l(2), r(2)\mathbf{x}_2 r(1)) \leftarrow E(\mathbf{x}_1, \mathbf{x}_2)$
$E(\varepsilon, \varepsilon) \leftarrow$

=

$K'(\mathbf{x}_1\mathbf{x}_2) \leftarrow K(\mathbf{x}_1, \mathbf{x}_2)$
$K(\mathbf{x}_1, \mathbf{x}_2) \leftarrow E(\mathbf{x}_1, \mathbf{x}_2)$
$E(a\mathbf{x}_1 b, \bar{b}S\mathbf{x}_2\bar{a}S) \leftarrow E(\mathbf{x}_1, \mathbf{x}_2)$
$E(\varepsilon, \varepsilon) \leftarrow$

32

Generates strings with nonterminals.



$\langle (1,R), l, r \rangle(K) = \{ l(w)r(w^R) \mid w \in K \}$

$K'(\mathbf{x}_1\mathbf{x}_2) \leftarrow K(\mathbf{x}_1, \mathbf{x}_2)$
$K(\mathbf{x}_1, \mathbf{x}_2) \leftarrow E(\mathbf{x}_1, \mathbf{x}_2)$
$E(a\mathbf{x}_1 b, \bar{b}S\mathbf{x}_2\bar{a}S) \leftarrow E(\mathbf{x}_1, \mathbf{x}_2)$
$E(\varepsilon, \varepsilon) \leftarrow$

$S \leftarrow K'$ iterated substitution
$\cap \Sigma^*$

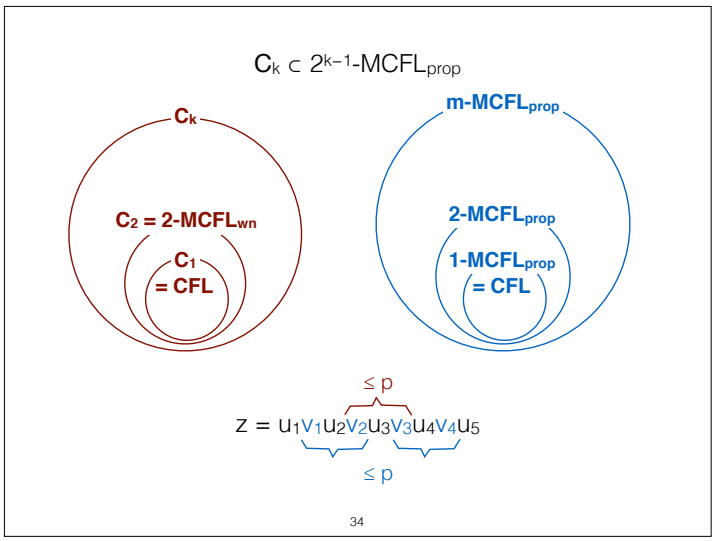
L(G, K)

$K'(\mathbf{x}_1\mathbf{x}_2) \leftarrow K(\mathbf{x}_1, \mathbf{x}_2)$
$K(\mathbf{x}_1, \mathbf{x}_2) \leftarrow E(\mathbf{x}_1, \mathbf{x}_2)$
$E(a\mathbf{x}_1 b, \bar{b}\mathbf{y}\mathbf{x}_2\bar{a}\mathbf{z}) \leftarrow E(\mathbf{x}_1, \mathbf{x}_2), K'(\mathbf{y}), K'(\mathbf{z})$
$E(\varepsilon, \varepsilon) \leftarrow$

2m-MCFG

33

The construction doubles the dimension, preserves properness.



The different requirement shows the properness of the inclusion.

Summary

- Pumping doesn't imply Ogden: There is no Ogden-like theorem for $3\text{-MCFL}_{wn} \cap 2\text{-MCFL}$
- There is a natural Ogden's lemma for $m\text{-MCFL}_{prop}$
- Covers Weir's control language hiearchy