D^3 as a 2-MCFL

Abstract. We discuss the open problem of parsing the Dyck language of 3 symbols, D^3, using a 2-Multiple Context-Free Grammar. We tackle this problem by implementing a number of novel techniques and present the associated software packages we developed.

Keywords: Dyck Language; multiple context free grammars (MCFG)

1 Introduction

Our goal with this paper is the analysis of the 3-dimensional Dyck language, D^3, under the scope of a 2-multiple context-free language, 2-MCFL. For brevity's sake, this section only serves as a brief introductory guide towards relevant papers, where the interested reader will find definitions, properties and various correspondences of the problem.

1.1 Preliminaries

We use D^3 to refer to the Dyck language over the lexicographically ordered alphabet $a < b < c$, which generalizes well-bracketed parentheses over three symbols. Denoting with $\#x(w)$ the number of occurrences of symbol x within word w, any word in D^3 satisfies the following conditions:

(D1) $\#a(w) = \#b(w) = \#c(w)$
(D2) $\#a(v) \geq \#b(v) \geq \#c(v),\ \forall v \in \textit{PrefixOf}(w)$

Eliding the second condition (D2), we get the MIX language, which represents free word order over the same alphabet. MIX has already been proven expressible by a 2-MCFG[10]; the class of multiple context-free grammars that operate on pairs of strings[2].

1.2 Motivation

Static Analysis Interestingly, the 2-symbol Dyck language is used in the *static analysis* of programming languages, where a large number of analyses are formulated as *language-reachability* problems[9].

For instance, when considering interprocedural calls as part of the source language, high precision can only be achieved by examining only control-flow paths that respect the fact that a procedure call always returns to the site of its current caller[8]. By associating the program point *before* a procedure call

f_k with $(_k$, and the one *after* the call with $)_k$, the validity problem is reduced to recognizing D^2 words.

Alas, the 2-dimensional case cannot accommodate richer control-flow structures, such as exception handling via `try/catch` and Python generators via the `yield` keyword. To achieve this, one must lift the Dyck-reachability problem to a higher dimension which, given the computational cost that context-sensitive parsing induces, is currently prohibited. If D^3 is indeed a 2-MCFL, parsing it would become computationally attainable for these purposes and eventually allow scalable analysis for non-standard control-flow mechanisms by exploiting the specific structure of analysed programs, as has been recently done in the 2-dimensional case[1].

Last but not least, future research directions will open up in a multitude of analyses that are currently restrained to two dimensions, such as *program slicing*, *flow-insensitive points-to analysis* and *shape approximation*[9].

Linguistics The extreme degree of 'scrambling' permitted by MIX is considered linguistically irrelevant[3]. On the other hand, D^3 enforces structural constraints, bringing it closer to natural languages. Hence, it is reasonable to examine whether D^3 can also be modelled by a 2-MCFG. Such an endeavour proved quite challenging, necessitating careful study of correspondences with other mathematical constructs.

1.3 Correspondences

Young Tableaux A standard Young Tableau is defined as an assortment of n boxes into a ragged (or jagged, i.e. non-rectangular) matrix containing the integers 1 through n and arranged in such a way that the entries are strictly increasing over the rows (from left to right) and columns (from top to bottom). Reading off the entries of the boxes, one may obtain the *Yamanouchi* word by placing (in order) each character's index to the row corresponding to its lexicographical ordering.

In the case of D^3, the Tableau associated with these words is in fact *rectangular* of size $n \times 3$, and the length of the corresponding word (called a *balanced or dominant Yamanouchi word* in this context) is $3n$, where n is the number of occurrences of each unique symbol[6]. Practically, the rectangular shape ensures constraint (D1), while the ascending order of elements over rows and columns ensures constraint (D2). In that sense, a rectangular standard Young tableau of size $n \times 3$ is, as a construct, an alternative way of uniquely representing the different words of D^3. We present an example tableau in Fig.1.

Promotions and Orbits There is an interesting transformation on Young Tableaux, namely the *Jeu-de-taquin* algorithm. When operating on a rectangular tableau $T(n, 3)$, Jeu-de-taquin consists of the following steps:

a:	1	3	4	8	9	10
b:	2	5	7	11	13	15
c:	6	12	14	16	17	18

Fig. 1: Young tableau for *"abaabcbaaabcbcbccc"*

(1) Reduce all elements of T by 1 and replace the first item of the first row with an empty box $\square(x, y) := (1, 1)$.
(2) While the empty box is not at the bottom right corner of T, $\square(x, y) \neq (n, 3)$, do:
   - Pick the minimum of the elements directly to the right and below the empty box, and swap the empty box with it. $T(x, y) := min(T_{(x+1,y)}, T_{(x,y+1)})$, $\square(x', y') := (x + 1, y)$ (in the case of a right-swap) or $\square(x', y') := (x, y + 1)$ (in the case of a down-swap).
(3) Replace the empty box with $3n$.

The tableau obtained through Jeu-de-taquin on T is called its promotion $p(T)$. We denote by $p^k(T)$, k successive applications of Jeu-de-taquin. It has been proven that $p^{3n}(T) = T$[7]. In other words, the promotion defines an equivalence class, which we name an *orbit*, which cycles back to itself. Orbits dissect the space of D^3 into disjoint sets, i.e. every word w belongs to a particular orbit, obtained by promotions of T_w.

A_2 *Combinatorial Spider Webs* The A_2 *irreducible combinatorial spider web* is a directed planar graph embedded in a disk that satisfies certain conditions [4]. Spider webs can be obtained through the application of a set of rules, known as the *Growth Algorithm*[7]. These operate on pairs of neighbouring nodes, collapsing them into a singular intermediate node, transforming them into a new pair or eliminating them altogether. Growth rules will be examined from a grammatical perspective in Section 2.2. Upon reaching a fixpoint, the growth process produces a well-formed Spider Web, which, in the context of D^3, can be interpreted as a visual representation of parsing a word[6,7].

A bijection also links Young Tableaux with Spider Webs. More specifically, the act of promotion is isomorphic to a combinatorial action on spider webs, namely *web rotation*[7].

Constrained Walk A Dyck word can also be visualized as a constrained *walk* within the first quadrant of $\mathbb{Z}^2$. We can assign each alphabet symbol x a vector value $\boldsymbol{v_x} \in \mathbb{Z}^2$ such that all pairs of $(\boldsymbol{v_x}, \boldsymbol{v_y})$ are linearly independent and:

$$\boldsymbol{v_a} + \boldsymbol{v_b} + \boldsymbol{v_c} = \boldsymbol{0} \tag{1}$$

$$\kappa\boldsymbol{v_a} + \lambda\boldsymbol{v_b} + \mu\boldsymbol{v_c} \geq \boldsymbol{0}, (\forall\kappa \geq \lambda \geq \mu) \tag{2}$$

We can then picture Dyck words as routes starting from $(0, 0)$. (1) means that each route must also end at $(0, 0)$ ($\cong$ (D1)), while (2) means that the x and

y axes may never be crossed (≅ (D2)). An example walk is depicted in Fig.2.

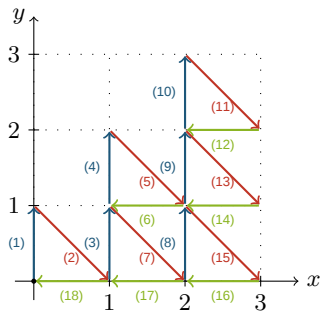

Fig. 2: The constrained walk of *"abaabcbaaabcbcbccc"* with vector value assignments $\boldsymbol{v}_a = (1, 0)$, $\boldsymbol{v}_b = (-1, 1)$, $\boldsymbol{v}_c = (0, -1)$

2 Modeling Techniques

We now present a number of novel techniques that we developed as an attempt to solve the problem at hand, incrementally moving towards more complex and abstract grammars. For the purpose of experimentation we have implemented these techniques, based on a software library for parsing MCFGs[5]. The resulting Python code is open-source and available online[1].

2.1 Triple Insertion

To set things off, we start with the grammar of *triple insertion* in Fig.3. This grammar operates on non-terminals W(x, y), producing W(x', y') with an additional triplet *a, b, c* that respects the partial orders $x < y$ and $a < b < c$. The end-word is produced through the concatenation of (x, y).

Despite being conceptually simple, this grammar consists of a large number of rules. Its expressivity is also limited; the prominent weak point is its inability to manage the effect of *straddling*, namely the generation of words whose substituents display complex interleaving patterns. Refer to Fig.10 for an example.

2.2 Meta-Grammars

To address the issue of rule size, we introduce the notion of *meta-grammars*, loosely inspired by Van Wijngaarden's work[11], which allows a more abstract

[1] https://github.com/omelkonian/dyck

$$\mathsf{S}(xy) \leftarrow \mathsf{W}(x, y). \tag{1}$$
$$\mathsf{W}(\epsilon, xy\mathbf{abc}) \leftarrow \mathsf{W}(x, y). \tag{2}$$
$$\dots$$
$$\mathsf{W}(\mathbf{abc}xy, \epsilon) \leftarrow \mathsf{W}(x, y). \tag{61}$$
$$\mathsf{W}(\epsilon, \mathbf{abc}). \tag{62}$$
$$\dots$$
$$\mathsf{W}(\mathbf{abc}, \epsilon). \tag{65}$$

Fig. 3: Grammar of triple insertions

view of the grammar as a whole. Specifically, we define $\mathcal{O}$ as the *meta-rule* which, given a rule format, a set of partial orders (over the tuple indices of its premises and/or newly added terminal symbols), and the MCFG dimensionality, automatically generates all the order-respecting permutations. An example of how we can abstract away from explicitly enumerating the entirety of our initial rules is showcased in Fig.4.

$$\mathsf{S}(xy) \leftarrow \mathsf{W}(x, y).$$
$$\mathcal{O}_2[\![\mathsf{W} \leftarrow \epsilon \mid \{a < b < c\}]\!].$$
$$\mathcal{O}_2[\![\mathsf{W} \leftarrow \mathsf{W} \mid \{x < y,\ a < b < c\}]\!].$$

Fig. 4: $\mathcal{G}_0$: Meta-grammar of triple insertions

This approach enhances the potential expressivity of our grammars as well. For instance, we can now extend the previous with a single meta-rule that allows two non-terminals $\mathsf{W}(x, y)$, $\mathsf{W}(z, w)$ to interleave with one another, producing rearranged tuple concatenations and allowing some degree of straddling to be generated:

$$\mathcal{G}_1 : \mathcal{G}_0 + \mathcal{O}_2[\![\mathsf{W} \leftarrow \mathsf{W}, \mathsf{W} \mid \{x < y,\ z < w\}]\!].$$

The addition of this rule gets us closer to completeness, but we are still not quite there. We have thus far only used a single non-terminal, not utilizing the expressivity that an MCFG allows. To that end, we propose non-terminals to represent incomplete word *states*; that is, words that either have an extra symbol or miss one. The former are *positive* states, whereas the latter are *negative*. The inclusion of these extra states would allow for more intricate interactions.

Interestingly, there is a direct correspondence between these non-terminals and the nodes of Petersen's growth algorithm[7]. Fig.5 depicts the growth rules in the exact same web form as proposed by Petersen, modulo node branding. A subset of these web-reduction rules are, in fact, precisely modelled by the meta-grammar $\mathcal{G}_2$ presented in Fig.6. In section 4, we briefly explains our inability to model the whole set of rules with a 2-MCFG, hence rendering our grammar complete.

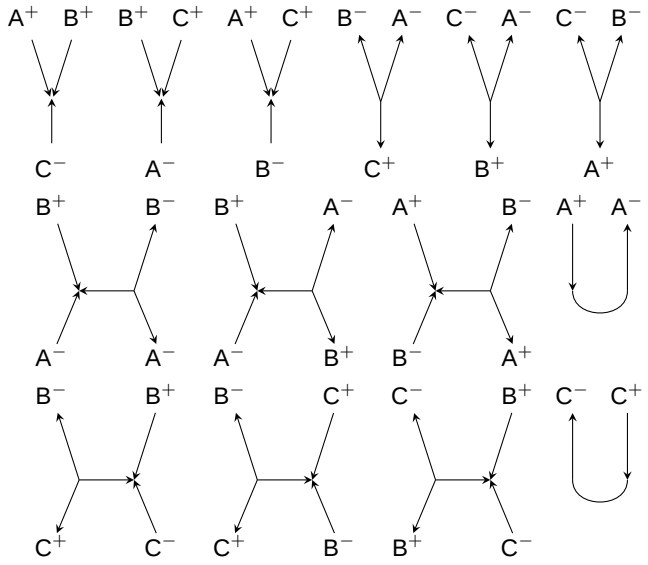

Fig. 5: Growth rules

$\mathcal{G}_2$ consists of base cases for positive states, possible state interactions, closures of pairs of inverse polarity and a universally quantified meta-rule that allows the combination of any incomplete state with a well-formed one (i.e. non-terminal W).

A further extension can be achieved through universally quantifying the notion of triple insertion, which is unique in the sense that it can insert three different terminals, each at a different position:

$$\mathcal{G}_3 : \mathcal{G}_2 + \forall\, \mathsf{K} \in \{\mathsf{A}^{+/-},\ \mathsf{B}^{+/-},\ \mathsf{C}^{+/-}\} : \mathcal{O}_2[\![\mathsf{K} \leftarrow \mathsf{K} \mid \{x < y,\ a < b < c\}]\!].$$

2.3 Rule Inference

The improved performance of the above approaches again proved insufficient to completely parse D^3. Our meta-rules are over-constrained by imposing a

$\mathsf{S}(xy) \leftarrow \mathsf{W}(x, y).$
$\mathcal{O}_2[\![\mathsf{W} \leftarrow \epsilon \mid \{a < b < c\}]\!].$
$\mathcal{O}_2[\![\mathsf{A}^+ \leftarrow \epsilon \mid \{a\}]\!].$
$\mathcal{O}_2[\![\mathsf{B}^+ \leftarrow \epsilon \mid \{b\}]\!].$
$\mathcal{O}_2[\![\mathsf{C}^+ \leftarrow \epsilon \mid \{c\}]\!].$
$\mathcal{O}_2[\![C^- \leftarrow A^+, B^+ \mid \{x < y < z < w\}]\!].$
$\mathcal{O}_2[\![B^- \leftarrow A^+, C^+ \mid \{x < y < z < w\}]\!].$
$\mathcal{O}_2[\![A^- \leftarrow B^+, C^+ \mid \{x < y < z < w\}]\!].$
$\mathcal{O}_2[\![A^+ \leftarrow C^-, B^- \mid \{x < y < z < w\}]\!].$
$\mathcal{O}_2[\![B^+ \leftarrow C^-, A^- \mid \{x < y < z < w\}]\!].$
$\mathcal{O}_2[\![C^+ \leftarrow B^-, A^- \mid \{x < y < z < w\}]\!].$
$\mathcal{O}_2[\![W \leftarrow A^+, A^- \mid \{x < y < z < w\}]\!].$
$\mathcal{O}_2[\![W \leftarrow C^-, C^+ \mid \{x < y < z < w\}]\!].$
$\forall\, \mathsf{K} \in \{\mathsf{A}^{+/-},\ \mathsf{B}^{+/-},\ \mathsf{C}^{+/-}\}:$
$\quad \mathcal{O}_2[\![\mathsf{K} \leftarrow \mathsf{K}, \mathsf{W} \mid \{x < y,\ z < w\}]\!].$

Fig. 6: $\mathcal{G}_2$: Meta-grammar of incomplete states

total order on the tuple elements, due to their inability to keep track of where the extra character(s) is. To overcome this, we split each state into multiple position-aware, *refined* states. Doing so revealed a vast amount of new interactions, as evidenced by the below alteration to the original A^+, B^+ interaction (where y can now occur after z or w):

$$\mathcal{O}_2[\![C^- \leftarrow A^+_{left}, B^+ \mid \{x < y, x < z < w\}]\!].$$

In order to accommodate the interactions between this increased number of states, we need to keep track of both internal and external order constraints. At this point, the abstraction offered by our meta-grammar approach does not cover our needs any more. The same difficulty that we had encountered before is prominent once more, except now at an even higher level.

As a solution to the aforementioned limitation, we propose a system that can automatically create a full-blown m-MCFG given only the states it consists of. To accomplish this, we assign each state a unique *descriptor* that specifies the content of its tuple's elements. Aligning these descriptors with the tuple, we can then infer the descriptor of the resulting tuple of every possible state interaction. For the subset of those interactions whose resulting descriptor is matched with a state, we can now automatically infer the rule.

Formally, the system is initialized with a map $\mathcal{D}$, such as the one illustrated in Fig.7. Its domain, $dom(\mathcal{D})$, is a set of *state identifiers* and its codomain, $codom(\mathcal{D})$, is the set of their corresponding *state descriptors*.

$$
\begin{aligned}
\mathsf{W} &\mapsto (\epsilon, \epsilon) \\
\mathsf{A}_l^+ &\mapsto (a, \epsilon) \\
\mathsf{A}_r^+ &\mapsto (\epsilon, a) \\
&\vdots \\
\mathsf{C}_r^- &\mapsto (\epsilon, ab) \\
\mathsf{C}_{l,r}^- &\mapsto (a, b)
\end{aligned}
$$

Fig. 7: Map $\mathcal{D}$ for refined states

Algorithm 1 ARIS: Automatic Rule Inference System

```
procedure aris(𝒟)
    for X ↦ (d_1, ..., d_n) ∈ 𝒟 do
        yield X(d_1, ..., d_n).
    for X, Y ∈ dom(𝒟)^2 do
        (X_ord, Y_ord) ← (x < y < ..., z < w < ...)
        for (d_1, ..., d_n) ∈ 𝒪_2⟦_ ← X, Y | {X_ord, Y_ord}⟧ do
            for S' ∈ eliminate((d_1, ..., d_n), 𝒟) do
                yield S'(d_1, ..., d_n) ← X, Y.

procedure eliminate((d_1, ..., d_n), 𝒟)
    for matches ∈ all_abc_triplets(d_1, ..., d_n) do
        for i ∈ 0 ... n/3 do
            for S' ∈ remove_abc_triplets(matches, i) do
                if S' ∈ codom(𝒟) then
                    yield S'
```

Meta-grammars accelerated the process of creating grammars, by letting us simply describe rules instead of explicitly defining them. ARIS builds upon this notion to raise the level of abstraction even further; one needs only specify a grammar's states and its descriptors, thus eliminating the need to define rules or even meta-rules.

3 Tools & Results

3.1 Grammar Utilities

We have implemented the modelling techniques described in Section 2 and distributed a Python package, called **dyck**, which provides the programmer with a *domain-specific language* close to this paper's mathematical notation.

To facilitate experimentation, our package includes features such as grammar selection, time measurements, word generation and soundness/completeness checking. The following example demonstrates the definition of $\mathcal{G}_1$:

```
from dyck import *
G_1 = Grammar([
    ('S <- W', {(x, y)}),
    O('W', {(a, b, c)}),
    O('W <- W', {(x, y), (a, b, c)}),
    O('W <- W, W', {(x, y), (z, w)})
])
```

3.2 Visualization

As counter-examples began to grow in size and number, we realised the necessity of a visualization tool to assist us in identifying properties they may exhibit. To that end, we distribute another Python package, called **dyckviz**, which allows the simultaneous visualization of tableau-promotion and web-rotation (grouped in their corresponding equivalence classes). An example of a web as rendered by our tool is given in Fig.8.

Young tableaux in an orbit are colour-grouped by their column indices, which sheds some light on how the *jeu-de-taquin* actually influences the structure of the corresponding Dyck words. Interesting patterns have began to emerge, which still remain to be properly investigated.

3.3 Grammar Comparisons

We display three charts, depicting the number of rules, percentage of counter-examples and computation times of each of our grammars for D_n^3 with n ranging from 2 to 6 (where n denotes the number of abc triplets). Even though none of our proposed grammars is complete, we observe that as grammars get more abstract, the number of failing parses steadily declines. This however comes at the cost of rule size growth, which in turn is associated with an increase in computation times. What this practically means is that we are unable to continue testing more elaborate grammars or scale our results to higher orders of n (note that $\|D_n^3\|$ also has a very rapid rate of expansion)[6].

4 Road to Completeness

To our knowledge, no other attempt has come so close to modelling D^3 with a 2-MCFG. We attribute this to the combination of a pragmatic approach with results from existing theoretical work. In this section, we present a collection of additional ideas, which we consider worthy of further exploration.

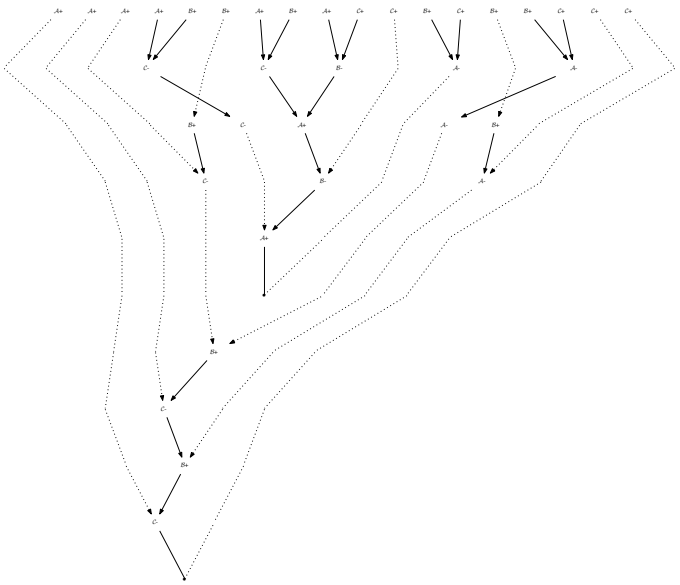

Fig. 8: Spider web of *"abaacbbacbabaccbcc"*

First-Match Policy and Relinking Possibly the most intuitive way of checking whether a word w is part of D_n^3 is checking whether a pair of links occur that match a_i to b_i and b_i to c_i $\forall i \in n$. We call this process of matching the *first-match policy*. The question arises whether a grammar can accomplish inserting a triplet of a, b, c, that would abide by the first-match policy. If that were the case, it would be relatively easy to generalize this ability by induction to every $n \in \mathbb{N}$. Unfortunately, the answer is seemingly negative; the expressiveness provided by a 2-MCFG does not allow for the arbitrary insertions required. On a related note, being able to produce a word state $W(x, y)$ where $w = xy$ and x any possible prefix of w, gives no guarantee of being able to produce the same word with an extra triplet inserted due to the straddling property.

However, if rules existed that would allow for match-making and breaking, i.e. match *relinking*, an inserted symbol could be temporarily matched with what might be its first match-policy in a local scope, and then relink it to its correct match when merging two words together.

Growth Rules Although $\mathcal{G}_2$ comes close to realizing the growth algorithm, not all of the growth rules can be translated into a 2-MCFG setting. It would be an interesting endeavour to attempt to model the element-swapping behaviour of these rules that produce two output states, without resorting to more expressive formalisms (e.g. context-sensitive grammars).

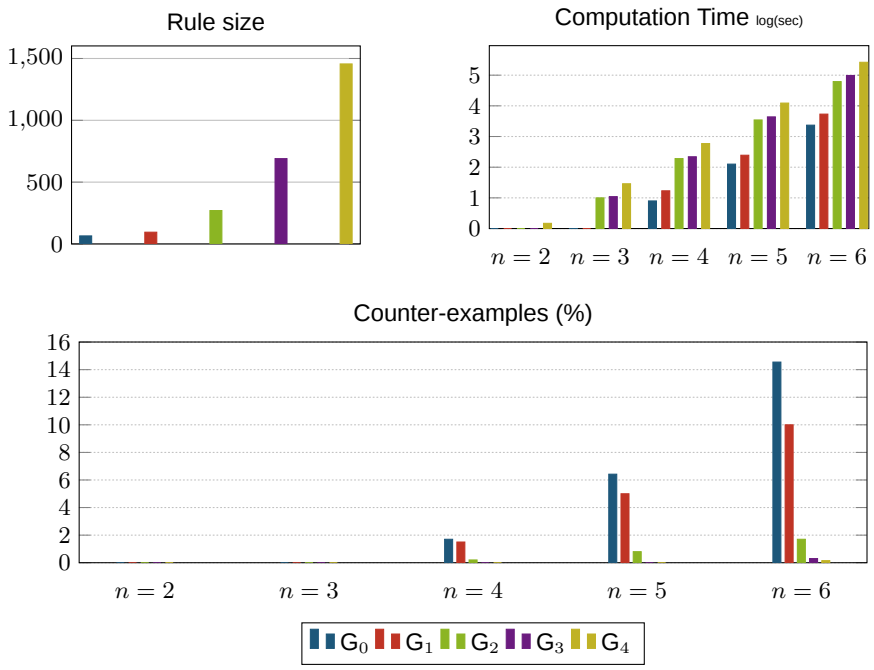

Fig. 9: Performance measures

Fig. 10: First-match policy for *"ababacbcabcc"*

Insights from promotion An interesting question is whether promotion can be handled by a 2-MCFG (as a *context-free rewriting system*). If so, it could be worth looking into the properties of orbits, to test for instance if there are promotions within an orbit that can be easier to solve than others. Solving a single promotion and transducing the solution to all equivalent words could then be a guideline towards completeness.

5 Conclusion

We tried to accurately present the intricacies of D^3 and the difficulties that arise when attempting to model it under the scope of a 2-MCFL. We have developed and introduced some novel techniques and tools, which we believe can be of

use even outside the problem's narrow domain. We have largely expanded on the existing tools to accommodate MIX-style languages and systems of meta-grammars in general.

Despite our best efforts, the question of whether D^3 can actually be encapsulated within a 2-MCFG still remains unanswered. Regardless, this problem has been very rewarding to pursue, and we hope to have intrigued the interested reader enough to further research the subject, use our code, or strive for a solution on his/her own.

Acknowledgements

We would like to thank Dr. Michael Moortgat for introducing us to the problem, providing insightful feedback and motivating us throughout the process, as well as Dr. Jurriaan Hage for suggesting the use of multi-dimensional Dyck languages in static analysis.

References

1. Chatterjee, K., Choudhary, B., Pavlogiannis, A.: Optimal dyck reachability for data-dependence and alias analysis. Proceedings of the ACM on Programming Languages 2(POPL), 30 (2017)
2. Götzmann, D.N.: Multiple context-free grammars
3. Kanazawa, M., Salvati, S.: Mix is not a tree-adjoining language. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. pp. 666–674. Association for Computational Linguistics (2012)
4. Kuperberg, G.: Spiders for rank 2 lie algebras. Communications in mathematical physics 180(1), 109–151 (1996)
5. Ljunglöf, P.: Practical parsing of parallel multiple context-free grammars. In: Workshop on Tree Adjoining Grammars and Related Formalisms. p. 144 (2012)
6. Moortgat, M.: A note on multidimensional dyck languages. In: Categories and Types in Logic, Language, and Physics. pp. 279–296 (2014)
7. Petersen, T.K., Pylyavskyy, P., Rhoades, B.: Promotion and cyclic sieving via webs. Journal of Algebraic Combinatorics 30(1), 19–41 (2009)
8. Reps, T., Horwitz, S., Sagiv, M.: Precise interprocedural dataflow analysis via graph reachability. pp. 49–61. ACM (1995)
9. Reps, T.W.: Program analysis via graph reachability. In: Logic Programming, Proceedings of the 1997 International Symposium, Port Jefferson, Long Island, NY, USA, October 13-16, 1997. pp. 5–19 (1997)
10. Salvati, S.: Mix is a 2-mcfl and the word problem in z^2 is solved by a third-order collapsible pushdown automaton. Journal of Computer and System Sciences 81(7), 1252–1277 (2015)
11. van Wijngaarden, A.: The generative power of two-level grammars. In: ICALP (1974)