Journal of Automata, Languages and Combinatorics **21** (2016) 3, 203 – 246
© Institut für Informatik · Justus-Liebig-Universität Giessen

NON-SELF-EMBEDDING LINEAR CONTEXT-FREE TREE GRAMMARS GENERATE REGULAR TREE LANGUAGES

Mark-Jan Nederhof$^{(A)}$ Markus Teichmann$^{(B,C)}$ Heiko Vogler$^{(B)}$

$^{(A)}$*School of Computer Science, University of St Andrews*
North Haugh, KY16 9SX, UK
markjan.nederhof@gmail.com

$^{(B)}$*Department of Computer Science, Technische Universität Dresden*
01062 Dresden, Germany
markus.teichmann@mailbox.tu-dresden.de heiko.vogler@tu-dresden.de

ABSTRACT

For the class of linear context-free tree grammars, we define a decidable property called self-embedding. We prove that each non-self-embedding grammar in this class generates a regular tree language and show how to construct the equivalent regular tree grammar.

Keywords: context-free tree grammar, regular tree grammar, self-embedding, natural language processing

1. Introduction

In natural language processing (NLP), formal string grammars are used to approximate the set of all syntactically valid sentences of a language. Two important and successful grammar classes are the regular grammars (REGs) and the context-free grammars (CFGs) [16]. For these two classes there is a clear trade-off between expressive power and cost of processing, e. g., for parsing. It is undecidable whether an arbitrary given CFG generates a regular language [13, Thm. 8.15], but one may approximate a given context-free language by a REG, for example, in order to achieve better parsing complexity [22]. Alternatively, one may restrict CFGs to satisfy a decidable property that guarantees that they generate regular languages. Chomsky [3] defined such a property called non-self-embedding. A CFG is self-embedding if there are a nonterminal A and non-empty strings v and w over terminals and nonterminals such that $A \Rightarrow^* vAw$. He proved that each non-self-embedding CFG generates a regular language [3, Thm. 11]. In [22] self-embedding was expressed as a syntactic criterion, accompanied by a direct construction of a REG from a non-self-embedding CFG.

$^{(C)}$Financially supported by DFG Graduiertenkolleg 1763 (QuantLA).