ISSN 1346-5597

NII National Institute of Informatics

NII Technical Report

Weak vs. Strong Finite Context and Kernel Properties

Makoto Kanazawa

NII-2016-006E
July 2016

URL: http://www.nii.ac.jp/ Tokyo 101-8430, JAPAN +81-3-4212-2000 FAX: +81-3-3556-1916

Weak vs. Strong Finite Context and Kernel Properties

Makoto Kanazawa
National Institute of Informatics and SOKENDAI

Let $L \subseteq \Sigma^*$ be given. For $C \subseteq \Sigma^* \times \Sigma^*$, we put

$$C^{\langle L|} = \{ x \in \Sigma^* \mid uxv \in L \text{ for all } (u, v) \in C \}.$$

For $K \subseteq \Sigma^*$, we put

$$K^{|L\rangle} = \{ (u, v) \in \Sigma^* \times \Sigma^* \mid uKv \subseteq L \}.$$

When the language L is understood from context, these are simply written $C^{\triangleleft}$ and $K^{\triangleright}$. A subset K of Σ^* is *closed* if $K = K^{\triangleright\triangleleft}$. Equivalently, K is closed if and only if there exists a $C \subseteq \Sigma^* \times \Sigma^*$ such that $K = C^{\triangleleft}$. (The notion of a closed set as well as the operations $\triangleleft$ and $\triangleright$ are always relative to the given language L.)

We allow a context-free grammar (CFG) to have multiple initial nonterminals. If X is a nonterminal of a context-free grammar G over the terminal alphabet Σ, we write $L(G, X)$ for $\{ x \in \Sigma^* \mid X \Rightarrow_G^* x \}$, the set of terminal strings derivable from X. If $\mathscr{I}$ is the set of initial nonterminals of a context-free grammar G, then $L(G) = \bigcup_{X \in \mathscr{I}} L(G, X)$.

Let $G = (N, \Sigma, P, \mathscr{I})$ be a CFG, and let the operators $\triangleleft$ and $\triangleright$ be understood relative to $L(G)$. We say that G has the *weak finite context property (FCP)* if for each nonterminal X of G, there is a finite $C_X \subseteq \Sigma^* \times \Sigma^*$ such that $L(G, X)^{\triangleright\triangleleft} = C_X^{\triangleleft}$. If $L(G, X) = C_X^{\triangleleft}$ for each nonterminal X, then G has the *strong FCP*. We say that G has the *weak finite kernel property (FKP)* if for each nonterminal X of G, there is a finite set $K_X \subseteq \Sigma^*$ such that $L(G, X)^{\triangleright\triangleleft} = K_X^{\triangleright\triangleleft}$. If $L(G, X) = K_X^{\triangleright\triangleleft}$ for each nonterminal X, then G has the *strong FKP*. Clearly, a CFG G has the strong FCP (FKP) if and only if G has the weak FCP (FKP) and moreover $L(G, X)$ is a closed set for each nonterminal X of G.

What Clark (2010) called the finite context property was what we here call the strong finite context property. The weaker definition was adopted by Yoshinaka (2011) and Leiß (2014). According to Yoshinaka (2015), it has been an open question whether or not every language that has a CFG satisfying the weak FCP has a CFG satisfying the strong FCP. The present note settles this question in the negative, and establishes a similar separation between the two variants of the FKP.

We write x^R for the reversal of a string x, and $|x|_a$ for the number of occurrences of a symbol a in x. Let

$$\begin{aligned}
\Sigma &= \{a, b, c, d, e, \#, \$\},\\
L_* &= L_1 \cup L_2 \cup L_3,\\
L_1 &= \{\, w_1 \# w_2 \# \dots \# w_n \$ w_n^R \dots w_2^R w_1^R \mid n \geq 1, w_1, \dots, w_n \in \{a, b\}^* \,\},\\
L_2 &= \{\, w y c^i d^i e^j z \mid w, z \in \{a, b\}^*, y \in (\#\{a, b\}^*)^*, i, j \geq 0, |w|_a \geq |w|_b \,\},\\
L_3 &= \{\, w y c^i d^j e^j z \mid w, z \in \{a, b\}^*, y \in (\#\{a, b\}^*)^*, i, j \geq 0, |w|_a \leq |w|_b \,\}.
\end{aligned}$$

Lemma 1. *Every CFG G for L_* has a nonterminal X such that $L(G, X)$ is not a closed set.*

Proof. Let G be a CFG for L_*. By applying Ogden's (1968) lemma[1] to a derivation tree of a sufficiently long string in L_1 of the form

$$a^p b^p \# a^p \$ a^p b^p a^p,$$

we obtain

$$\begin{aligned}
S_1 &\Rightarrow^+ a^{m_1} A a^{l_1},\\
A &\Rightarrow^+ a^{n_1} A a^{n_1},\\
A &\Rightarrow^+ a^{m_2} b^{m_3} B b^{l_3} a^{l_2},\\
B &\Rightarrow^+ b^{n_2} B b^{n_2},\\
B &\Rightarrow^+ b^{m_4} \# a^{m_5} D a^{l_5} b^{l_4},\\
D &\Rightarrow^+ a^{m_6} \$ a^{l_6},
\end{aligned}$$

for some $n_1, n_2, m_1, m_2, m_3, m_4, m_5, m_6, l_1, l_2, l_3, l_4, l_5, l_6 \geq 1$ such that $m_1 + n_1 + m_2 = m_3 + n_2 + m_4 = m_5 + m_6 = l_1 + n_1 + l_2 = l_3 + n_2 + l_4 = l_5 + l_6 = p$, where S_1 is an initial nonterminal. We show that $L(G, D)$ is not a closed set.

Let $(u, v) \in L(G, D)^{\triangleright}$. Then $u a^{m_6} \$ a^{l_6} v \in L_*$. Since $y \$ z \in L_*$ implies $y c^i d^i e^i z \in L_*$ for every $y, z \in \Sigma^*$ and $i \geq 0$, we have $u a^{m_6} c^i d^i e^i a^{l_6} v \in L_*$ for every $i \geq 0$. This shows

$$\{\, a^{m_6} c^i d^i e^i a^{l_6} \mid i \geq 0 \,\} \subseteq L(G, D)^{\triangleright}. \tag{1}$$

On the other hand, since

$$S_1 \Rightarrow^* a^{m_1} (a^{n_1})^i a^{m_2} b^{m_3} (b^{n_2})^j b^{m_4} \# a^{m_5} D a^{l_5} b^{l_4} (b^{n_2})^j b^{l_3} a^{l_2} (a^{n_1})^i a^{l_1}$$

for all $i, j \geq 0$, there are $w, w', z, z' \in \{a, b\}^*$ such that

$$\begin{aligned}
S_1 &\Rightarrow^* w \# a^{m_5} D z\\
S_1 &\Rightarrow^* w' \# a^{m_5} D z'
\end{aligned} \tag{2}$$

[1] It is clear from Ogden's proof that the lemma is really about one particular derivation tree of a context-free grammar. If p is the constant of Ogden's lemma for G, we obtain the required decomposition of the derivation tree by first marking the initial a^p, then the b^p preceding $\#$, and then the a^p immediately following $\#$.

$$|w|_a > |w|_b,$$
$$|w'|_a < |w'|_b.$$

Now suppose $a^{m_6}c^i d^j e^k a^{l_6} \in L(G,D)^{\triangleright\triangleleft}$. Since (2) implies

$$(w\#a^{m_5}, z) \in L(G,D)^{\triangleright}$$
$$(w'\#a^{m_5}, z') \in L(G,D)^{\triangleright},$$

we must have

$$w\#a^{m_5}a^{m_6}c^i d^j e^k a^{l_6} z \in L_2,$$
$$w'\#a^{m_5}a^{m_6}c^i d^j e^k a^{l_6} z \in L_3.$$

It follows that

$$a^{m_6}c^i d^j e^k a^{l_6} \in L(G,D)^{\triangleright\triangleleft} \text{ only if } i = j = k. \tag{3}$$

By (1) and (3),

$$L(G,D)^{\triangleright\triangleleft} \cap a^{m_6}c^* d^* e^* a^{l_6} = \{\, a^{m_6}c^i d^i e^i a^{l_6} \mid i \geq 0 \,\},$$

which implies that $L(G,D)^{\triangleright\triangleleft}$ is not context-free. Therefore, $L(G,D) \neq L(G,D)^{\triangleright\triangleleft}$ and $L(G,D)$ is not a closed set. $\square$

The above lemma implies that L_* has no CFG that has either the strong FCP or the strong FKP.

Lemma 2. *There is a CFG for L_* that has both the weak FCP and the weak FKP.*

Proof. Let G be the following CFG, where S_1, S_2, S_3 are the initial nonterminals.

$$\begin{aligned}
S_1 &\to \$ \mid aS_1a \mid bS_1b \mid \#S_1 \\
Q &\to \varepsilon \mid aQbQ \mid bQaQ \\
F &\to Q\# \mid Fa \mid Fb \mid F\# \\
H &\to \varepsilon \mid cHd \\
E &\to \varepsilon \mid Ee \\
C &\to \varepsilon \mid cC \\
J &\to \varepsilon \mid dJe \\
S_2 &\to HE \mid FS_2 \mid QS_2 \mid aS_2 \mid S_2a \mid S_2b \\
S_3 &\to CJ \mid FS_3 \mid QS_3 \mid bS_3 \mid S_3a \mid S_3b
\end{aligned}$$

We have

$$\begin{aligned}
L(G,S_1) &= L_1, \\
L(G,S_1)^{\triangleright} &= \{\, (w_1\#w_2\#\ldots\#w_n, w_n^R \ldots w_2^R w_1^R) \mid n \geq 1, w_1, \ldots, w_n \in \{a,b\}^* \},
\end{aligned}$$

$$
\begin{aligned}
L(G, S_1)^{\rhd\lhd} &= L_1 \cup \{\, yc^i d^i e^i z \mid y \in \{a, b, \#\}^*, z \in \{a, b\}^*, i \geq 0 \,\} \\
&= \{(a\#, a), (b\#, b)\}^{\lhd} \\
&= \{\$\}^{\rhd\lhd}, \\
L(G, Q) &= \{\, w \in \{a, b\}^* \mid |w|_a = |w|_b \,\} \\
&= \{(\varepsilon, \#cd), (a, b\#de)\}^{\lhd} \\
&= \{ab\}^{\rhd\lhd}, \\
L(G, F) &= \{\, w\#y \mid w \in \{a, b\}^*, |w|_a = |w|_b, y \in \{a, b, \#\}^* \,\} \\
&= \{(\varepsilon, cd), (\varepsilon, ade)\,\}^{\lhd} \\
&= \{\#\}^{\rhd\lhd}, \\
L(G, H) &= \{\, c^i d^i \mid i \geq 0 \,\} \\
&= \{(a\#c, d)\}^{\lhd} \\
&= \{\varepsilon, cd\}^{\rhd\lhd}, \\
L(G, E) &= e^* \\
&= \{(a\#cd, e)\}^{\lhd} \\
&= \{\varepsilon, e\}^{\rhd\lhd}, \\
L(G, C) &= c^* \\
&= \{(b\#c, de)\}^{\lhd} \\
&= \{\varepsilon, c\}^{\rhd\lhd}, \\
L(G, J) &= \{\, d^i e^i \mid i \geq 0 \,\} \\
&= \{(b\#d, e)\}^{\lhd} \\
&= \{\varepsilon, de\}^{\rhd\lhd}, \\
L(G, S_2) &= L_2 \\
&= \{(\varepsilon, \varepsilon), (a\#, b)\}^{\lhd} \\
&= \{cda\}^{\rhd\lhd}, \\
L(G, S_3) &= L_3 \\
&= \{(\varepsilon, \varepsilon), (b\#, a)\}^{\lhd} \\
&= \{\#de\}^{\rhd\lhd}.
\end{aligned}
$$

(Recall that $K = C^{\lhd}$ implies $K = K^{\rhd\lhd}$.) This shows that G has both the weak FCP and the weak FKP.[2] □

Theorem 3. *There is a language that is generated by a CFG that has both the weak FCP and the weak FKP but is not generated by any CFG that has either the strong FCP or the strong FKP.*

A CFG G is said to have the *weak k-FCP* if for each nonterminal X, there is a $C_X \subseteq \Sigma^* \times \Sigma^*$ such that $|C_X| \leq k$ and $L(G, X)^{\rhd\lhd} = C_X^{\lhd}$. Similarly, G is

[2] Sometimes the definition of the weak FCP requires that $L(G, X)^{\rhd\lhd} = C_X^{\lhd}$ for some finite subset C_X of $\{\, (u, v) \mid S \Rightarrow^* uXv$ for some initial nonterminal $S \,\}$ (Kanazawa and Yoshinaka, to appear). This property is satisfied by the present grammar.

said to have the *weak k-FKP* if for each nonterminal X, there is a $K_X \subseteq \Sigma^*$ such that $|K_X| \leq k$ and $L(G,X)^{\triangleright\triangleleft} = K_X^{\triangleright\triangleleft}$. (The *strong* k-FCP (k-FKP) may be defined similarly.) The CFG in the proof of Lemma 2 has both the weak 2-FCP and the weak 2-FKP.

The language L_* used in this note has the following two notable properties:

(i) *L_* is inherently ambiguous.* In particular, using Ogden's lemma in a familiar way, one can show that every CFG for L_* assigns more than one derivation tree to some string in $L_2 \cap L_3$.

(ii) *L_* has no CFG that has the weak* 1*-FCP.* This follows from the proof of Lemma 1 since if L is context-free and $|C| \leq 1$, then $C^{\langle L|}$ is also context-free.

It would be interesting to see whether one or both of these properties of L_* are essential for Lemma 1 to hold. In particular, if a context-free language L has a CFG that has the weak 1-FCP, does it follow that L has a CFG that has the strong 1-FCP?

References

Clark, A. (2010). Learning context free grammars with the syntactic concept lattice. In J. M. Sempere and P. García (Eds.), *Grammatical Inference: Theoretical Results and Applications, 10th International Colloquium, ICGI 2010*, Lecture Notes in Artificial Intelligence, Berlin, pp. 38–51. Springer-Verlag.

Kanazawa, M. and R. Yoshinaka (to appear). Distributional learning and context/substructure enumerability in nonlinear tree grammars. In A. Foret, G. Morrill, R. Muskens, R. Osswald, and S. Pogodalla (Eds.), *Formal Grammar 2015/2016*, Lecture Notes in Computer Science, Berlin. Springer-Verlag.

Leiß, H. (2014). Learning context free grammars with the finite context property: A correction of A. Clark's algorithm. In G. Morrill, R. Muskens, R. Osswald, and F. Richter (Eds.), *Formal Grammar 2014*, Lecture Notes in Computer Science, Berlin, pp. 121–137. Springer-Verlag.

Ogden, W. (1968). A helpful result for proving inherent ambiguity. *Mathematical Systems Theory 2*(3), 191–194.

Yoshinaka, R. (2011). Towards dual approaches for learning context-free grammars based on syntactic concept lattices. In G. Mauri and A. Leporati (Eds.), *Developments in Language Theory, 15th International Conference, DLT 2011*, Lecture Notes in Computer Science, Berlin, pp. 429–440. Springer-Verlag.

Yoshinaka, R. (2015). Learning conjunctive grammars and contextual binary feature grammars. In A.-H. Dediu, E. Formenti, C. Martín-Vide, and B. Truthe (Eds.), *Language and Automata Theory and Applications: LATA 2015*, Lecture Notes in Computer Science, pp. 623–635. Springer.