





DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

# **Formal Verification of an Earley Parser**

Martin Rau



DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

# **Formal Verification of an Earley Parser**

## **Formale Verifikation eines Earley Parsers**

Author:	Martin Rau
Supervisor:	Tobias Nipkow
Advisor:	Tobias Nipkow
Submission Date:	15.06.2023

I confirm that this master's thesis in informatics is my own work and I have documented all sources and material used.

Munich, 15.06.2023

Martin Rau

## Acknowledgments

I owe an enormous debt of gratitude to my family which always supported me throughout my studies. Thank you. I also would like to thank Prof. Tobias Nipkow for introducing me to the world of formal verification through Isabelle and for supervising both my Bachelor's and my Master's thesis. It was a pleasure to learn from and to work with you.

# Abstract

TODO

# Contents

<b>Acknowledgments</b>	<b>iii</b>
<b>Abstract</b>	<b>iv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Related Work . . . . .	1
1.2 Structure . . . . .	2
1.3 Contributions . . . . .	2
<b>2 Earley Recognizer</b>	<b>4</b>
<b>3 Earley Recognizer Formalization</b>	<b>7</b>
3.1 Context-free grammars and Isabelle/HOL . . . . .	7
3.2 The Formalized Algorithm . . . . .	10
3.3 Well-formedness . . . . .	14
3.4 Soundness . . . . .	15
3.5 Completeness . . . . .	18
3.6 Finiteness . . . . .	25
<b>4 Earley Recognizer Implementation</b>	<b>26</b>
4.1 The Executable Algorithm . . . . .	26
4.2 A Word on Performance . . . . .	31
4.3 Sets or Bins as Lists . . . . .	33
4.4 Well-formedness and Termination . . . . .	35
4.5 Soundness . . . . .	37
4.6 Completeness . . . . .	40
4.7 Correctness . . . . .	48
<b>5 Earley Parser Implementation</b>	<b>49</b>
5.1 A Single Parse Tree . . . . .	50
5.1.1 Pointer Lemmas . . . . .	51
5.1.2 The Parse Tree Algorithm . . . . .	54
5.1.3 Termination . . . . .	56

5.1.4	Correctness . . . . .	57
5.2	A Parse Forest . . . . .	58
5.2.1	The Parse Forest Algorithm . . . . .	59
5.2.2	Termination . . . . .	60
5.2.3	Soundness . . . . .	60
5.3	A Word on Performance and Completeness . . . . .	61
<b>6</b>	<b>The Running Example</b>	<b>63</b>
<b>7</b>	<b>Conclusion</b>	<b>65</b>
7.1	Summary . . . . .	65
7.2	Future Work . . . . .	65



# 1 Introduction

some introduction about parsing, formal development of correct algorithms: an example based on earley’s recogniser, the benefits of formal methods, LocalLexing and the Bachelor thesis.

## 1.1 Related Work

Tomita [**Tomita:1987**] presents an generalized LR parsing algorithm for augmented context-free grammars that can handle arbitrary context-free grammars.

Izmaylova *et al* [**Izmaylova:2016**] develop a general parser combinator library based on memoized Continuation-Passing Style (CPS) recognizers that supports all context-free grammars and constructs a Shared Packed Parse Forest (SPPF) in worst case cubic time and space.

Obua *et al* [**Obua:2017**] introduce local lexing, a novel parsing concept which interleaves lexing and parsing whilst allowing lexing to be dependent on the parsing process. They base their development on Earley’s algorithm and have verified the correctness with respect to its local lexing semantics in the theorem prover Isabelle/HOL. The background theory of this Master’s thesis is based upon the local lexing entry [**LocalLexing-AFP**] in the Archive of Formal Proofs.

Lasser *et al* [**Lasser:2019**] verify an LL(1) parser generator using the Coq proof assistant.

Barthwal *et al* [**Barthwal:2009**] formalize background theory about context-free languages and grammars, and subsequently verify an SLR automaton and parser produced by a parser generator.

Blaudeau *et al* [**Blaudeau:2020**] formalize the metatheory on Parsing expression grammars (PEGs) and build a verified parser interpreter based on higher-order parsing combinators for expression grammars using the PVS specification language and verification system. Koprowski *et al* [**Koprowski:2011**] present TRX: a parser interpreter formally developed in Coq which also parses expression grammars.

Jourdan *et al* [**Jourdan:2012**] present a validator which checks if a context-free grammar and an LR(1) parser agree, producing correctness guarantees required by verified compilers.

Lasser *et al* [Lasser:2021] present the verified parser CoStar based on the ALL(\*) algorithm. They proof soundness and completeness for all non-left-recursive grammars using the Coq proof assistant.

## 1.2 Structure

## 1.3 Contributions

SNIPPET:

Context-free grammars have been used extensively for describing the syntax of programming languages and natural languages. Parsing algorithms for context-free grammars consequently play a large role in the implementation of compilers and interpreters for programming languages and of programs which understand or translate natural languages. Numerous parsing algorithms have been developed. Some are general, in the sense that they can handle all context-free grammars, while others can handle only subclasses of grammars. The latter, restricted algorithms tend to be much more efficient. The algorithm described here seems to be the most efficient of the general algorithms, and also it can handle a larger class of grammars in linear time than most of the restricted algorithms.

SNIPPET:

The Computer Science community has been able to automatically generate parsers for a very wide class of context free languages. However, many parsers are still written manually, either using tool support or even completely by hand. This is partly because in some application areas such as natural language processing and bioinformatics we don't have the luxury of designing the language so that it is amendable to know parsing techniques, but also it is clear that left to themselves computer language designers do not naturally write LR(1) grammars. A grammar not only defines the syntax of a language, it is also the starting point for the definition of the semantics, and the grammar which facilitates semantics definition is not usually the one which is LR(1). Given this difficulty in constructing natural LR(1) grammars that support desired semantics, the general parsing techniques, such as the CYK Younger [Younger:1967], Earley [Earley:1970] and GLR Tomita [Tomita:1985] algorithms, developed for natural language processing are also of interest to the wider computer science community. When using grammars as the starting point for semantics definition, we distinguish between recognizers which simply determine whether or not a given string is in the language defined by a given grammar, and parser which also return some form of derivation of the string, if one exists. In their basic form the CYK and Earley algorithms are recognizers while GLR-style algorithms are designed with derivation tree construction, and hence parsing, in mind.

There is no known linear time parsing or recognition algorithm that can be used with all

context free grammars. In their recognizer forms the CYK algorithm is worst case cubic on grammars in Chomsky normal form and Earley's algorithm is worst case cubic on general context free grammars and worst case  $n^2$  on non-ambiguous grammars. General recognizers must, by definition, be applicable to ambiguous grammars. Tomita's GLR algorithm is of unbounded polynomial order in the worst case. Expanding general recognizers to parser raises several problems, not least because there can be exponentially many or even infinitely many derivations for a given input string. A cubic recognizer which was modified to simply return all derivations could become an unbounded parser. Of course, it can be argued that ambiguous grammars reflect ambiguous semantics and thus should not be used in practice. This would be far too extreme a position to take. For example, it is well known that the if-else statement in the ANSI-standard grammar for C is ambiguous, but a longest match resolution results in a linear time parser that attach the else to the most recent if, as specified by the ANSI-C semantics. The ambiguous ANSI-C grammar is certainly practical for parser implementation. However, in general ambiguity is not so easily handled, and it is well known that grammar ambiguity is in fact undecidable Hopcroft *et al* [Hopcroft:2006], thus we cannot expect a parser generator simply to check for ambiguity in the grammar and report the problem back to the user. Another possibility is to avoid the issue by just returning one derivation. However, if only one derivation is returned then this creates problems for a user who wants all derivations and, even in the case where only one derivation is required, there is the issue of ensuring that it is the required derivation that is returned. A truly general parser will return all possible derivations in some form. Perhaps the most well known representation is the shared packed parse forest SPPF described and used by Tomita [Tomita:1985]. Tomita's description of the representation does not allow for the infinitely many derivations which arise from grammars which contain cycles, the source adapt the SPPF representation to allow these. Johnson [Johnson:1991] has shown that Tomita-style SPPFs are worst case unbounded polynomial size. Thus using such structures will also turn any cubic recognition technique into a worst case unbounded polynomial parsing technique. Leaving aside the potential increase in complexity when turning a recogniser into a parser, it is clear that this process is often difficult to carry out correctly. Earley gave an algorithm for constructing derivations of a string accepted by his recognizer, but this was subsequently shown by Tomita [Tomita:1985] to return spurious derivations in certain cases. Tomita's original version of his algorithm failed to terminate on grammars with hidden left recursion and, as remarked above, had no mechanism for constructing complete SPPFs for grammars with cycles.

## 2 Earley Recognizer

We present a slightly simplified version of Earley’s original recognizer algorithm [Earley:1970], omitting Earley’s proposed look-ahead since its primary purpose is to increase the efficiency of the resulting recognizer. Throughout this thesis we are working with a running example. The considered grammar is a tiny excerpt of a toy arithmetic expression grammar:  $\mathcal{G} ::= S \rightarrow x \mid S \rightarrow S + S$  and the, rather trivial, input is  $\omega = x + x + x$ .

Intuitively, Earley’s recognizer works in principle like a top-down parser carrying along all possible parses simultaneously in an efficient manner. In detail, the algorithm works as follows: it parses the input  $\omega = a_0, \dots, a_n$ , constructing  $n + 1$  Earley bins  $B_i$  that are sets of Earley items. An initial bin  $B_0$  and one bin  $B_{i+1}$  for each symbol  $a_i$  of the input. In general, an Earley item  $A \rightarrow \alpha \bullet \beta, i, j$  consists of four parts: a production rule of the grammar that we are currently considering, a bullet signalling how much of the productions right-hand side we have recognized so far, an origin  $i$  describing the position in  $\omega$  where we started parsing, and an end  $j$  indicating the position in  $\omega$  we are currently considering next for the remaining right-hand side of the production rule. Note that there will be only one set of earley items or only one bin  $B$  and we say an item is conceptually part of bin  $B_j$  if its end is the index  $j$ . Table 2.1 lists the items for our example grammar. Bin  $B_4$  contains for example the item  $S \rightarrow S + \bullet S, 2, 4$ . Or, we are considering the rule  $S \rightarrow S + S$ , have recognized the substring from 2 to 4 (the first index being inclusive the second one exclusive) of  $\omega$  by  $\alpha = S +$ , and are trying to parse  $\beta = S$  from position 4 in  $\omega$ .

The algorithm initializes  $B$  by applying the *Init* operation. It then proceeds to execute the *Scan*, *Predict* and *Complete* operations listed in Figure 2.1 until there are no more new items being generated and added to  $B$ . Next we describe these four operations in detail:

1. The *Init* operation adds items  $S \rightarrow \bullet \alpha, 0, 0$  for each production rule containing the start symbol  $S$  on its left-hand side.

For our example *Init* adds the items  $S \rightarrow \bullet x, 0, 0$  and  $S \rightarrow \bullet S + S, 0, 0$ .

2. The *Scan* operation applies if there is a terminal to the right-hand side of the bullet, or items of the form  $A \rightarrow \alpha \bullet a \beta, i, j$ , and the  $j$ -th symbol of  $\omega$  matches the terminal symbol following the bullet. We add one new item  $A \rightarrow \alpha a \bullet \beta, i, j + 1$  to  $B$  moving the bullet over the parsed terminal symbol.

Considering our example, bin  $B_3$  contains the item  $S \rightarrow S \bullet + S, 2, 3$ , the third symbol of  $\omega$  is the terminal  $+$ , so we add the item  $S \rightarrow S + \bullet S, 2, 4$  to the conceptual bin  $B_4$ .

3. The *Predict* operation is applicable to an item when there is a non-terminal to the right-hand side of the bullet or items of the form  $A \rightarrow \alpha \bullet B\beta, i, j$ . It adds one new item  $B \rightarrow \bullet \gamma, j, j$  to the bin for each alternate  $B \rightarrow \gamma$  of that non-terminal.

E.g. for the item  $S \rightarrow S + \bullet S, 0, 2$  in  $B_2$  we add the two items  $S \rightarrow \bullet x, 2, 2$  and  $S \rightarrow \bullet S + S, 2, 2$  corresponding to the two alternates of  $S$ . The bullet is set to the beginning of the right-hand side of the production rule, the origin and end are set to  $j = 2$  to indicate that we are starting to parse in the current bin and have not parsed anything so far.

4. The *Complete* operation applies if we process an item with the bullet at the end of the right-hand side of its production rule. For an item  $B \rightarrow \gamma \bullet, j, k$  we have successfully parsed the substring  $\omega[j..k)$ , as mentioned before indices  $j$  and  $k$  being inclusive respectively exclusive, and are now going back to the origin bin  $B_j$  where we predicted this non-terminal. There we look for any item of the form  $A \rightarrow \alpha \bullet B\beta, i, j$  containing a bullet in front of the non-terminal we completed, or the reason we predicted it on the first place. Since we parsed the predicted non-terminal successfully, we are allowed to move over the bullet, resulting in one new item  $A \rightarrow \alpha B \bullet \beta, i, k$ . Note in particular the origin and end indices.

Looking back at our example, we can add the item  $S \rightarrow S + S \bullet, 0, 5$  for two different reasons corresponding to the two different ways we can derive  $\omega$ . When processing  $S \rightarrow x \bullet, 4, 5$  we find  $S \rightarrow S + \bullet S, 0, 4$  in the origin bin  $B_4$  which corresponds to recognizing  $(x + x) + x$ . We would add the same item again while applying the *Complete* operation to  $S \rightarrow S + S \bullet, 2, 5$  and  $S \rightarrow S + \bullet S, 0, 2$  which corresponds to recognizing the input as  $x + (x + x)$ .

If the algorithm encounters an item of the form  $S \rightarrow \alpha \bullet, 0, |\omega| + 1$ , it returns *true*, otherwise it returns *false*. For the tiny arithmetic expression grammar we generate the item  $S \rightarrow S + S \bullet, 0, 5$  and return the correct answer *true*, since there exist derivations for  $\omega = x + x + x$ , e.g.  $S \Rightarrow S + S \Rightarrow x + S \Rightarrow x + S + S \xRightarrow{*} x + x + x$  or  $S \Rightarrow S + S \Rightarrow S + x \Rightarrow S + S + x \xRightarrow{*} x + x + x$ .

To proof the correctness of Earley's recognizer algorithm we need to show the following theorem:

$$S \rightarrow \alpha \bullet, 0, |\omega| + 1 \in B \text{ iff } S \xRightarrow{*} \omega$$

It follows from the following three lemmas:

1. Soundness: for every generated item there exists an according derivation:  
 $A \rightarrow \alpha \bullet \beta, i, j \in B$  implies  $A \Rightarrow^* \omega[i..j] \beta$
2. Completeness: for every derivation we generate an according item:  
 $A \Rightarrow^* \omega[i..j] \beta$  implies  $A \rightarrow \alpha \bullet \beta, i, j \in B$
3. Finiteness: there only exist a finite number of Earley items

$$\begin{array}{c}
 \text{INIT} \\
 \hline
 S \rightarrow \bullet \alpha, 0, 0
 \end{array}
 \qquad
 \begin{array}{c}
 \text{SCAN} \\
 \hline
 \frac{A \rightarrow \alpha \bullet a \beta, i, j \quad \omega[j] = a}{A \rightarrow \alpha a \bullet \beta, i, j+1}
 \end{array}$$
  

$$\begin{array}{c}
 \text{PREDICT} \\
 \hline
 \frac{A \rightarrow \alpha \bullet B \beta, i, j \quad (B \rightarrow \gamma) \in \mathcal{G}}{B \rightarrow \bullet \gamma, j, j}
 \end{array}
 \qquad
 \begin{array}{c}
 \text{COMPLETE} \\
 \hline
 \frac{A \rightarrow \alpha \bullet B \beta, i, j \quad B \rightarrow \gamma \bullet, j, k}{A \rightarrow \alpha B \bullet \beta, i, k}
 \end{array}$$

Figure 2.1: Earley inference rules

Table 2.1: Earley items for the grammar  $\mathcal{G}: S \rightarrow x, S \rightarrow S + S$

$B_0$	$B_1$	$B_2$
$S \rightarrow \bullet x, 0, 0$ $S \rightarrow \bullet S + S, 0, 0$	$S \rightarrow x \bullet, 0, 1$ $S \rightarrow S \bullet + S, 0, 1$	$S \rightarrow S + \bullet S, 0, 2$ $S \rightarrow \bullet x, 2, 2$ $S \rightarrow \bullet S + S, 2, 2$
$B_3$	$B_4$	$B_5$
$S \rightarrow x \bullet, 2, 3$ $S \rightarrow S + S \bullet, 0, 3$ $S \rightarrow S \bullet + S, 2, 3$ $S \rightarrow S \bullet + S, 0, 3$	$S \rightarrow S + \bullet S, 2, 4$ $S \rightarrow S + \bullet S, 0, 4$ $S \rightarrow \bullet x, 4, 4$ $S \rightarrow \bullet S + S, 4, 4$	$S \rightarrow x \bullet, 4, 5$ $S \rightarrow S + S \bullet, 2, 5$ $S \rightarrow S + S \bullet, 0, 5$ $S \rightarrow S \bullet + S, 4, 5$ $S \rightarrow S \bullet + S, 2, 5$ $S \rightarrow S \bullet + S, 0, 5$

## 3 Earley Recognizer Formalization

In this chapter we shortly introduce the interactive theorem prover Isabelle/HOL [Nipkow:2002] used as the tool for verification in this thesis and recap some of the formalism of context-free grammars and their representation in Isabelle. Then we formalize the simplified Earley recognizer algorithm presented in Chapter 2; discussing the implementation and the proofs for soundness, completeness, and finiteness. Note that most of the basic definitions of Sections 3.1 and 3.2 are not our own work but only slightly adapted from Obua’s work on *Local Lexing* [Obua:2017] [LocalLexing-AFP]. All of the proofs in this chapter are our own work.

### 3.1 Context-free grammars and Isabelle/HOL

Isabelle/HOL [Nipkow:2002] is an interactive theorem prover based on a fragment of higher-order logic. It supports the core concepts commonly known from functional programming languages. The notation  $t :: \tau$  means that term  $t$  has type  $\tau$ . Basic types include *bool*, *nat*; type variables are written  $'a$ ,  $'b$ , etc. Pairs are written  $(a, b)$ ; triples are written  $(a, b, c)$  and so forth but are internally represented as nested pairs; the nesting is on the first component of a pair. Functions *fst* and *snd* return the first and second component of a pair; the operator  $(\times)$  represents pairs at the type level. Most type constructors are written postfix, e.g.  $'a \text{ set}$  and  $'a \text{ list}$ ; the function space arrow is  $\Rightarrow$ ; function *set* converts a list into a set. Type synonyms are introduced via the *type\_synonym* command. Algebraic data types are defined with the keyword *datatype*. Non-recursive definitions are introduced with the *definition* keyword.

It is standard to define a language as a set of strings over a finite set of symbols. We deviate slightly by introducing a type variable  $'a$  for the type of symbols. Thus a string corresponds to a list of symbols and a language is formalized as a set of lists of symbols, a symbol being either a terminal or a non-terminal. We represent a context-free grammar as the datatype *CFG*. An instance  $\mathcal{G}$  consists of (1) a list of non-terminals ( $\mathfrak{N} \mathcal{G}$ ), (2) a list of terminals ( $\mathfrak{T} \mathcal{G}$ ), (3) a list of production rules ( $\mathfrak{R} \mathcal{G}$ ), and a start symbol ( $\mathfrak{S} \mathcal{G}$ ) where  $\mathfrak{N}$ ,  $\mathfrak{T}$ ,  $\mathfrak{R}$  and  $\mathfrak{S}$  are projections accessing the specific part of an instance  $\mathcal{G}$  of the datatype *CFG*. Each rule consists of a left-hand side or *rule-head*, a single symbol, and a right-hand side or *rule-body*, a list of symbols. The productions with a particular non-terminal  $N$  on their left-hand sides are called the alternatives of  $N$ . We make the

usual assumptions about the well-formedness of a context-free grammar: the intersection of the set of terminals and the set of non-terminals is empty; the start symbol is a non-terminal; the rule head of a production is a non-terminal and its rule body consists of only symbols. Additionally, since we are working with a list of productions, we make the assumption that this list is distinct.

**type-synonym**  $'a \text{ rule} = 'a \times 'a \text{ list}$

**type-synonym**  $'a \text{ rules} = 'a \text{ rule list}$

**datatype**  $'a \text{ cfg} =$

$CFG (\mathfrak{N} : 'a \text{ list}) (\mathfrak{T} : 'a \text{ list}) (\mathfrak{R} : 'a \text{ rules}) (\mathfrak{S} : 'a)$

**definition**  $\text{rule-head} :: 'a \text{ rule} \Rightarrow 'a \text{ where}$

$\text{rule-head} = \text{fst}$

**definition**  $\text{rule-body} :: 'a \text{ rule} \Rightarrow 'a \text{ list where}$

$\text{rule-body} = \text{snd}$

**definition**  $\text{disjunct-symbols} :: 'a \text{ cfg} \Rightarrow \text{bool where}$

$\text{disjunct-symbols } \mathcal{G} \equiv \text{set } (\mathfrak{N} \mathcal{G}) \cap \text{set } (\mathfrak{T} \mathcal{G}) = \{\}$

**definition**  $\text{wf-startsymbol} :: 'a \text{ cfg} \Rightarrow \text{bool where}$

$\text{wf-startsymbol } \mathcal{G} \equiv \mathfrak{S} \mathcal{G} \in \text{set } (\mathfrak{N} \mathcal{G})$

**definition**  $\text{wf-rules} :: 'a \text{ cfg} \Rightarrow \text{bool where}$

$\text{wf-rules } \mathcal{G} \equiv \forall (N, \alpha) \in \text{set } (\mathfrak{R} \mathcal{G}). N \in \text{set } (\mathfrak{N} \mathcal{G}) \wedge (\forall s \in \text{set } \alpha. s \in \text{set } (\mathfrak{N} \mathcal{G}) \cup \text{set } (\mathfrak{T} \mathcal{G}))$

**definition**  $\text{distinct-rules} :: 'a \text{ cfg} \Rightarrow \text{bool where}$

$\text{distinct-rules } \mathcal{G} \equiv \text{distinct } (\mathfrak{R} \mathcal{G})$

**definition**  $\text{wf-}\mathcal{G} :: 'a \text{ cfg} \Rightarrow \text{bool where}$

$\text{wf-}\mathcal{G} \mathcal{G} \equiv \text{disjunct-symbols } \mathcal{G} \wedge \text{wf-startsymbol } \mathcal{G} \wedge \text{wf-rules } \mathcal{G} \wedge \text{distinct-rules } \mathcal{G}$

Furthermore, in Isabelle, lists are constructed from the empty list  $[]$  via the infix cons-operator  $(\#)$ ; the operator  $(@)$  appends two lists;  $|xs|$  denotes the length and  $xs ! n$  returns the  $n$ -th item of the list  $xs$ . Sets follow the standard mathematical notation including the commonly found set builder notation or set comprehensions  $\{x \mid P x\}$ . Sets can also be defined inductively using the keyword *inductive\_set*.

Next we formalize the concept of a derivation. We use lowercase letters  $a, b, c$  indicating terminal symbols; capital letters  $A, B, C$  denote non-terminals; lists of symbols are represented by greek letters:  $\alpha, \beta, \gamma$ , occasionally also by lowercase letters  $u, v, w$ . The empty list in the context of a language is  $\epsilon$ . A sentential is a list consisting of only



symbols. A sentence is a sentential if it only contains terminal symbols. We first define a predicate  $derives1 \mathcal{G} u v$  which expresses that we can derive  $v$  from  $u$  in a single step or the predicate holds if there exist  $\alpha, \beta, N$  and  $\gamma$  such that  $u = \alpha @ [N] @ \beta$ ,  $v = \alpha @ \gamma @ \beta$  and  $(N, \gamma)$  is a production rule. We also introduce some slightly more convenient notation:  $derives1 \mathcal{G} u v$  is written  $\mathcal{G} \vdash u \Rightarrow v$  in the following. We then can define the set of single-step derivations using  $derives1$ , and subsequently the set of all derivations given a particular grammar is the reflexive-transitive closure of the set of single-step derivations. Finally, we say  $v$  can be derived from  $u$  given a grammar  $\mathcal{G}$  or  $derives \mathcal{G} u v$  if  $(u, v) \in derivations \mathcal{G}$ . A slightly more convenient notation is again:  $derives \mathcal{G} u v = \mathcal{G} \vdash u \Rightarrow^* v$

**type-synonym**  $'a \text{ sentential} = 'a \text{ list}$

**definition**  $is\_terminal :: 'a \text{ cfg} \Rightarrow 'a \Rightarrow bool$  **where**  
 $is\_terminal \mathcal{G} s \equiv s \in set (\mathfrak{T} \mathcal{G})$

**definition**  $is\_nonterminal :: 'a \text{ cfg} \Rightarrow 'a \Rightarrow bool$  **where**  
 $is\_nonterminal \mathcal{G} s \equiv s \in set (\mathfrak{N} \mathcal{G})$

**definition**  $is\_symbol :: 'a \text{ cfg} \Rightarrow 'a \Rightarrow bool$  **where**  
 $is\_symbol \mathcal{G} s \equiv is\_terminal \mathcal{G} s \vee is\_nonterminal \mathcal{G} s$

**definition**  $wf\_sentential :: 'a \text{ cfg} \Rightarrow 'a \text{ sentential} \Rightarrow bool$  **where**  
 $wf\_sentential \mathcal{G} s \equiv \forall x \in set s. is\_symbol \mathcal{G} x$

**definition**  $is\_sentence :: 'a \text{ cfg} \Rightarrow 'a \text{ sentential} \Rightarrow bool$  **where**  
 $is\_sentence \mathcal{G} s \equiv \forall x \in set s. is\_terminal \mathcal{G} x$

**definition**  $derives1 :: 'a \text{ cfg} \Rightarrow 'a \text{ sentential} \Rightarrow 'a \text{ sentential} \Rightarrow bool$  **where**  
 $derives1 \mathcal{G} u v \equiv$   
 $\exists \alpha \beta N \gamma.$   
 $u = \alpha @ [N] @ \beta$   
 $\wedge v = \alpha @ \gamma @ \beta$   
 $\wedge (N, \gamma) \in set (\mathfrak{R} \mathcal{G})$

**definition**  $derivations1 :: 'a \text{ cfg} \Rightarrow ('a \text{ sentential} \times 'a \text{ sentential}) \text{ set}$  **where**  
 $derivations1 \mathcal{G} = \{ (u, v) \mid u v. \mathcal{G} \vdash u \Rightarrow v \}$

**definition**  $derivations :: 'a \text{ cfg} \Rightarrow ('a \text{ sentential} \times 'a \text{ sentential}) \text{ set}$  **where**  
 $derivations \mathcal{G} = (derivations1 \mathcal{G})^*$

**definition**  $derives :: 'a \text{ cfg} \Rightarrow 'a \text{ sentential} \Rightarrow 'a \text{ sentential} \Rightarrow bool$  **where**  
 $derives \mathcal{G} u v \equiv (u, v) \in derivations \mathcal{G}$

Potentially recursive but provably total functions that may make use of pattern matching are defined with the *fun* and *function* keywords; partial functions are defined via *partial\_function*. Take for example the function *slice* defined below. Term *slice xs i j* computes the slice of a list *xs* between indices *i* (inclusive) and *j* (exclusive), e.g. *slice [a, b, c, d, e] 2 4* evaluates to *[c, d]*. We also introduce a shorthand notation: e.g. *slice xs i j* is written *xs[i..j]* in the following.

```
fun slice :: 'a list  $\Rightarrow$  nat  $\Rightarrow$  nat  $\Rightarrow$  'a list where
  slice [] - - = []
| slice (x#xs) - 0 = []
| slice (x#xs) 0 (Suc b) = x # slice xs 0 b
| slice (x#xs) (Suc a) (Suc b) = slice xs a b
```

Lemmas, theorems and corollaries are presented using the keywords *lemma*, *theorem*, *corollary* respectively, followed by their names. They consist of zero or more assumptions marked by *assumes* keywords and one conclusion indicated by *shows*. E.g. we can proof a simple lemma about the interaction between the *slice* function and the append operator (@), stating the conditions under which we can split one slice into two.

**lemma** slice-append:

```
assumes i  $\leq$  j
assumes j  $\leq$  k
shows xs[i..j] @ xs[j..k] = xs[i..k]
```

## 3.2 The Formalized Algorithm

Next we formalize the algorithm presented in Chapter 2. First we define the datatype *item* representing Earley items. For example, the item  $S \rightarrow S + \bullet S, 2, 4$  consists of four parts: a production rule (*item-rule*), a natural number (*item-bullet*) indicating the position of the bullet in the production rule, and two natural numbers (*item-origin* inclusive, *item-end* exclusive) representing the portion of the input string  $\omega$  that has been parsed by the item. Additionally, we introduce a few useful abbreviations: the functions *item-rule-head* and *item-rule-body* access the *rule-head* respectively *rule-body* of an item. Functions *item- $\alpha$*  and *item- $\beta$*  split the production rule body at the bullet, e.g.  $S \rightarrow \alpha \bullet \beta$ . We call an item *complete* if the bullet is at the end of the production rule body. The next symbol (*next-symbol*) of an item is either *None* if it is complete, or *Some s* where *s* is the symbol in the production rule body following the bullet. An item is finished if the item rule head is the start symbol, the item is complete, and the whole input has been parsed or *item-origin item* = 0 and *item-end item* =  $|\omega|$ . Finally, we call a set of items *recognizing* if it contains at least one finished item, indicating that this set of items recognizes the input  $\omega$ .

**datatype** 'a item =  
*Item* (*item-rule*: 'a rule) (*item-bullet* : nat) (*item-origin* : nat) (*item-end* : nat)

**type-synonym** 'a items = 'a item set

**definition** *item-rule-head* :: 'a item  $\Rightarrow$  'a **where**  
*item-rule-head* x = rule-head (*item-rule* x)

**definition** *item-rule-body* :: 'a item  $\Rightarrow$  'a sentential **where**  
*item-rule-body* x = rule-body (*item-rule* x)

**definition** *item- $\alpha$*  :: 'a item  $\Rightarrow$  'a sentential **where**  
*item- $\alpha$*  x = take (*item-bullet* x) (*item-rule-body* x)

**definition** *item- $\beta$*  :: 'a item  $\Rightarrow$  'a sentential **where**  
*item- $\beta$*  x = drop (*item-bullet* x) (*item-rule-body* x)

**definition** *is-complete* :: 'a item  $\Rightarrow$  bool **where**  
*is-complete* x  $\equiv$  *item-bullet* x  $\geq$  |*item-rule-body* x|

**definition** *next-symbol* :: 'a item  $\Rightarrow$  'a option **where**  
*next-symbol* x  $\equiv$  if *is-complete* x then None else Some (*item-rule-body* x ! *item-bullet* x)

**definition** *is-finished* :: 'a cfg  $\Rightarrow$  'a sentential  $\Rightarrow$  'a item  $\Rightarrow$  bool **where**  
*is-finished*  $\mathcal{G}$   $\omega$  x  $\equiv$   
*item-rule-head* x =  $\mathcal{S}$   $\mathcal{G}$   $\wedge$   
*item-origin* x = 0  $\wedge$   
*item-end* x = | $\omega$ |  $\wedge$   
*is-complete* x

**definition** *recognizing* :: 'a items  $\Rightarrow$  'a cfg  $\Rightarrow$  'a sentential  $\Rightarrow$  bool **where**  
*recognizing* I  $\mathcal{G}$   $\omega$   $\equiv \exists x \in I. \text{is-finished } \mathcal{G} \omega x$

Normally we don't construct items directly via the *Item* constructor but use two auxiliary constructors: the function *init-item* is used by the *Init* and *Predict* operations. It sets the *item-bullet* to 0 or the beginning of the production rule body, initializes the *item-rule*, and indicates that this is an initial item by assigning *item-origin* and *item-end* to the current position in the input. The function *inc-item* returns a new item, moving the bullet over the next symbol (assuming there is one), and setting the *item-end* to the current position in the input, leaving the item rule and origin untouched. It is utilized by the *Scan* and *Complete* operations.

**definition** *init-item* :: 'a rule  $\Rightarrow$  nat  $\Rightarrow$  'a item **where**  
*init-item* r k = *Item* r 0 k k

**definition**  $inc\text{-}item :: 'a\ item \Rightarrow nat \Rightarrow 'a\ item$  **where**

$$inc\text{-}item\ x\ k = Item\ (item\text{-}rule\ x)\ (item\text{-}bullet\ x + 1)\ (item\text{-}origin\ x)\ k$$

There are different approaches of defining the set of Earley items in accordance with the rules of Figure 2.1. We can take an abstract approach and define the set inductively using Isabelle's inductive sets, or a more operational point of view. We take the latter approach and discuss the reasoning for this decision end the end of this section.

Note that, as mentioned previously, even though we are only constructing one set of Earley items, conceptually all items with the same item end form one Earley bin. Our operational approach is then the following: we generate Earley items bin by bin in ascending order, starting from the 0-th bin that contains all initial items computed by the *Init* operation. The three operations *Scan*, *Predict*, and *Complete* all take as arguments the index of the current bin and the current set of Earley items. For the  $k$ -th bin the *Scan* operation initializes the  $k + 1$ -st bin, whereas the *Predict* and *Complete* operations only generate items belonging to the  $k$ -th bin. We then define a function *Earley-step* that returns the union of the set itself and applying the three operations to a set of Earley items. We complete the  $k$ -th bin and initialize the  $k + 1$ -th bin by iterating *Earley-step* until the set of items converges, captured by the *Earley-bin* definition. The function *Earley* then generates the bins up to the  $n$ -th bin by applying the *Earley-bin* function first to the initial set of items *Init* and continuing in ascending order bin by bin. Finally, we compute the set of Earley items by applying function *Earley* to the length of the input.

**definition**  $bin :: 'a\ items \Rightarrow nat \Rightarrow 'a\ items$  **where**

$$bin\ I\ k = \{ x . x \in I \wedge item\text{-}end\ x = k \}$$

**definition**  $Init :: 'a\ cfg \Rightarrow 'a\ items$  **where**

$$Init\ \mathcal{G} = \{ init\text{-}item\ r\ 0 \mid r. r \in set\ (\mathfrak{R}\ \mathcal{G}) \wedge fst\ r = (\mathfrak{S}\ \mathcal{G}) \}$$

**definition**  $Scan :: nat \Rightarrow 'a\ sentential \Rightarrow 'a\ items \Rightarrow 'a\ items$  **where**

$$\begin{aligned} Scan\ k\ \omega\ I = \\ \{ inc\text{-}item\ x\ (k+1) \mid x\ a. \\ x \in bin\ I\ k \wedge \\ \omega!k = a \wedge \\ k < |\omega| \wedge \\ next\text{-}symbol\ x = Some\ a \} \end{aligned}$$

**definition**  $Predict :: nat \Rightarrow 'a\ cfg \Rightarrow 'a\ items \Rightarrow 'a\ items$  **where**

$$\begin{aligned} Predict\ k\ \mathcal{G}\ I = \\ \{ init\text{-}item\ r\ k \mid r\ x. \\ r \in set\ (\mathfrak{R}\ \mathcal{G}) \wedge \\ x \in bin\ I\ k \} \end{aligned}$$

$next-symbol\ x = Some\ (rule-head\ r)\ }$

**definition**  $Complete :: nat \Rightarrow 'a\ items \Rightarrow 'a\ items\ \mathbf{where}$

$Complete\ k\ I =$   
 $\{ inc-item\ x\ k \mid x\ y.$   
 $x \in bin\ I\ (item-origin\ y) \wedge$   
 $y \in bin\ I\ k \wedge$   
 $is-complete\ y \wedge$   
 $next-symbol\ x = Some\ (item-rule-head\ y)\ }$

**definition**  $Earley-step :: nat \Rightarrow 'a\ cfg \Rightarrow 'a\ sentential \Rightarrow 'a\ items \Rightarrow 'a\ items\ \mathbf{where}$

$Earley-step\ k\ \mathcal{G}\ \omega\ I = I \cup Scan\ k\ \omega\ I \cup Complete\ k\ I \cup Predict\ k\ \mathcal{G}\ I$

**fun**  $funpower :: ('a \Rightarrow 'a) \Rightarrow nat \Rightarrow ('a \Rightarrow 'a)\ \mathbf{where}$

$funpower\ f\ 0\ x = x$   
 $\mid funpower\ f\ (Suc\ n)\ x = f\ (funpower\ f\ n\ x)$

**definition**  $natUnion :: (nat \Rightarrow 'a\ set) \Rightarrow 'a\ set\ \mathbf{where}$

$natUnion\ f = \bigcup \{ f\ n \mid n. True \}$

**definition**  $limit :: ('a\ set \Rightarrow 'a\ set) \Rightarrow 'a\ set \Rightarrow 'a\ set\ \mathbf{where}$

$limit\ f\ x = natUnion\ (\lambda\ n. funpower\ f\ n\ x)$

**definition**  $Earley-bin :: nat \Rightarrow 'a\ cfg \Rightarrow 'a\ sentential \Rightarrow 'a\ items \Rightarrow 'a\ items\ \mathbf{where}$

$Earley-bin\ k\ \mathcal{G}\ \omega\ I = limit\ (Earley-step\ k\ \mathcal{G}\ \omega)\ I$

**fun**  $Earley :: nat \Rightarrow 'a\ cfg \Rightarrow 'a\ sentential \Rightarrow 'a\ items\ \mathbf{where}$

$Earley\ 0\ \mathcal{G}\ \omega = Earley-bin\ 0\ \mathcal{G}\ \omega\ (Init\ \mathcal{G})$   
 $\mid Earley\ (Suc\ n)\ \mathcal{G}\ \omega = Earley-bin\ (Suc\ n)\ \mathcal{G}\ \omega\ (Earley\ n\ \mathcal{G}\ \omega)$

**definition**  $\mathcal{E}arley :: 'a\ cfg \Rightarrow 'a\ sentential \Rightarrow 'a\ items\ \mathbf{where}$

$\mathcal{E}arley\ \mathcal{G}\ \omega = Earley\ |\omega|\ \mathcal{G}\ \omega$

We follow the operational approach of defining the set of Earley items primarily for two reasons: first of all, we reuse and only slightly adapt most of the basic definitions of this chapter from the work of Obua on *Local Lexing* [Obua:2017] [LocalLexing-AFP], who takes the more operational approach and already defines useful lemmas, for example on function iteration. Secondly, the operational approach maps more easily to the list-based implementation of the next chapter that necessarily takes an ordered approach to generating Earley items. Nonetheless, in hindsight, defining the set of Earley items inductively seems to be not only the more elegant approach but also might simplify some of the proofs of this chapter, and is consequently future work worth considering.

### 3.3 Well-formedness

Due to the operational view of generating the set of Earley items, the proofs of, not only, well-formedness, but also soundness and completeness follow a similar structure: we first proof a property about the basic building blocks, the *Init*, *Scan*, *Predict*, and *Complete* operations. Then we proof that this property is maintained iterating the function *Earley-step*, and thus holds for the *Earley-bin* operation. Finally, we show that the function *Earley* maintains this property for all bins and thus for the *Earley* definition, or the set of Earley items.

Before we start to proof soundness and completeness of the generated set of Earley items, especially the completeness proof is more involved, we highlight the general proof structure once in detail, for a simpler property: well-formedness of the items, allowing us to concentrate only on the core aspects for the soundness and completeness proofs.

An Earley item is well-formed (*wf-item*) if the item rule is a rule of the grammar; the item bullet is bounded by the length of the item rule body; the item origin does not exceed the item end, and finally the item end is at most the length of the input.

**definition** *wf-item* :: 'a cfg  $\Rightarrow$  'a sentential  $\Rightarrow$  'a item  $\Rightarrow$  bool **where**

*wf-item*  $\mathcal{G} \ \omega \ x \equiv$   
 $\text{item-rule } x \in \text{set } (\mathfrak{R} \ \mathcal{G}) \wedge$   
 $\text{item-bullet } x \leq |\text{item-rule-body } x| \wedge$   
 $\text{item-origin } x \leq \text{item-end } x \wedge$   
 $\text{item-end } x \leq |\omega|$

**definition** *wf-items* :: 'a cfg  $\Rightarrow$  'a sentential  $\Rightarrow$  'a items  $\Rightarrow$  bool **where**

*wf-items*  $\mathcal{G} \ \omega \ I \equiv \forall x \in I. \text{wf-item } \mathcal{G} \ \omega \ x$

**lemma** *wf-Init*:

**shows** *wf-items*  $\mathcal{G} \ \omega \ (\text{Init } \mathcal{G})$

**lemma** *wf-Scan-Predict-Complete*:

**assumes** *wf-items*  $\mathcal{G} \ \omega \ I$

**shows** *wf-items*  $\mathcal{G} \ \omega \ (\text{Scan } k \ \omega \ I \cup \text{Predict } k \ \mathcal{G} \ I \cup \text{Complete } k \ I)$

**lemma** *wf-Earley-step*:

**assumes** *wf-items*  $\mathcal{G} \ \omega \ I$

**shows** *wf-items*  $\mathcal{G} \ \omega \ (\text{Earley-step } k \ \mathcal{G} \ \omega \ I)$

Lemmas *wf-Init*, *wf-Scan-Predict-Complete*, and *wf-Earley-step* follow trivially by definition of the respective operations.

**lemma** *wf-funpower*:  
**assumes** *wf-items*  $\mathcal{G} \ \omega \ I$   
**shows** *wf-items*  $\mathcal{G} \ \omega \ (\text{funpower } (\text{Earley-step } k \ \mathcal{G} \ \omega) \ n \ I)$

**lemma** *wf-Earley-bin*:  
**assumes** *wf-items*  $\mathcal{G} \ \omega \ I$   
**shows** *wf-items*  $\mathcal{G} \ \omega \ (\text{Earley-bin } k \ \mathcal{G} \ \omega \ I)$

**lemma** *wf-Earley-bin0*:  
**shows** *wf-items*  $\mathcal{G} \ \omega \ (\text{Earley-bin } 0 \ \mathcal{G} \ \omega \ (\text{Init } \mathcal{G}))$

We proof the lemma *wf-funpower* by induction on  $n$  using lemma *wf-Earley-step*, and lemmas *wf-Earley-bin* and *wf-Earley-bin0* follow immediately using additionally the fact that  $x \in \text{limit } f \ X \equiv \exists n. x \in \text{funpower } f \ n \ X$  and lemma *wf-Init*.

**lemma** *wf-Earley*:  
**shows** *wf-items*  $\mathcal{G} \ \omega \ (\text{Earley } n \ \mathcal{G} \ \omega)$

**lemma** *wf-Earley*:  
**shows** *wf-items*  $\mathcal{G} \ \omega \ (\mathcal{E} \text{arley } \mathcal{G} \ \omega)$

Finally, lemma *wf-Earley* is proved by induction on  $n$  using lemmas *wf-Earley-bin* and *wf-Earley-bin0*; lemma *wf-Earley* follows by definition of *E*arley.

### 3.4 Soundness

Next we proof the soundness of the generated items, or:  $A \rightarrow \alpha \bullet \beta, i, j \in B$  implies  $A \xRightarrow{*} \omega[i..j]\beta$  which is stated in terms of our formalization by the *sound-item* definition below. As mentioned previously, the general proof structure follows the proof for well-formedness. Thus, we only highlight one slightly more involved lemma stating the soundness of the *Complete* operation while stating the remaining lemmas without explicit proof. Additionally, proving lemma *sound-Complete* provides some insight into working with and proving properties about derivations.

**definition** *sound-item* :: '*a* cfg  $\Rightarrow$  '*a* sentential  $\Rightarrow$  '*a* item  $\Rightarrow$  bool **where**  
*sound-item*  $\mathcal{G} \ \omega \ x = \mathcal{G} \vdash [\text{item-rule-head } x] \Rightarrow^* \omega[\text{item-origin } x.. \text{item-end } x] @ \text{item-}\beta \ x$

**definition** *sound-items* :: '*a* cfg  $\Rightarrow$  '*a* sentential  $\Rightarrow$  '*a* items  $\Rightarrow$  bool **where**  
*sound-items*  $\mathcal{G} \ \omega \ I \equiv \forall x \in I. \text{sound-item } \mathcal{G} \ \omega \ x$

Obua [Obua:2017] [LocalLexing-AFP] defines derivations at two different abstraction levels. The first representation is as the reflexive-transitive closure of the set of one-step derivations as introduced earlier in this chapter. The second representation is again more operational. He defines a predicate *Derives1*  $\mathcal{G} \ u \ i \ r \ v$  that is conceptually

analogous to the predicate  $\mathcal{G} \vdash u \Rightarrow v$  but also captures the rule  $r$  used for a single rewriting step and the position  $i$  in  $u$  where the rewriting occurs.

**definition** *Derives1* :: 'a cfg  $\Rightarrow$  'a sentential  $\Rightarrow$  nat  $\Rightarrow$  'a rule  $\Rightarrow$  'a sentential  $\Rightarrow$  bool **where**  
*Derives1*  $\mathcal{G} \ u \ i \ r \ v \equiv$   
 $\exists \alpha \ \beta \ N \ \gamma.$   
 $u = \alpha @ [N] @ \beta$   
 $\wedge v = \alpha @ \gamma @ \beta$   
 $\wedge (N, \gamma) \in \text{set } (\mathfrak{R} \ \mathcal{G})$   
 $\wedge r = (N, \gamma) \wedge i = |\alpha|$

He then defines the type of a *derivation* as a list of pairs representing precisely the positions and rules used to apply each rewrite step. The predicate *Derivation* is defined recursively as follows: *Derivation*  $\alpha \ [] \ \beta$  holds only if  $\alpha = \beta$ . If the derivation consists of at least one rewrite pair  $(i, r)$ , or *Derivation*  $\mathcal{G} \ \alpha \ ((i, r) \# D) \ \beta$ , then there must exist a  $\gamma$  such that *Derives1*  $\mathcal{G} \ \alpha \ i \ r \ \gamma$  and *Derivation*  $\mathcal{G} \ \gamma \ D \ \beta$ . Note that we introduce once again a more convenient notation: e.g. *Derivation*  $\alpha \ D \ \beta$  is written  $\mathcal{G} \vdash \alpha \Rightarrow^D \beta$  in the following. Obua then proves that both notions of a derivation are equivalent (lemma *derives-equiv-Derivation*)

**type-synonym** 'a derivation = (nat  $\times$  'a rule) list

**fun** *Derivation* :: 'a cfg  $\Rightarrow$  'a sentential  $\Rightarrow$  'a derivation  $\Rightarrow$  'a sentential  $\Rightarrow$  bool **where**  
*Derivation* -  $\alpha \ [] \ \beta = (\alpha = \beta)$   
 $| \text{Derivation } \mathcal{G} \ \alpha \ (d \# D) \ \beta = (\exists \gamma. \text{Derives1 } \mathcal{G} \ \alpha \ (\text{fst } d) \ (\text{snd } d) \ \gamma \wedge \text{Derivation } \mathcal{G} \ \gamma \ D \ \beta)$

**lemma** *derives-equiv-Derivation*:

**shows**  $\mathcal{G} \vdash \alpha \Rightarrow^* \beta \equiv \exists D. \mathcal{G} \vdash \alpha \Rightarrow^D \beta$

Next we state a small but useful lemma about rewriting derivations using the more operational definition of derivations defined above without explicit proof.

**lemma** *Derivation-append-rewrite*:

**assumes**  $\mathcal{G} \vdash \alpha \Rightarrow^D \beta @ \gamma @ \delta$   
**assumes**  $\mathcal{G} \vdash \gamma \Rightarrow^E \gamma'$   
**shows**  $\exists F. \mathcal{G} \vdash \alpha \Rightarrow^F \beta @ \gamma' @ \delta$

And finally, we proof soundness of the *Complete* operation:

**lemma** *sound-Complete*:

**assumes** *wf*: wf-items  $\mathcal{G} \ \omega \ I$   
**assumes** *sound*: sound-items  $\mathcal{G} \ \omega \ I$   
**shows** sound-items  $\mathcal{G} \ \omega \ (\text{Complete } k \ I)$

*Proof.* Let  $z$  denote an arbitrary but fixed item of *Complete*  $k \ I$ . By the definition of the *Complete* operation there exist items  $x$  and  $y$  such that:



$$x \in \text{bin } I \text{ (item-origin } y) \quad (1) \quad \text{next-symbol } x = \text{Some (item-rule-head } y) \quad (2)$$

$$y \in \text{bin } I \ k \quad (3) \quad \text{is-complete } y \quad (4)$$

$$z = \text{inc-item } x \ k \quad (5)$$

Since  $y$  is in bin  $k$  (3), it is complete (4) and the set  $I$  is sound (assumption *sound*), there exists a derivation  $E$  such that

$$\mathcal{G} \vdash [\text{item-rule-head } y] \Rightarrow^E \omega[\text{item-origin } y.. \text{item-end } y] \quad (6)$$

by lemma *derives-equiv-Derivation*. Similarly, since  $x$  is in bin *item-origin*  $y$  (1) and due to assumption *sound*, there exists a derivation  $D$  such that

$$\mathcal{G} \vdash [\text{item-rule-head } x] \Rightarrow^D \omega[\text{item-origin } x.. \text{item-origin } y] @ \text{item-}\beta \ x \quad (7)$$

Note that  $\text{item-}\beta \ x = \text{item-rule-head } y \# \text{tl (item-}\beta \ x)$  since the next symbol of  $x$  is equal to the item rule head of  $y$  (2). Thus, by lemma *Derivation-append-rewrite*, and the definition of  $D$  (7) and  $E$  (6), there exists a derivation  $F$  such that

$$\begin{aligned} \mathcal{G} \vdash [\text{item-rule-head } x] \Rightarrow^F \omega[\text{item-origin } x.. \text{item-origin } y] @ \\ \omega[\text{item-origin } y.. \text{item-end } y] @ \text{tl (item-}\beta \ x) \end{aligned}$$

Additionally, we know that  $x$  and  $y$  are well-formed items due to the facts that  $x$  is in bin *item-origin*  $y$  (1),  $y$  is in bin  $k$  (3), and the assumption *wf-items*  $\mathcal{G} \ \omega \ I$ . Thus, we can discharge the assumptions of lemma *slice-append* ( $\text{item-origin } x \leq \text{item-origin } y$  and  $\text{item-origin } y \leq \text{item-end } y$ ) and have

$$\mathcal{G} \vdash [\text{item-rule-head } x] \Rightarrow^F \omega[\text{item-origin } x.. \text{item-end } y] @ \text{tl (item-}\beta \ x)$$

Moreover, since  $z = \text{inc-item } x \ k$  (5) and the next symbol of  $x$  is the item rule head of  $y$  (2), it follows that  $\text{tl (item-}\beta \ x) = \text{item-}\beta \ z$ , and ultimately *sound-item*  $\mathcal{G} \ \omega \ z$ , again by the definition of  $z$  (5) and lemma *derives-equiv-Derivation*. □

**lemma** *sound-Init*:

**shows** *sound-items*  $\mathcal{G} \ \omega \ (\text{Init } \mathcal{G})$

**lemma** *sound-Scan*:

**assumes** *wf-items*  $\mathcal{G} \ \omega \ I$   
**assumes** *sound-items*  $\mathcal{G} \ \omega \ I$   
**shows** *sound-items*  $\mathcal{G} \ \omega \ (\text{Scan } k \ \omega \ I)$

**lemma** *sound-Predict*:

**assumes** *sound-items*  $\mathcal{G} \ \omega \ I$   
**shows** *sound-items*  $\mathcal{G} \ \omega \ (\text{Predict } k \ \mathcal{G} \ I)$

**lemma** *sound-Earley-step*:

**assumes** *wf-items*  $\mathcal{G} \ \omega \ I$   
**assumes** *sound-items*  $\mathcal{G} \ \omega \ I$   
**shows** *sound-items*  $\mathcal{G} \ \omega \ (\text{Earley-step } k \ \mathcal{G} \ \omega \ I)$

**lemma** *sound-funpower*:

**assumes** *wf-items*  $\mathcal{G} \ \omega \ I$   
**assumes** *sound-items*  $\mathcal{G} \ \omega \ I$   
**shows** *sound-items*  $\mathcal{G} \ \omega \ (\text{funpower } (\text{Earley-step } k \ \mathcal{G} \ \omega) \ n \ I)$

**lemma** *sound-Earley-bin*:

**assumes** *wf-items*  $\mathcal{G} \ \omega \ I$   
**assumes** *sound-items*  $\mathcal{G} \ \omega \ I$   
**shows** *sound-items*  $\mathcal{G} \ \omega \ (\text{Earley-bin } k \ \mathcal{G} \ \omega \ I)$

**lemma** *sound-Earley-bin0*:

**shows** *sound-items*  $\mathcal{G} \ \omega \ (\text{Earley-bin } 0 \ \mathcal{G} \ \omega \ (\text{Init } \mathcal{G}))$

**lemma** *sound-Earley*:

**shows** *sound-items*  $\mathcal{G} \ \omega \ (\text{Earley } k \ \mathcal{G} \ \omega)$

**lemma** *sound- $\mathcal{E}$ arley*:

**shows** *sound-items*  $\mathcal{G} \ \omega \ (\mathcal{E}arley \ \mathcal{G} \ \omega)$

Finally, using *sound- $\mathcal{E}$ arley* and the definitions of *sound-item*, *recognizing*, *is-finished* and *is-complete* the final theorem follows: if the generated set of Earley items is *recognizing*, or contains a *finished* item, then there exists a derivation of the input  $\omega$  from the start symbol of the grammar.

**theorem** *soundness*:

**assumes** *recognizing*  $(\mathcal{E}arley \ \mathcal{G} \ \omega) \ \mathcal{G} \ \omega$   
**shows**  $\mathcal{G} \vdash [\mathcal{S} \ \mathcal{G}] \Rightarrow^* \omega$

### 3.5 Completeness

Next we prove completeness and consequently obtain a concluded correctness proof using theorem *soundness*. The completeness proof is by far the most involved proof of this

chapter. Thus we present it in greater detail, and also slightly deviate from the proof structure of the well-formedness and soundness proofs presented previously. We directly start to prove three properties of the *Earley* function that correspond conceptually to the three different operations that can occur while generating the bins.

We need three simple lemmas concerning the *Earley-bin* function, stated without explicit proof: (1) *Earley-bin*  $k \mathcal{G} \omega I$  only (potentially) changes bins  $k$  and  $k + 1$  (lemma *Earley-bin-bin-idem*); (2) the *Earley-step* operation is subsumed by the *Earley-bin* operation, since it computes the limit of *Earley-step* (lemma *Earley-step-sub-Earley-bin*); and (3) the function *Earley-bin* is idempotent (lemma *Earley-bin-idem*).

**lemma** *Earley-bin-bin-idem*:

**assumes**  $i \neq k$

**assumes**  $i \neq k + 1$

**shows**  $\text{bin } (\text{Earley-bin } k \mathcal{G} \omega I) i = \text{bin } I i$

**lemma** *Earley-step-sub-Earley-bin*:

**shows**  $\text{Earley-step } k \mathcal{G} \omega I \subseteq \text{Earley-bin } k \mathcal{G} \omega I$

**lemma** *Earley-bin-idem*:

**shows**  $\text{Earley-bin } k \mathcal{G} \omega (\text{Earley-bin } k \mathcal{G} \omega I) = \text{Earley-bin } k \mathcal{G} \omega I$

Next, we proof lemma *Scan-Earley* in detail: it describes under which assumptions the function *Earley* generates a 'scanned' item:

**lemma** *Scan-Earley*:

**assumes**  $i + 1 \leq k$

**assumes**  $x \in \text{bin } (\text{Earley } k \mathcal{G} \omega) i$

**assumes**  $\text{next-symbol } x = \text{Some } a$

**assumes**  $k \leq |\omega|$

**assumes**  $\omega ! i = a$

**shows**  $\text{inc-item } x (i + 1) \in \text{Earley } k \mathcal{G} \omega$

*Proof.* The proof is by induction in  $k$  for arbitrary  $i$ ,  $x$ , and  $a$ :

The base case  $k = 0$  is trivial, since we have the assumption  $i + 1 \leq 0$ .

For the induction step we can assume

$$i + 1 \leq k + 1 \quad (1) \quad k + 1 \leq |\omega| \quad (2)$$

$$x \in \text{bin } (\text{Earley } (k + 1) \mathcal{G} \omega) i \quad (3) \quad \text{next-symbol } x = \text{Some } a \quad (4)$$

$$\omega ! i = a \quad (5)$$

Assumptions (1) and (3) imply that  $x \in \text{bin } (\text{Earley } k \mathcal{G} \omega) i$  by lemma *Earley-bin-bin-idem*.

We then consider two cases:

- $i + 1 \leq k$ : We can apply the induction hypothesis using assumptions (2), (4), (5), and fact  $x \in \text{bin } (\text{Earley } k \mathcal{G} \omega) \ i$  and have  $\text{inc-item } x \ (i + 1) \in \text{Earley } k \mathcal{G} \omega$ . The statement to proof follows by lemma *Earley-step-sub-Earley-bin* and the definition of *Earley-step*.
- $k < i + 1$ : hence we have  $i = k$  by assumption (1). Thus, we have  $\text{inc-item } x \ (i + 1) \in \text{Scan } k \omega \ (\text{Earley } k \mathcal{G} \omega)$  using assumptions (2), (4), (5), and fact  $x \in \text{bin } (\text{Earley } k \mathcal{G} \omega) \ i$  by the definition of the *Scan* operation. This in turn implies  $\text{inc-item } x \ (i + 1) \in \text{Earley-step } k \mathcal{G} \omega \ (\text{Earley } k \mathcal{G} \omega)$  by lemma *Earley-step-sub-Earley-bin* and the definition of *Earley-step*. Since the function *Earley-bin* is idempotent (lemma *Earley-bin-idem*), we have  $\text{inc-item } x \ (i + 1) \in \text{Earley } k \mathcal{G} \omega$  by definition of *Earley*. And again, the final statement follows by lemma *Earley-step-sub-Earley-bin* and the definition of *Earley-step*.

□

**lemma** *Predict-Earley*:

**assumes**  $i \leq k$   
**assumes**  $x \in \text{bin } (\text{Earley } k \mathcal{G} \omega) \ i$   
**assumes**  $\text{next-symbol } x = \text{Some } N$   
**assumes**  $(N, \alpha) \in \text{set } (\mathfrak{R} \mathcal{G})$   
**shows**  $\text{init-item } (N, \alpha) \ i \in \text{Earley } k \mathcal{G} \omega$

**lemma** *Complete-Earley*:

**assumes**  $i \leq j$   
**assumes**  $j \leq k$   
**assumes**  $x \in \text{bin } (\text{Earley } k \mathcal{G} \omega) \ i$   
**assumes**  $\text{next-symbol } x = \text{Some } N$   
**assumes**  $(N, \alpha) \in \text{set } (\mathfrak{R} \mathcal{G})$   
**assumes**  $y \in \text{bin } (\text{Earley } k \mathcal{G} \omega) \ j$   
**assumes**  $\text{item-rule } y = (N, \alpha)$   
**assumes**  $i = \text{item-origin } y$   
**assumes**  $\text{is-complete } y$   
**shows**  $\text{inc-item } x \ j \in \text{Earley } k \mathcal{G} \omega$

The proof of lemmas *Predict-Earley* and *Complete-Earley* are similar in structure to the proof of lemma *Scan-Earley* with the exception of the base case that is in both cases non-trivial but can be proven with the help of lemmas *Earley-step-sub-Earley-bin* and *Earley-bin-idem*, the definition of *Earley-bin* and the definitions of *Predict* and *Complete*, respectively.

Next we give some intuition about the core idea of the completeness proof. Assume there exists an item  $N \rightarrow \bullet A_0 A_1 \dots A_n$  in a *complete* (we define what exactly this means)

set of items  $I$  where  $A_i$  are either terminal or non-terminal symbols. Furthermore, assume there exist the following derivations for  $i_0 \leq i_1 \leq \dots \leq i_n \leq i_{n+1}$ :

$$\begin{aligned} \mathcal{G} \vdash A_0 &\Rightarrow^* \omega[i_0..i_1] \\ \mathcal{G} \vdash A_1 &\Rightarrow^* \omega[i_1..i_2] \\ &\dots \\ \mathcal{G} \vdash A_n &\Rightarrow^* \omega[i_n..i_{n+1}] \end{aligned}$$

We have one derivation to move the bullet over each terminal or non-terminal  $A_i$  and consequently the item  $N \rightarrow A_0 A_1 \dots A_n \bullet$  should be in  $I$  if  $I$  is a *complete* set of items.

We formalize this idea as follows: a set  $I$  is *partially-completed* if for each non-complete item  $x$  in  $I$ , the existence of a derivation  $D$  from the next symbol of  $x$  implies, that the item that can be obtained by moving the bullet over the next symbol of  $x$ , is also present in  $I$ . The full definition of *partially-completed* below is slightly more involved since we need to keep track of the validity of the indices. Note that the definition also requires that an arbitrary predicate  $P$  holds for the derivation  $D$ . This predicate is necessary since the completeness proof requires a proof on the length of the derivation  $D$ , and thus we sometimes need to limit the *partially-completed* property to derivations that don't exceed a certain length.

Lemma *partially-completed-upto* then formalizes the core idea: if the item  $N \rightarrow \alpha \bullet \beta, i, j$  exists in a set of items  $I$  and there exists a derivation  $\beta \xRightarrow{D} \omega[j..k]$ , then  $I$  also contains the complete item  $N \rightarrow \alpha \beta \bullet, i, k$ . Note that this holds only if  $j \leq k$ ,  $k \leq |\omega|$ , all items of  $I$  are well-formed and most importantly  $I$  must be *partially-completed* up to the length of the derivation  $D$ .

**definition** *partially-completed* ::  $\text{nat} \Rightarrow 'a \text{ cfg} \Rightarrow 'a \text{ sentential} \Rightarrow 'a \text{ items} \Rightarrow ('a \text{ derivation} \Rightarrow \text{bool}) \Rightarrow \text{bool}$  **where**

$$\begin{aligned} \text{partially-completed } k \mathcal{G} \omega I P \equiv & \\ \forall i j x a D. & \\ i \leq j \wedge j \leq k \wedge k \leq |\omega| \wedge & \\ x \in \text{bin } I \ i \wedge & \\ \text{next-symbol } x = \text{Some } a \wedge & \\ \mathcal{G} \vdash [a] \Rightarrow^D \omega[i..j] \wedge P D \longrightarrow & \\ \text{inc-item } x \ j \in I & \end{aligned}$$

To proof lemma *partially-completed-upto*, we need two auxiliary lemmas: The first one is about splitting derivations (lemma *Derivation-append-split*): a derivation  $\alpha \beta \xRightarrow{D} \gamma$ , can be split into two derivations  $E$  and  $F$  whose length is bounded by the length of  $D$ , and there exist  $\alpha'$  and  $\beta'$  such that  $\alpha \xRightarrow{E} \alpha'$ ,  $\beta \xRightarrow{F} \beta'$  and  $\gamma = \alpha' @ \beta'$ . The proof is by induction on  $D$  for arbitrary  $\alpha$  and  $\beta$  and quite technical since we need to manipulate the exact indices where each rewriting rule is applied in  $\alpha$  and  $\beta$ , and thus we omit it.

The second one is a, in spirit similar, lemma about splitting slices (lemma *slice-append-split*). The proof is straightforward by induction on the computation of the *slice* function, we also omit it, and move on to the proof of lemmas *partially-completed-upto* and *partially-completed-Earley*.

**lemma** *Derivation-append-split*:

**assumes**  $\mathcal{G} \vdash (\alpha @ \beta) \Rightarrow^D \gamma$   
**shows**  $\exists E F \alpha' \beta'. \mathcal{G} \vdash \alpha \Rightarrow^E \alpha' \wedge \mathcal{G} \vdash \beta \Rightarrow^F \beta' \wedge \gamma = \alpha' @ \beta' \wedge |E| \leq |D| \wedge |F| \leq |D|$

**lemma** *slice-append-split*:

**assumes**  $i \leq k$   
**assumes**  $xs[i..k] = ys @ zs$   
**shows**  $\exists j. ys = xs[i..j] \wedge zs = xs[j..k] \wedge i \leq b \wedge b \leq k$

**lemma** *partially-completed-upto*:

**assumes** *wf-items*  $\mathcal{G} \ \omega \ I$   
**assumes**  $j \leq k$   
**assumes**  $k \leq |\omega|$   
**assumes**  $x = \text{Item } (N, \alpha) \ b \ i \ j$   
**assumes**  $x \in I$   
**assumes**  $\mathcal{G} \vdash (\text{item-}\beta \ x) \Rightarrow^D \omega[j..k]$   
**assumes** *partially-completed*  $k \ \mathcal{G} \ \omega \ I \ (\lambda D'. |D'| \leq |D|)$   
**shows**  $\text{Item } (N, \alpha) \ |\alpha| \ i \ k \in I$

*Proof.* The proof is by induction on  $(\text{item-}\beta \ x)$  for arbitrary  $b, i, j, k, N, \alpha, x$ , and  $D$ :

For the base case we have  $\text{item-}\beta \ x = []$  and need to show that  $\text{Item } (N, \alpha) \ |\alpha| \ i \ k \in I$ :

The bullet of  $x$  is right before  $\text{item-}\beta \ x$ , or  $\text{item-}\alpha \ x = \alpha$ . Thus, the value of the bullet must be equal to the length of  $\alpha$ , which implies  $x = \text{Item } (N, \alpha) \ |\alpha| \ i \ j$ , since  $x$  is a well-formed item and  $\text{item-}\beta \ x = []$ .

We also know that  $j = k$ : we have  $\mathcal{G} \vdash \text{item-}\beta \ x \Rightarrow^D \omega[j..k]$  and  $\text{item-}\beta \ x = []$  which in turn implies that  $\omega[j..k] = []$ , and thus  $j = k$  as trivial fact about the function *slice* follows.

Hence, the statement follows from the assumption  $x \in I$  and the fact that  $x = \text{Item } (N, \alpha) \ |\alpha| \ i \ j$ .

For the induction step we need to show that  $\text{Item } (N, \alpha) \ |\alpha| \ i \ k \in I$  using assumptions:

$$\begin{aligned}
 a \# as &= item\text{-}\beta \ x & (1) & \quad wf\text{-}items \ \mathcal{G} \ \omega \ I & (2) \\
 j &\leq k & (3) & \quad k \leq |\omega| & (4) \\
 x &= Item \ (N, \alpha) \ b \ i \ j & (5) & \quad x \in I & (6) \\
 \mathcal{G} \vdash item\text{-}\beta \ x &\Rightarrow^D \omega[j..k] & (7) & \\
 partially\text{-}completed \ k \ \mathcal{G} \ \omega \ I &(\lambda D'. |D'| \leq |D|) & (8) &
 \end{aligned}$$

Using assumptions (1), (3), and (7) there exists an index  $j'$  and derivations  $E$  and  $F$  by lemmas *Derivation-append-split* and *slice-append-split* such that

$$\mathcal{G} \vdash [a] \Rightarrow^E \omega[j..j'] \quad (9) \quad |E| \leq |D| \quad (10)$$

$$\mathcal{G} \vdash as \Rightarrow^F \omega[j'..k] \quad (11) \quad |F| \leq |D| \quad (12)$$

$$j \leq j' \quad (13) \quad j' \leq k \quad (14)$$

We have *next-symbol*  $x = Some \ a$  due to assumption (1), consequently we have *inc-item*  $x \ j' \in I$  using additionally the facts about derivation  $E$  (9-10), the bounds on  $j'$  (13-14) and the assumptions (4-7) by the definition of *partially-completed*. Note that *inc-item*  $x \ j' = Item \ (N, \alpha) \ (b + 1) \ i \ j'$ , which we will from now on refer to as item  $x'$ .

From assumption (8) and fact (12) follows *partially-completed*  $k \ \mathcal{G} \ \omega \ I \ (\lambda D'. |D'| \leq |F|)$ . We also have  $as = item\text{-}\beta \ x'$  and  $x' \in I$  by the definition of  $x'$  and  $x$  and the assumptions (1,5,6). Hence, we can apply the induction hypothesis for  $x'$  using additionally the assumptions (2,4), and the facts about derivation  $F$  (11-12) from above, and have *Item*  $(N, \alpha) \ |\alpha| \ i \ k \in I$ , what we intended to show.  $\square$

**lemma** *partially-completed-Earley*:

**assumes**  $wf\text{-}\mathcal{G} \ \mathcal{G}$

**shows** *partially-completed*  $k \ \mathcal{G} \ \omega \ (Earley \ k \ \mathcal{G} \ \omega) \ (\lambda\text{-}. \ True)$

*Proof.* Let  $x, i, a, D$ , and  $j$  be arbitrary but fixed.

By definition of *partially-completed* we need to show *inc-item*  $x \ j \in Earley \ k \ \mathcal{G} \ \omega$  and can assume

$$i \leq j \quad (1) \quad j \leq k \quad (2)$$

$$k \leq |\omega| \quad (3) \quad x \in bin \ (Earley \ k \ \mathcal{G} \ \omega) \ i \quad (4)$$

$$next\text{-}symbol \ x = Some \ a \quad (5) \quad \mathcal{G} \vdash [a] \Rightarrow^D \omega[i..j] \quad (6)$$

We proof this by complete induction on  $|D|$  for arbitrary  $x, i, a, j$ , and  $D$ , and split the proof into two different cases:

- $D = []$ : Since  $\mathcal{G} \vdash [a] \Rightarrow^D \omega[i..j]$ , we have  $[a] = \omega[i..j]$ , and consequently  $\omega ! i = a$  and  $j = i + 1$ . Now we discharge the assumptions of lemma *Scan-Earley*, by assumptions (4,5) and the fact  $j \leq |\omega|$  (that follows from assumptions (2,3)), and have *inc-item*  $x (i + 1) \in \text{Earley } k \mathcal{G} \omega$  which finishes the proof since  $j = i + 1$ .
- $D = d \# \mathcal{D}$ : Due to assumption  $\mathcal{G} \vdash [a] \Rightarrow^D \omega[i..j]$ , there exists an  $\alpha$  such that *Derives1*  $\mathcal{G} [a] (fst d) (snd d) \alpha$  and  $\mathcal{G} \vdash \alpha \Rightarrow^D \omega[i..j]$  by the definition of *Derivation*. From the definition of *Derives1* we see that there exists a non-terminal  $N$  such that  $a = N$ ,  $(N, \alpha) \in \text{set } (\mathfrak{R} \mathcal{G})$ ,  $fst d = 0$ , and  $snd d = (N, \alpha)$ .

Let  $y$  denote *Item*  $(N, \alpha) 0 i i$ . Since we have  $i \leq k$  (assumptions (1,2)), and assumptions (4,5), and we showed that  $a = N$  and  $(N, \alpha) \in \text{set } (\mathfrak{R} \mathcal{G})$ , and  $y$  is an initial item, we have  $y \in \text{Earley } k \mathcal{G} \omega$  by lemma *Predict-Earley*.

Next, we use lemma *partially-completed-upto* to show that the completed version of item  $y$  is also present in the  $j$ -th bin of  $\text{Earley } k \mathcal{G} \omega$  since we have a derivation  $\mathcal{G} \vdash \alpha \Rightarrow^D \omega[i..j]$ , or *Item*  $(N, \alpha) |\alpha| i j \in \text{bin } (\text{Earley } k \mathcal{G} \omega) j$ : we use assumptions (1-3); have proven  $y \in \text{Earley } k \mathcal{G} \omega$ ; and have *wf-items*  $\mathcal{G} \omega (\text{Earley } k \mathcal{G} \omega)$  by lemma *wf-Earley*. Additionally, we know  $\mathcal{G} \vdash \text{item-}\beta y \Rightarrow^D \omega[i..j]$  since  $\mathcal{G} \vdash [a] \Rightarrow^D \omega[i..j]$  and  $a = N$ , by the definition of item  $y$ . Finally, we use the induction hypothesis to show *partially-completed*  $k \mathcal{G} \omega (\text{Earley } k \mathcal{G} \omega) (\lambda E. |E| \leq |\mathcal{D}|)$ , since  $|\mathcal{D}| \leq |D|$  by definition of *partially-completed*, using once again all of our assumptions. This in turn implies *partially-completed*  $j \mathcal{G} \omega (\text{Earley } k \mathcal{G} \omega) (\lambda E. |E| \leq |\mathcal{D}|)$  since  $j \leq k$  by definition of *partially-completed*. Now we can use lemma *partially-completed-upto*, and the statement follows from the definition of a bin.

Finally, we prove *inc-item*  $x j \in \text{Earley } k \mathcal{G} \omega$  by lemma *Complete-Earley*: Once again we use assumptions (1,2,4), we also know that *next-symbol*  $x = \text{Some } N$ , due to assumption (5) and the fact  $a = N$ . Moreover, we have  $(N, \alpha) \in \text{set } (\mathfrak{R} \mathcal{G})$  and most importantly *Item*  $(N, \alpha) |\alpha| i j \in \text{bin } (\text{Earley } k \mathcal{G} \omega) j$ , which concludes this proof.

□

Lemma *partially-completed-Earley* follows trivially from *partially-completed-Earley* by definition of *Earley*.

**lemma** *partially-completed-Earley*:

**assumes** *wf-G*  $\mathcal{G}$

**shows** *partially-completed*  $|\omega| \mathcal{G} \omega (\text{Earley } \mathcal{G} \omega) (\lambda -. \text{True})$

And finally, we can proof completeness of Earley's algorithm, obtaining corollary *correctness-Earley* due to lemma *soundness*.



**theorem** *completeness:*

**assumes** *wf- $\mathcal{G}$*   $\mathcal{G}$   
**assumes** *is-sentence*  $\mathcal{G}$   $\omega$   
**assumes**  $\mathcal{G} \vdash [\mathfrak{S} \mathcal{G}] \Rightarrow^* \omega$   
**shows** *recognizing* ( $\mathcal{E}arley \mathcal{G} \omega$ )  $\mathcal{G} \omega$

*Proof.* We know that there exists an  $\alpha$  and a derivation  $D$  such that  $(\mathfrak{S} \mathcal{G}, \alpha) \in \text{set } (\mathfrak{R} \mathcal{G})$  and  $\mathcal{G} \vdash \alpha \Rightarrow^D \omega$ , since  $\mathcal{G} \vdash [\mathfrak{S} \mathcal{G}] \Rightarrow^* \omega$ . Let  $x$  denote the item  $\text{Item } (\mathfrak{S} \mathcal{G}, \alpha) \ 0 \ 0 \ 0$ . By definition of  $x$  and the *Init* operation and  $\mathcal{E}arley$  function, and the fact that  $\text{Init } \mathcal{G} \subseteq \mathcal{E}arley \ k \ \mathcal{G} \ \omega$ , we have  $x \in \mathcal{E}arley \ \mathcal{G} \ \omega$ , moreover we have *partially-completed*  $|\omega| \ \mathcal{G} \ \omega$  ( $\mathcal{E}arley \ \mathcal{G} \ \omega$ ) ( $\lambda\cdot$ . *True*) using lemma *partially-completed- $\mathcal{E}arley$*  and assumption *wf- $\mathcal{G}$*   $\mathcal{G}$ , and thus have  $\text{Item } (\mathfrak{S} \mathcal{G}, \alpha) \ |\alpha| \ 0 \ |\omega| \in \mathcal{E}arley \ \mathcal{G} \ \omega$  by lemmas *partially-completed-upto* and *wf- $\mathcal{E}arley$*  and the definition of *partially-completed*. The statement *recognizing* ( $\mathcal{E}arley \ \mathcal{G} \ \omega$ )  $\mathcal{G} \ \omega$  follows immediately by the definition of *recognizing*, *is-finished*, and *is-complete*.  $\square$

**corollary** *correctness- $\mathcal{E}arley$ :*

**assumes** *wf- $\mathcal{G}$*   $\mathcal{G}$   
**assumes** *is-sentence*  $\mathcal{G}$   $\omega$   
**shows** *recognizing* ( $\mathcal{E}arley \ \mathcal{G} \ \omega$ )  $\mathcal{G} \ \omega \longleftrightarrow \mathcal{G} \vdash [\mathfrak{S} \mathcal{G}] \Rightarrow^* \omega$

### 3.6 Finiteness

At last, we prove that the set of Earley items is finite. In Chapter 4 we are using this result to prove the termination of an executable version of the algorithm.

Since  $\mathcal{E}arley \ \mathcal{G} \ \omega$  only generates well-formed items (lemma *wf- $\mathcal{E}arley$* ) it suffices to prove that there only exists a finite number of well-formed items. Define

$$T = \text{set } (\mathfrak{R} \mathcal{G}) \times \{0..m\} \times \{0..|\omega|\} \times \{0..|\omega|\}$$

where  $m = \text{Max } \{|rule-body \ r| \mid r \in \text{set } (\mathfrak{R} \mathcal{G})\}$ . The set  $T$  is finite since there exists only a finite number of production rules and  $\{x \mid \text{wf-item } \mathcal{G} \ \omega \ x\}$  is a subset of mapping the *Item* constructor over  $T$  (strictly speaking we need to first unpack the quadruple).

**lemma** *finiteness-UNIV-wf-item:*

**shows** *finite*  $\{x \mid x. \text{wf-item } \mathcal{G} \ \omega \ x\}$

**theorem** *finiteness:*

**shows** *finite* ( $\mathcal{E}arley \ \mathcal{G} \ \omega$ )

## 4 Earley Recognizer Implementation

### 4.1 The Executable Algorithm

In Chapter 3 we proved correctness of a set-based, non-executable version of Earley's simplified recognizer algorithm. In this chapter we implement an executable algorithm. But instead of re-proving soundness and completeness for the executable algorithm, we follow the approach of Jones [Jones:1972]. We refine our set-based approach from Chapter 3 to a *functional* list-based implementation and prove subsumption in both directions, or each item generated by the list-based approach is also generated by the set-based approach which implies soundness of the executable algorithm, and vice versa which implies in turn completeness. We extend the algorithm of Chapter 3 in a second orthogonal way by already adding the necessary information to construct parse trees. We only introduce and explain the needed data structures but refrain from presenting any proofs in this chapter since constructing parse trees is the primary subject of Chapter 5.

First we introduce a new data representation: instead of a set of Earley items we work with the data structure *bins*: a list of static length  $(|\omega| + 1)$  containing in turn bins implemented as variable length lists of Earley *entries*. An entry consists of an Earley item and a new data type *pointer* representing conceptually an imperative pointer describing the origin of its accompanying item. Table 4.1 illustrates the bins for our running example. There are three possible reasons, corresponding to the three basic operations, for the existence of an entry with Earley item  $x$  in a specific bin  $k$ :

- It was predicted. In that case we consider it created from thin air and do not need to track any additional information, thus the pointer is *Null*. For our example, bin  $B_0$  contains the entry  $S \rightarrow \bullet x, 0, 0; \perp$  consisting of the item  $S \rightarrow \bullet x, 0, 0$  and a *Null* pointer denoted by  $\perp$ .
- It was scanned. Then there exists another Earley item  $x'$  in the previous bin  $k - 1$  from which this item was computed. Hence, we keep a predecessor pointer *Pre pre* where *pre* is a natural number indicating the index of item  $x'$  in bin  $k - 1$ . Table 4.1 contains the entry  $S \rightarrow x \bullet, 2, 3; 1$  in bin  $B_3$ , the predecessor pointer is 1 (we omit the *Pre* constructor for readability) since this item was created by the item  $S \rightarrow \bullet x, 2, 2$  of the entry at index 1 in  $B_2$ .
- It was completed. Note that an item might be completed in more than one way.

In each case the item  $x$  has a complete reduction item  $y$  in the current bin and a predecessor item  $x'$  in the origin bin of  $y$ . We track this information by at least one reduction pointer ( $PreRed$  ( $k', pre, red$ )  $reds$ ) where  $k'$ ,  $pre$ , and  $red$  are respectively the origin index of the complete item  $y$  or the bin of item  $x'$ ,  $pre$  is the index of  $x'$  in bin  $k'$ , and  $red$  is the index of  $y$  in the current bin  $k$ . The list  $reds$  contains other valid reduction triples for this item. This is illustrated by the entry  $S \rightarrow S + S \bullet, 0, 5; (4, 1, 0), (2, 0, 1)$  in bin  $B_5$  of Table 4.1. We omit the  $PreRed$  and list constructors again for readability. This entry (without the second reduction triple) was first created due to the complete item  $S \rightarrow x \bullet, 4, 5$  at index 0 in bin  $B_5$  and the predecessor item  $S \rightarrow S + \bullet S, 0, 4$  at index 1 in bin  $B_4$ , but we can also create it by the complete item  $S \rightarrow S + S \bullet, 2, 5$  at index 1 in bin  $B_5$  and the predecessor item  $S \rightarrow S + \bullet S, 0, 2$  at index 0 in bin  $B_2$ , or the two possible ways to derive the input  $\omega = (x + x) + x$  and  $\omega = x + (x + x)$ .

Additionally, we define two useful abbreviations *items* and *pointers* that map a given bin to the list of items respectively pointers it consists of.

 Table 4.1: Earley items with pointers for the grammar  $\mathcal{G}$ :  $S \rightarrow x, S \rightarrow S + S$ 

	$B_0$	$B_1$	$B_2$
0	$S \rightarrow \bullet x, 0, 0; \perp$	$S \rightarrow x \bullet, 0, 1; 0$	$S \rightarrow S + \bullet S, 0, 2; 1$
1	$S \rightarrow \bullet S + S, 0, 0; \perp$	$S \rightarrow S \bullet + S, 0, 1; (0, 1, 0)$	$S \rightarrow \bullet x, 2, 2; \perp$
2			$S \rightarrow \bullet S + S, 2, 2; \perp$
	$B_3$	$B_4$	$B_5$
0	$S \rightarrow x \bullet, 2, 3; 1$	$S \rightarrow S + \bullet S, 2, 4; 2$	$S \rightarrow x \bullet, 4, 5; 2$
1	$S \rightarrow S + S \bullet, 0, 3; (2, 0, 0)$	$S \rightarrow S + \bullet S, 0, 4; 3$	$S \rightarrow S + S \bullet, 2, 5; (4, 0, 0)$
2	$S \rightarrow S \bullet + S, 2, 3; (2, 2, 0)$	$S \rightarrow \bullet x, 4, 4; \perp$	$S \rightarrow S + S \bullet, 0, 5; (4, 1, 0), (2, 0, 1)$
3	$S \rightarrow S \bullet + S, 0, 3; (0, 1, 1)$	$S \rightarrow \bullet S + S, 4, 4; \perp$	$S \rightarrow S \bullet + S, 4, 5; (4, 3, 0)$
4			$S \rightarrow S \bullet + S, 2, 5; (2, 2, 1)$
5			$S \rightarrow S \bullet + S, 0, 5; (0, 1, 2)$

```

datatype pointer =
  Null
| Pre nat — pre
| PreRed nat × nat × nat (nat × nat × nat) list — (k', pre, red) reds

```

```

datatype 'a entry =
  Entry (item : 'a item) (pointer : pointer)

```

**type-synonym**  $'a \text{ bin} = 'a \text{ entry list}$

**type-synonym**  $'a \text{ bins} = 'a \text{ bin list}$

**definition**  $\text{items} :: 'a \text{ bin} \Rightarrow 'a \text{ item list}$  **where**  
 $\text{items } b = \text{map item } b$

**definition**  $\text{pointers} :: 'a \text{ bin} \Rightarrow \text{pointer list}$  **where**  
 $\text{pointers } b = \text{map pointer } b$

Next we implement list-based versions of the *Init*, *Scan*, *Predict*, and *Complete* operations. Function *Init-list* creates a list of  $(|\omega| + 1)$  empty lists or bins. Subsequently, it constructs an initial bin containing entries consisting of initial items for all the production rules that have the start symbol on their left-hand sides, and finally it overwrites the 0-th bin with this initial bin.

**definition**  $\text{Init-list} :: 'a \text{ cfg} \Rightarrow 'a \text{ sentential} \Rightarrow 'a \text{ bins}$  **where**  
 $\text{Init-list } \mathcal{G} \ \omega \equiv$   
 $\text{let } bs = \text{replicate } (|\omega| + 1) \ ([\ ])$  *in*  
 $\text{let } rs = \text{filter } (\lambda r. \text{rule-head } r = \mathfrak{S} \ \mathcal{G}) \ (\mathfrak{R} \ \mathcal{G})$  *in*  
 $\text{let } b0 = \text{map } (\lambda r. (\text{Entry } (\text{init-item } r \ 0) \ \text{Null})) \ rs$  *in*  $bs[0 := b0]$

Functions *Scan-list*, *Predict-list*, and *Complete-list* are defined analogously to the definitions of *Scan*, *Predict*, and *Complete* and we only highlight noteworthy differences. The set-based implementations take accumulated as arguments the index  $k$  of the current bin, the grammar  $\mathcal{G}$ , the input  $\omega$ , and the current set of Earley items  $I$ . The list-based definitions are more specific. The  $k$ -th bin is no longer only conceptual and we replace the argument  $I$  in the following ways: function *Scan-list* takes as arguments the currently considered item  $x$ , its next *terminal* symbol  $a$  (as plain value and not wrapped in an option) and the index *pre* of  $x$  in the current bin  $k$ , and sets the predecessor pointer accordingly. Function *Predict-list* only needs access to the next non-terminal symbol  $N$  of  $x$ , and returns only entries with *Null* pointers. The implementation of *Complete-list* is slightly more involved. It takes as arguments again  $x$  and the index *red* of  $x$  in the current bin  $k$  (since  $x$  is a complete reduction item this time around), but also the complete bins  $bs$ , since it needs to find all potential predecessor items as well as their indices in the origin bin of  $x$  (see *find-with-index*), and sets the reduction triples accordingly.

**definition**  $\text{Scan-list} :: \text{nat} \Rightarrow 'a \text{ sentential} \Rightarrow 'a \Rightarrow 'a \text{ item} \Rightarrow \text{nat} \Rightarrow 'a \text{ entry list}$  **where**  
 $\text{Scan-list } k \ \omega \ a \ x \ \text{pre} \equiv$   
 $\text{if } \omega!k = a \ \text{then}$   
 $\text{let } x' = \text{inc-item } x \ (k+1)$  *in*  
 $[\text{Entry } x' \ (\text{Pre } \text{pre})]$   
 $\text{else } []$

**definition**  $\text{Predict-list} :: \text{nat} \Rightarrow 'a \text{ cfg} \Rightarrow 'a \Rightarrow 'a \text{ entry list}$  **where**

*Predict-list*  $k \mathcal{G} N \equiv$   
 $\text{let } rs = \text{filter } (\lambda r. \text{rule-head } r = N) (\mathfrak{R} \mathcal{G}) \text{ in}$   
 $\text{map } (\lambda r. (\text{Entry } (\text{init-item } r \ k) \ \text{Null})) \ rs$

**fun** *filter-with-index'* ::  $\text{nat} \Rightarrow ('a \Rightarrow \text{bool}) \Rightarrow 'a \text{ list} \Rightarrow ('a \times \text{nat}) \text{ list}$  **where**  
 $\text{filter-with-index}' \ - \ - \ [] = []$   
 $| \text{filter-with-index}' \ i \ P \ (x\#xs) = ($   
 $\quad \text{if } P \ x \text{ then } (x, i) \# \text{filter-with-index}' \ (i+1) \ P \ xs$   
 $\quad \text{else } \text{filter-with-index}' \ (i+1) \ P \ xs)$

**definition** *filter-with-index* ::  $('a \Rightarrow \text{bool}) \Rightarrow 'a \text{ list} \Rightarrow ('a \times \text{nat}) \text{ list}$  **where**  
 $\text{filter-with-index} \ P \ xs = \text{filter-with-index}' \ 0 \ P \ xs$

**definition** *Complete-list* ::  $\text{nat} \Rightarrow 'a \text{ item} \Rightarrow 'a \text{ bins} \Rightarrow \text{nat} \Rightarrow 'a \text{ entry list}$  **where**  
 $\text{Complete-list} \ k \ x \ bs \ red \equiv$   
 $\text{let } orig = bs ! \text{item-origin } x \text{ in}$   
 $\text{let } is = \text{filter-with-index} \ (\lambda x'. \text{next-symbol } x' = \text{Some } (\text{item-rule-head } x)) \ (\text{items } orig) \text{ in}$   
 $\text{map } (\lambda (x', pre). (\text{Entry } (\text{inc-item } x' \ k) \ (\text{PreRed } (\text{item-origin } x, pre, red) \ []))) \ is$

In our data representation a bin is just a simple list but it implements a set. Hence, we need to make sure that updating a bin (*bin-upd*) or inserting an additional entry into a bin maintains its set properties. Additionally, since it is possible to generate multiple reduction pointers for the same item, we have to take care to update the pointer information accordingly, in particular merge reduction triples, if the item of the entry to be inserted matches the item of an already present entry. Function *bin-upds* inserts multiple entries into a specific bin. Finally, function *bins-upd* updates the  $k$ -th bin by inserting the given list of entries using function *bin-upds*. Note that an alternative but equivalent implementation of *bin-upds* is *fold bin-upd es b*. We primarily choose the explicit definition since it simplified some of the proofs, but overall the choice is stylistic in nature.

**fun** *bin-upd* ::  $'a \text{ entry} \Rightarrow 'a \text{ bin} \Rightarrow 'a \text{ bin}$  **where**  
 $\text{bin-upd} \ e' \ [] = [e']$   
 $| \text{bin-upd} \ e' \ (e\#es) = ($   
 $\quad \text{case } (e', e) \text{ of}$   
 $\quad (\text{Entry } x \ (\text{PreRed } px \ xs), \text{Entry } y \ (\text{PreRed } py \ ys)) \Rightarrow$   
 $\quad \quad \text{if } x = y \text{ then } \text{Entry } x \ (\text{PreRed } py \ (px\#xs@ys)) \# \ es$   
 $\quad \quad \text{else } e \# \text{bin-upd } e' \ es$   
 $\quad | \ - \Rightarrow$   
 $\quad \quad \text{if item } e' = \text{item } e \text{ then } e \# \ es$   
 $\quad \quad \text{else } e \# \text{bin-upd } e' \ es)$

**fun** *bin-upds* ::  $'a \text{ entry list} \Rightarrow 'a \text{ bin} \Rightarrow 'a \text{ bin}$  **where**  
 $\text{bin-upds} \ [] \ b = b$

|  $\text{bin-upds } (e\#es) \ b = \text{bin-upds } es \ (\text{bin-upd } e \ b)$

**definition**  $\text{bins-upd} :: 'a \text{ bins} \Rightarrow \text{nat} \Rightarrow 'a \text{ entry list} \Rightarrow 'a \text{ bins}$  **where**  
 $\text{bins-upd } bs \ k \ es = bs[k := \text{bin-upds } es \ (bs!k)]$

The central piece for the list-based implementation is the function *Earley-bin-list'*. A function call of the form  $\text{Earley-bin-list}' \ k \ \mathcal{G} \ \omega \ bs \ i$  completes the  $k$ -th bin starting from index  $i$ . For the current item  $x$  under consideration the function first computes the possible new entries depending on the next symbol of  $x$  which can either be some terminal symbol - we scan -, or non-terminal symbol - we predict -, or *None* - we complete. It then updates the bins  $bs$  appropriately using the function  $\text{bins-upd}$ . We have to define the function as a *partial-function*, since it might never terminate if it keeps appending newly generated items to the  $k$ -th bin it currently operates on. We prove termination and highlight the Isabelle specific details in Section 4.4. The function *Earley-bin-list* then fully completes the  $k$ -th bin, or starts its computation at index 0, and thus corresponds in functionality to the function *Earley-bin* of Chapter 3.

**partial-function** (*tailrec*)  $\text{Earley-bin-list}' :: \text{nat} \Rightarrow 'a \text{ cfg} \Rightarrow 'a \text{ sentential} \Rightarrow 'a \text{ bins} \Rightarrow \text{nat} \Rightarrow 'a \text{ bins}$  **where**

$\text{Earley-bin-list}' \ k \ \mathcal{G} \ \omega \ bs \ i =$   
 if  $i \geq |\text{items } (bs!k)|$  then  $bs$   
 else  
 let  $x = \text{items } (bs!k) ! i$  in  
 let  $bs' =$   
 case  $\text{next-symbol } x$  of  
 Some  $a \Rightarrow$   
 if  $\text{is-terminal } \mathcal{G} \ a$  then  
 if  $k < |\omega|$  then  $\text{bins-upd } bs \ (k+1) \ (\text{Scan-list } k \ \omega \ a \ x \ i)$   
 else  $bs$   
 else  $\text{bins-upd } bs \ k \ (\text{Predict-list } k \ \mathcal{G} \ a)$   
 |  $\text{None} \Rightarrow \text{bins-upd } bs \ k \ (\text{Complete-list } k \ x \ bs \ i)$   
 in  $\text{Earley-bin-list}' \ k \ \mathcal{G} \ \omega \ bs' \ (i+1)$

**definition**  $\text{Earley-bin-list} :: \text{nat} \Rightarrow 'a \text{ cfg} \Rightarrow 'a \text{ sentential} \Rightarrow 'a \text{ bins} \Rightarrow 'a \text{ bins}$  **where**  
 $\text{Earley-bin-list } k \ \mathcal{G} \ \omega \ bs = \text{Earley-bin-list}' \ k \ \mathcal{G} \ \omega \ bs \ 0$

Finally, functions *Earley-list* and *Earley-list* are structurally identical to functions *Earley* respectively *Earley* of Chapter 3, differing only in the type of the used operations and the return type: bins or lists of items instead of set of items.

```

fun Earley-list :: nat  $\Rightarrow$  'a cfg  $\Rightarrow$  'a sentential  $\Rightarrow$  'a bins where
  Earley-list 0  $\mathcal{G}$   $\omega$  = Earley-bin-list 0  $\mathcal{G}$   $\omega$  (Init-list  $\mathcal{G}$   $\omega$ )
| Earley-list (Suc n)  $\mathcal{G}$   $\omega$  = Earley-bin-list (Suc n)  $\mathcal{G}$   $\omega$  (Earley-list n  $\mathcal{G}$   $\omega$ )

```

```

definition  $\mathcal{E}$ arley-list :: 'a cfg  $\Rightarrow$  'a sentential  $\Rightarrow$  'a bins where
   $\mathcal{E}$ arley-list  $\mathcal{G}$   $\omega$  = Earley-list  $|\omega|$   $\mathcal{G}$   $\omega$ 

```

## 4.2 A Word on Performance

Earley [Earley:1970] implements his recognizer algorithm in the imperative programming paradigm and provides an informal argument for the running time  $\mathcal{O}(n^3)$  where  $n = |\omega|$ . In contrast, our implementation is purely functional, and one might expect a quite significant decrease in performance. In this section we provide an informal argument showing that, although we cannot quite achieve the time complexity of an imperative implementation, we are 'only' one order of magnitude slower or the running time of our implementation is  $\mathcal{O}(n^4)$ . Then we summarize Earley's imperative implementation approach and the additional steps that are needed to achieve the desired running time. Additionally, we sketch a slightly different and more complicated functional implementation that achieves a theoretical running time of  $\mathcal{O}(n^3 \log n)$ , and highlight possible further performance improvements. Finally, we discuss why we choose our particular implementation.

We state the running time of our implementation of the algorithm in terms of the length  $n$  of the input  $\omega$ , and provide an informal argument that its running time is  $\mathcal{O}(n^4)$ . Each bin  $B_j$  ( $0 \leq j \leq n$ ) contains only items of the form *Item*  $r$   $b$   $i$   $j$ . The number of possible production rules  $r$ , and possible bullet positions  $b$  are both independent of  $n$  and can thus be considered (possible large) constants. The origin  $i$  is bounded by  $0 \leq i \leq j$  and thus depends on  $j$  which is in turn dependent on  $n$ . Thus, the number of items in each bin  $B_j$  is overall  $\mathcal{O}(n)$ .

We also have  $\text{Init-list} \in \mathcal{O}(n)$  since the function *replicate* takes time linear in the length of  $\omega$ , and functions *filter* and *map* operate at most on the size of the grammar  $\mathcal{G}$  or constant time. We also know  $\text{Scan-list} \in \mathcal{O}(n)$ . The dominating term is surprisingly  $\omega ! k$ , since  $0 \leq k \leq n$ , and it computes at most one new entry. Function *Predict-list* takes time in the the size of the grammar  $\mathcal{G}$ , due to the *filter* and *map* functions, or constant time, and computes at most  $|\mathcal{G}|$  new items. Function *Complete-list* again takes linear time, since finding the origin bin of the given item  $x$  takes linear time, and functions *items*, *filter-with-index*, and *map* operate on the origin bin which is - in the worst case - of linear size as argued in the previous paragraph. Consequently, the function also computes at most  $\mathcal{O}(n)$  new items.

Updating a bin (*bin-upd*) with a single entry takes at most linear time, inserting  $e$  new

entries (*bin-upds*) thus takes time  $e \cdot \mathcal{O}(n)$ , and hence function *bins-upd* also runs in time  $e \cdot \mathcal{O}(n)$ . The analysis of function *Earley-bin-list'* is slightly more involved. It computes the contents of a bin  $B_j$ , or it calls itself recursively at most  $n$  times, since the number of items in any bin is  $\mathcal{O}(n)$ . The time for one function execution is dominated by the time it takes to update the bins with the newly created items whose number in turn depends on the operation we applied but is bounded in the worst case by  $n$  during the *Complete-list* operation. All the other operations of the function body run in at most linear time. Overall we have for the body of *Earley-bin-list'*:  $\mathcal{O}(n) + e \cdot \mathcal{O}(n) = \mathcal{O}(n^2)$ . And thus *Earley-bin-list'*  $\in \mathcal{O}(n^3)$ . The same bound holds trivially for *Earley-bin-list*. Since functions *Earley-list* or *Earley-list* call *Earley-bin-list* once for each bin  $B_j$  and  $0 \leq j \leq n$ , the overall running time is  $\mathcal{O}(n^4)$ .

One might be tempted to think that the decrease in performance compared to an imperative implementation is due to the fact that we are representing bins as functional lists and appending to and indexing into bins which takes linear time and not constant time. This is not the case. Earley implements the algorithm as follows. On the top-level bins are no longer a list but an array. Each bin is a singly-linked list, and pointers are no longer represented by the type *pointer* but by actual pointers between entries. The worst case running time of the algorithm is still  $\mathcal{O}(n^4)$ . The algorithm still iterates over  $n$  bins, traverses in the worst case  $\mathcal{O}(n)$  items in each bin and for each item, the worst case operation, completion, still generates  $\mathcal{O}(n)$  new items that all have to be inserted into the current bin which takes linear time for *each* new item. To achieve the running time of  $\mathcal{O}(n^3)$  we need to find a way to add a new item into a bin in constant time. In an imperative setting one obvious way is to not only keep a singly-linked list of items and pointers but additionally a map. The keys are the items of the list and the map stores as value for a specific item a pointer to itself or its position in the list. Insertion of a new item into a bin then works as follows: if the item is already present in the map, we follow the pointer to the item and update the parse tree pointers of the item in the list accordingly depending on the kind of item. Otherwise we just append the item and its corresponding parse tree pointers to the list and insert the item and a pointer to its position in the linked list into the map.

Sadly, this approach does not work in a functional setting. Appending an item to a list takes linear and not constant time. But even if we prepend the new item onto the list there is another problem. We cannot simply store pointers in the map that we can chase in constant time to the location of the item in the list, but still have to store the index of the corresponding item. And consequently updating the pointer information takes again linear time due to the indexing. One possible solution is to change one's point of view. In the imperative approach the list serves two purposes: it represents the bin and is at the same time a worklist for the algorithm. The map only optimizes performance. We can obtain a  $\mathcal{O}(n^3 \log n)$  functional implementation if we consider the



list only a worklist and the map (or its keys) the bin. We also need to adapt the pointer datatype. Instead of wrapping indices representing predecessor or reduction items in the list, a pointer should contain the actual items. E.g. a pointer is either *Null*, or *Pre*  $x'$ , or *PreRed*  $(x', y)$  *xy*s where  $x'$  is respectively the predecessor item and  $y$  is the complete reduction item. Overall the running time for inserting a new item into a bin consists of prepending the item onto the worklist, or constant time, and inserting the item into the map which can be done in logarithmic time. Thus, the overall running time of this approach is  $\mathcal{O}(n^3 \log n)$ .

Since we are already talking about performance, we highlight some of the more common performance improvements. We can predict faster if we organize the grammar in a more efficient manner. Currently, the *Predict* operation needs to pass through the whole grammar to find the alternatives for a specific non-terminal. The first performance improvement is to group the production rules by their left-hand side non-terminals. We can also complete more efficiently. The *Complete* operation scans through the origin bin of an complete item, searching for items where the next symbol matches the rule head of the production rule of the complete item. We can optimize this search by keeping an additional map from 'next symbol' non-terminals to their corresponding items for each bin. Finally, as mentioned earlier, we omit implementing a lookahead terminal. Note that, although these performance improvements might speed up the algorithm quite considerably, particularly the lookahead terminal, none of them improve the worst case running time.

We decided against implementing the map-based functional approach with a running time of  $\mathcal{O}(n^3 \log n)$  and 'settle' for the current approach with a running time of  $\mathcal{O}(n^4)$  due to two reasons. The map-based functional approach is more complicated and the improvement of the running time, although significant, still does not reach the optimum. If we optimize our approach only to achieve better performance, we would like to achieve optimal performance, at least asymptotically. The current approach, appending items to the list and using natural numbers as pointers, maps more easily to the imperative approach. And our original intention was to refine the algorithm once more to an imperative version. This exceeded the scope of this thesis but is worthwhile future work.

### 4.3 Sets or Bins as Lists

In this section we prove that the list representation of bins, in particular updating a bin or bins with the functions *bin-upd*, *bin-upds*, and *bins-upd*, fulfills the required set semantics. We define a function *bins* that accumulates all bins into one set of Earley items. Note that a call of the form *Earley-bin-list'*  $k \mathcal{G} \omega bs i$  iterates through the entries of the  $k$ -th bin or the current worklist in ascending order starting at index  $i$ . All items at

indices  $i \leq j$  are untouched and thus should already have been processed accordingly. We make two further definitions capturing the set of items which should already be 'done'. The term  $\text{bin-upto } b \ i$  represents the items of a bin  $b$  up to but not including the  $i$ -th index. Similarly, function  $\text{bins-upto}$  computes the set of items consisting of the  $k$ -th bin up to but not including the  $i$ -th index and the items of all previous bins.

**definition**  $\text{bins} :: 'a \text{ bins} \Rightarrow 'a \text{ items where}$   
 $\text{bins } bs = \bigcup \{ \text{set } (\text{items } (bs!k)) \mid k. k < |bs| \}$

**definition**  $\text{bin-upto} :: 'a \text{ bin} \Rightarrow \text{nat} \Rightarrow 'a \text{ items where}$   
 $\text{bin-upto } b \ i = \{ \text{items } b \ ! \ j \mid j. j < i \wedge j < |\text{items } b| \}$

**definition**  $\text{bins-upto} :: 'a \text{ bins} \Rightarrow \text{nat} \Rightarrow \text{nat} \Rightarrow 'a \text{ items where}$   
 $\text{bins-upto } bs \ k \ i = \bigcup \{ \text{set } (\text{items } (bs!l)) \mid l. l < k \} \cup \text{bin-upto } (bs!k) \ i$

The next six lemmas then proof the set semantics of updating one bin with one item ( $\text{bin-upd}$ ), multiple items ( $\text{bin-upds}$ ), or updating a particular bin with multiple items ( $\text{bins-upd}$ ). The proofs are straightforward and respectively by induction on the bin  $b$  for an arbitrary item  $e$ , by induction on the items  $es$  to be inserted for an arbitrary bin  $b$ , or by definition of  $\text{bin-upds}$  and  $\text{bins}$ , each time using previously proven lemmas in the appropriate proofs.

**lemma**  $\text{set-items-bin-upd}$ :  
 $\text{set } (\text{items } (\text{bin-upd } e \ b)) = \text{set } (\text{items } b) \cup \{ \text{item } e \}$

**lemma**  $\text{distinct-bin-upd}$ :  
**assumes**  $\text{distinct } (\text{items } b)$   
**shows**  $\text{distinct } (\text{items } (\text{bin-upd } e \ b))$

**lemma**  $\text{set-items-bin-upds}$ :  
 $\text{set } (\text{items } (\text{bin-upds } es \ b)) = \text{set } (\text{items } b) \cup \text{set } (\text{items } es)$

**lemma**  $\text{distinct-bin-upds}$ :  
**assumes**  $\text{distinct } (\text{items } b)$   
**shows**  $\text{distinct } (\text{items } (\text{bin-upds } es \ b))$

**lemma**  $\text{bins-bins-upd}$ :  
**assumes**  $k < |bs|$   
**shows**  $\text{bins } (\text{bins-upd } bs \ k \ es) = \text{bins } bs \cup \text{set } (\text{items } es)$

**lemma**  $\text{distinct-bins-upd}$ :  
**assumes**  $\text{distinct } (\text{items } (bs!k))$   
**shows**  $\text{distinct } (\text{items } (\text{bins-upd } bs \ k \ es \ ! \ k))$

In our formalization we prove further basic lemmas about functions  $\text{bin-upd}$ ,  $\text{bin-upds}$ , and  $\text{bins-upd}$ . In particular how updating bins changes the length of a bin, interacts with

indexing into a bin or does not change the ordering of the items in a bin. Furthermore, we prove similar lemmas about functions *bin-upto* and *bins-upto* and their interplay with *bin(s)* updates. We omit them for brevity.

## 4.4 Well-formedness and Termination

We also need to refine the notion of well-formed items to well-formed *bin* items. An item is a well-formed bin item for the *k*-th bin if it is a well-formed item and its end index coincides with *k*. We call a bin well-formed if it only contains well-formed bin items and its items are distinct, and lift this notion of well-formedness to the toplevel list of bins.

**definition** *wf-bin-item* :: 'a cfg  $\Rightarrow$  'a sentential  $\Rightarrow$  nat  $\Rightarrow$  'a item  $\Rightarrow$  bool **where**  
*wf-bin-item*  $\mathcal{G} \ \omega \ k \ x \equiv \text{wf-item } \mathcal{G} \ \omega \ x \wedge \text{item-end } x = k$

**definition** *wf-bin-items* :: 'a cfg  $\Rightarrow$  'a sentential  $\Rightarrow$  nat  $\Rightarrow$  'a item list  $\Rightarrow$  bool **where**  
*wf-bin-items*  $\mathcal{G} \ \omega \ k \ xs \equiv \forall x \in \text{set } xs. \text{wf-bin-item } \mathcal{G} \ \omega \ k \ x$

**definition** *wf-bin* :: 'a cfg  $\Rightarrow$  'a sentential  $\Rightarrow$  nat  $\Rightarrow$  'a bin  $\Rightarrow$  bool **where**  
*wf-bin*  $\mathcal{G} \ \omega \ k \ b \equiv \text{distinct } (\text{items } b) \wedge \text{wf-bin-items } \mathcal{G} \ \omega \ k \ (\text{items } b)$

**definition** *wf-bins* :: 'a cfg  $\Rightarrow$  'a list  $\Rightarrow$  'a bins  $\Rightarrow$  bool **where**  
*wf-bins*  $\mathcal{G} \ \omega \ bs \equiv \forall k < |bs|. \text{wf-bin } \mathcal{G} \ \omega \ k \ (bs!k)$

Next we prove that inserting well-formed bin items maintains the well-formedness of a bin or bins. The proofs are structurally analogous to those of Section 4.3.

**lemma** *wf-bin-bin-upd*:  
**assumes** *wf-bin*  $\mathcal{G} \ \omega \ k \ b$   
**assumes** *wf-bin-item*  $\mathcal{G} \ \omega \ k \ (\text{item } e)$   
**shows** *wf-bin*  $\mathcal{G} \ \omega \ k \ (\text{bin-upd } e \ b)$

**lemma** *wf-bin-bin-upds*:  
**assumes** *wf-bin*  $\mathcal{G} \ \omega \ k \ b$   
**assumes**  $\forall x \in \text{set } (\text{items } es). \text{wf-bin-item } \mathcal{G} \ \omega \ k \ x$   
**assumes** *distinct* (*items* *es*)  
**shows** *wf-bin*  $\mathcal{G} \ \omega \ k \ (\text{bin-upds } es \ b)$

**lemma** *wf-bins-bins-upd*:  
**assumes** *wf-bins*  $\mathcal{G} \ \omega \ bs$   
**assumes**  $\forall x \in \text{set } (items \ es). \text{wf-bin-item } \mathcal{G} \ \omega \ k \ x$   
**assumes** *distinct*  $(items \ es)$   
**shows** *wf-bins*  $\mathcal{G} \ \omega \ (bins\text{-}upd \ bs \ k \ es)$

At this point we would like to proof that function *Earley-bin-list'* also maintains the well-formedness of the bins. But since it is a partial function we first need to take a short excursion into function definitions in Isabelle. Intuitively, a recursive function terminates if for every recursive call the size of its input strictly decreases. And normally all functions defined in Isabelle must be total. But there are different ways to define a recursive function depending on the complexity of its termination: (1) with the *fun* keyword. Isabelle then tries to find a measure of the input that proves termination. If successful we obtain an induction schema corresponding to the function definition. (2) via the *function* keyword. We then need to define and prove a suitable measure by hand. (3) if the function is a partial function we need to define it with the keyword *partial-function*. For tail-recursive functions the definition is straightforward, otherwise we have to wrap the return type in an option to signal possible non-termination. But contrary to total functions we do *not* obtain the usual induction schema. To prove anything useful about a partial function we have to define the set of inputs and a corresponding measure for which the function terminates and subsequently prove an appropriate induction schema by hand.

As previously explained, in Section 4.1 we defined the function *Earley-bin-list'* as a partial function since a call of the form *Earley-bin-list' k G ω bs i* might never terminate if the function keeps appending arbitrary new items to the *k*-th bin it currently operates on. But we know that the newly generated items are not arbitrary but well-formed bin items. From lemma *finiteness* of Chapter 3 we also know that the set of well-formed items is finite. Since we made sure to only add each item once to a bin, the function *Earley-bin-list'* will eventually run out of new items to insert into the bin it currently operates on and terminate.

In Isabelle we define the set of well-formed earley inputs as a set of quadruples consisting of the index *k* of the current bin, the grammar  $\mathcal{G}$ , the input  $\omega$ , and the bins *bs*. Note that we not only require the bins to be well-formed but also suitable bounds on *k* and the length of the bins to make sure that we are not indexing outside the input or the bins as well as a well-formed grammar to ensure we only generate well-formed bin items. We then define a suitable measure for the termination of *Earley-bin-list' k G ω bs i* that intuitively corresponds to the number of well-formed bin items that are still possible to generate from index *i* onwards. Finally, we prove an induction schema (*earley induction*) for the function by complete induction on the measure of the input. We omit showing the schema explicitly since it is rather verbose. But intuitively it partitions the function

into five cases: the base (*Base*) case where we have run out of items to operate on and terminate; one case for completion (*Complete*) and prediction (*Predict*) each; and two cases for scanning covering the normal (*Scan*) and the special case (*Pass*) where  $k$  exceeds the length of the input.

**definition** *wf-earley-input* ::  $(\text{nat} \times 'a \text{ cfg} \times 'a \text{ sentential} \times 'a \text{ bins}) \text{ set}$  **where**

$$\begin{aligned} \text{wf-earley-input} = \{ \\ & (k, \mathcal{G}, \omega, bs) \mid k \mathcal{G} \omega bs. \\ & k \leq |\omega| \wedge |bs| = |\omega| + 1 \wedge \\ & \text{wf-}\mathcal{G} \mathcal{G} \wedge \\ & \text{wf-bins } \mathcal{G} \omega bs \} \end{aligned}$$

**fun** *earley-measure* ::  $\text{nat} \times 'a \text{ cfg} \times 'a \text{ sentential} \times 'a \text{ bins} \Rightarrow \text{nat} \Rightarrow \text{nat}$  **where**

$$\text{earley-measure } (k, \mathcal{G}, \omega, bs) \ i = \text{card } \{ x \mid x. \text{wf-bin-item } \mathcal{G} \omega k x \} - i$$

Concluding this section, we prove that we maintain the well-formedness of the input for the function *Earley-bin-list'*. The proof is by *earley induction*, lemma *wf-bins-bins-upd* and - straightforward and thus omitted - auxiliary lemmas stating that the scanning, predicting and completing only generates well-formed bin items. The proofs for functions *Earley-bin-list*, *Earley-list*, and *Earley-list* are respectively by definition, by induction on  $k$  using additionally the fact that the initial bins are well-formed, and once more by definition, each time using previously proven lemmas appropriately.

**lemma** *wf-earley-input-Earley-bin-list'*:

**assumes**  $(k, \mathcal{G}, \omega, bs) \in \text{wf-earley-input}$

**shows**  $(k, \mathcal{G}, \omega, \text{Earley-bin-list}' k \mathcal{G} \omega bs) \in \text{wf-earley-input}$

**lemma** *wf-earley-input-Earley-bin-list*:

**assumes**  $(k, \mathcal{G}, \omega, bs) \in \text{wf-earley-input}$

**shows**  $(k, \mathcal{G}, \omega, \text{Earley-bin-list } k \mathcal{G} \omega bs) \in \text{wf-earley-input}$

**lemma** *wf-earley-input-Earley-list*:

**assumes** *wf- $\mathcal{G}$*   $\mathcal{G}$

**assumes**  $k \leq |\omega|$

**shows**  $(k, \mathcal{G}, \omega, \text{Earley-list } k \mathcal{G} \omega) \in \text{wf-earley-input}$

**lemma** *wf-earley-input-Earley-list*:

**assumes** *wf- $\mathcal{G}$*   $\mathcal{G}$

**assumes**  $k \leq |\omega|$

**shows**  $(k, \mathcal{G}, \omega, \text{Earley-list } \mathcal{G} \omega) \in \text{wf-earley-input}$

## 4.5 Soundness

Now we are ready to prove subsumption in both directions. Since functions *Earley-list* and *Earley-list* are structurally identical to *Earley* respectively *Earley*, the main task

for the next two sections is to show that function *Earley-bin-list* or *Earley-bin-list'* computes the same items as the function *Earley-bin* that computes in turn the fixpoint of *Earley-step*. We start with the easier direction: every item generated by the list-based approach is also present in the set-based approach which implies soundness of the list-based algorithm. This is the 'easier' direction due to the fact that during execution of the body of *Earley-bin-list'* we only consider a single item  $x$  in bin  $k$  at position  $i$  and apply the appropriate operation. In contrast, one execution of function *Earley-step* applies the scan, predict and complete operations for all previously computed items.

We start the soundness proof with three auxiliary lemmas proving subsumption of the three basic operations. The proofs of lemmas *Scan-list-sub-Scan*, *Predict-list-sub-Predict*, and *Complete-list-sub-Complete* are each straightforward by definition of the corresponding functions.

**lemma** *Scan-list-sub-Scan*:

**assumes**  $wf\text{-}bins \ \mathcal{G} \ \omega \ bs$   
**assumes**  $bins \ bs \subseteq I$   
**assumes**  $k < |bs|$   
**assumes**  $k < |\omega|$   
**assumes**  $x \in set \ (items \ (bs!k))$   
**assumes**  $next\text{-}symbol \ x = Some \ a$   
**shows**  $set \ (items \ (Scan\text{-}list \ k \ \omega \ a \ x \ pre)) \subseteq Scan \ k \ \omega \ I$

**lemma** *Predict-list-sub-Predict*:

**assumes**  $wf\text{-}bins \ \mathcal{G} \ \omega \ bs$   
**assumes**  $bins \ bs \subseteq I$   
**assumes**  $k < |bs|$   
**assumes**  $x \in set \ (items \ (bs!k))$   
**assumes**  $next\text{-}symbol \ x = Some \ N$   
**shows**  $set \ (items \ (Predict\text{-}list \ k \ \mathcal{G} \ N)) \subseteq Predict \ k \ \mathcal{G} \ I$

**lemma** *Complete-list-sub-Complete*:

**assumes**  $wf\text{-}bins \ \mathcal{G} \ \omega \ bs$   
**assumes**  $bins \ bs \subseteq I$   
**assumes**  $k < |bs|$   
**assumes**  $x \in set \ (items \ (bs!k))$   
**assumes**  $is\text{-}complete \ x$   
**shows**  $set \ (items \ (Complete\text{-}list \ k \ x \ bs \ red)) \subseteq Complete \ k \ I$

We then proof that all items generated by the function *Earley-bin-list'* are also present in the set produced by the function *Earley-bin*. The proof is by *earley induction* for an arbitrary set of items  $I$ . The cases *Base* and *Pass* are trivial. The other three cases follow the same structure and we only highlight the *Complete* case. Lemma *Earley-bin-list-sub-Earley-bin* follows from *Earley-bin-list'-sub-Earley-bin* by definition.

**lemma** *Earley-bin-list'-sub-Earley-bin*: — Detailed: only complete

**assumes**  $(k, \mathcal{G}, \omega, bs) \in \text{wf-earley-input}$

**assumes**  $\text{bins } bs \subseteq I$

**shows**  $\text{bins } (\text{Earley-bin-list}' k \mathcal{G} \omega bs i) \subseteq \text{Earley-bin } k \mathcal{G} \omega I$

*Proof.* We are in the case *Complete*. Hence, the item  $x$  in the  $k$ -th bin at index  $i$  is complete and the new bins  $bs'$  are  $\text{bins-upd } bs k (\text{Complete-list } k x bs i)$ . We can discharge the assumptions of lemma *Complete-list-sub-Complete* by our assumptions of well-formed earley input and  $\text{bins } bs \subseteq I$  and the additional assumption that we are in the *Complete* case, and have  $\text{bins } bs' \subseteq I \cup \text{Complete } k I$ . Since updating the bins maintains well-formedness of the input we can use the induction hypothesis and obtain the fact

$$\text{bins } (\text{Earley-bin-list}' k \mathcal{G} \omega bs i) \subseteq \text{Earley-bin } k \mathcal{G} \omega (I \cup \text{Complete } k I) \quad (1)$$

We also know that  $I \cup \text{Complete } k I \subseteq \text{Earley-bin } k \mathcal{G} \omega I$  since *Earley-bin* is the fixpoint iteration of *Earley-step* that is in turn defined as  $I \cup \text{Scan } k \omega I \cup \text{Complete } k I \cup \text{Predict } k \mathcal{G} I$ . Moreover we know that function *Earley-bin* is monotonic in the set it operates on. And thus we have

$$\text{Earley-bin } k \mathcal{G} \omega (I \cup \text{Complete } k I) \subseteq \text{Earley-bin } k \mathcal{G} \omega (\text{Earley-bin } k \mathcal{G} \omega I) \quad (2)$$

The statement to proof follows from (1), (2) and the fact that the function *Earley-bin* is idempotent. □

**lemma** *Earley-bin-list-sub-Earley-bin*:

**assumes**  $(k, \mathcal{G}, \omega, bs) \in \text{wf-earley-input}$

**assumes**  $\text{bins } bs \subseteq I$

**shows**  $\text{bins } (\text{Earley-bin-list } k \mathcal{G} \omega bs) \subseteq \text{Earley-bin } k \mathcal{G} \omega I$

We prove lemma *Earley-list-sub-Earley* by induction on  $k$  using the additional lemma *Init-list-eq-Init*, that shows that the set of items created by the *Init-list* and *Init* functions are identical, and lemma *Earley-bin-list-sub-Earley-bin*. Lemma *Earley-list-sub-Earley* follows by definition and concludes the first half of the subsumption proof that implies soundness of the list-based implementation due to the soundness proof of the set of Earley items of Chapter 3 (lemma *sound-Earley*).

**lemma** *Init-list-eq-Init*:

**shows**  $\text{bins } (\text{Init-list } \mathcal{G} \omega) = \text{Init } \mathcal{G}$

**lemma** *Earley-list-sub-Earley*:

**assumes**  $wf\text{-}\mathcal{G} \ \mathcal{G} \ k \leq |\omega|$   
**shows**  $bins \ (Earley\text{-}list \ k \ \mathcal{G} \ \omega) \subseteq Earley \ k \ \mathcal{G} \ \omega$

**lemma** *Earley-list-sub-Earley*:

**assumes**  $wf\text{-}\mathcal{G} \ \mathcal{G}$   
**shows**  $bins \ (Earley\text{-}list \ \mathcal{G} \ \omega) \subseteq Earley \ \mathcal{G} \ \omega$

## 4.6 Completeness

In this section we proof completeness of the list-based algorithm. The two main complications are the following. The function *Earley-bin-list'* starts its computation at a specific index  $i$  in the  $k$ -th bin. In contrast, while completing the  $k$ -th bin, the set-based approach of Chapter 3 applies the function *Earley-step* in each iteration of the fixpoint computation to all items. Hence, we have to generalize the proofs such that all items at indices  $j \leq i$  are already 'done'. The second problem is more severe: as stated the algorithm is incorrect, at least for some classes of grammars. In contrast to the fixpoint computation of the set-based approach the list-based implementation imposes an order on the creation of items, and sometimes order matters. Consider for example an item  $A \rightarrow \bullet, i, j$ , or an epsilon-rule  $A \rightarrow \epsilon$ , that the list-based implementation encounters during creation of bin  $B_j$ . Since the item is complete we apply the *Complete* operation. The algorithm first determines the origin bin  $i$  of the item which always coincides with  $j$  for epsilon rules. Consequently, we search the current bin  $B_j$  for any items of the form  $B \rightarrow \alpha \bullet A\beta, i', j$ . But bin  $B_j$  is only partially constructed at this point in time. Hence, we might be missing some of these items, either since they have not been predicted, or completed up to this point. Thus, if we apply the complete operation to item  $A \rightarrow \bullet, i, j$  immediately we might not generate all items of the form  $B \rightarrow \alpha A \bullet \beta, i', j$  and in turn not all items depending on those items. In essence, we might be missing potential derivation paths.

There exist various approaches to deal with this problem. Aho *et al* [Aho:1972] take a rather relaxed point of view and propose to keep interleaving the *Predict* and *Complete* operations until no more new items are being generated. Earley [Earley:1970] suggests to have the *Complete* operation note that we actually need to move the bullet over the non-terminal  $A$  when encountering the item  $A \rightarrow \bullet, i, j$ , and taking this information into account in the subsequent execution of the algorithm. Or, in essence, delaying the *Complete* operation for item  $A \rightarrow \bullet, i, j$  until we are sure that we have encountered all items of the form  $B \rightarrow \alpha \bullet A\beta, i', j$ . Earley suggests that the algorithm should keep an additional collection of non-terminals to look out for stored in an appropriate data structure. Aycock *et al* [Aycock:2002] propose yet another approach based on a slight modification of the *Predict* operation. Note that the problem during completion only



arises if the non-terminal  $A$  is nullable, or there exists a derivation such that  $\mathcal{G} \vdash A \Rightarrow^* \epsilon$ . The authors suggest the following approach. Pre-compute nullable non-terminals using well-know approaches [Appel:2003][Fischer:2009]. If the algorithm encounters an item of the form  $A \rightarrow \alpha \bullet B\beta, i, j$ , predict items  $B \rightarrow \bullet \gamma, j, j$  for each rule  $B \rightarrow \gamma$  of the grammar  $\mathcal{G}$ . But additionally add the item  $A \rightarrow \alpha B \bullet \beta, j, j$  if the non-terminal  $B$  is nullable.

Interleaving prediction and completion until we generate no new items seems rather impractical in our opinion. Thus, we only considered the approaches of Earley and Aycock *et al.* Both ideas are straightforward to implement in the context of a pure recognizer. But complications arise when we need to annotate the items with the needed information to construct parse trees. For the approach of Earley it is no longer sufficient to keep solely a list of nullable non-terminals to look out for but we need to maintain additional information of the origin of these non-terminals to update the reduction and predecessor pointers accordingly. The approach of Aycock *et al* implies even more complications. For a pure recognizer they construct an LR(0) automaton for the modified *Predict* operation, but for an Earley parser they introduce a new type of automaton, a split-epsilon DFA, and also slightly rewrite the grammar into *nihilist normal form* to encode the necessary information to reconstruct derivations.

In the end, we decided against implementing any of the approaches above and follow the approach of Jones [Jones:1972]. We restrict the grammar. If we disallow any non-terminal to derive  $\epsilon$ , the problem does not arise in the first place. Our justification for this approach is that it is by far the simplest solution while still being practical and allowing a wide enough range of grammars to be supported.

Overall, our obligation for the remainder of the section is to prove that restricting the grammar to not contain empty derivations ensures that the order of constructing items does not matter in the end, and that the list-based approach covers the fixpoint computation of Chapter 3.

**definition** *nonempty-derives* :: 'a cfg  $\Rightarrow$  bool **where**  
 $\text{nonempty-derives } \mathcal{G} \equiv \forall N. N \in \text{set } (\mathfrak{N} \mathcal{G}) \longrightarrow \neg (\mathcal{G} \vdash [N] \Rightarrow^* [])$

The core lemma is the following: if the grammar is well-formed and does not allow empty derivations, and a given item is well-formed, sound and complete, then its item origin and item end cannot coincide, which implies that the origin of the item is strictly smaller than the item end due to the well-formedness of the item. And consequently there do not exist any items of the form  $A \rightarrow \epsilon, i, j$  in any bin  $B_j$ .

**lemma** *impossible-complete-item*:

**assumes** *wf- $\mathcal{G}$*   $\mathcal{G}$   
**assumes** *nonempty-derives*  $\mathcal{G}$   
**assumes** *wf-item*  $\mathcal{G} \ \omega \ x$   
**assumes** *sound-item*  $\mathcal{G} \ \omega \ x$

**assumes** *is-complete*  $x$   
**assumes** *item-origin*  $x = k$  *item-end*  $x = k$   
**shows** *False*

*Proof.* From assumptions *sound-item*  $\mathcal{G} \omega x$ , *is-complete*  $x$ , *item-origin*  $x = k$ , and *item-end*  $x = k$  we have by definition of a sound and complete item that

$$\mathcal{G} \vdash \text{item-rule-head } x \Rightarrow^* []$$

Since the grammar  $\mathcal{G}$  and the item  $x$  are well-formed, we also know that the item rule head of  $x$  is indeed a non-terminal. The proof concludes by assumption *nonempty-derives*  $\mathcal{G}$  by definition. □

Lemma *Complete-Un-absorb* then captures the idea that it does not matter for the *Complete* operation if we add an additional item  $z$  of the form  $B \rightarrow \alpha \bullet A\beta, i, k$  to bin  $B_k$  while constructing the  $k$ -th bin under the assumption of well-formedness and non-empty derivations.

**lemma** *Complete-Un-absorb*:

**assumes** *wf- $\mathcal{G}$*   $\mathcal{G}$   
**assumes** *wf-items*  $\mathcal{G} \omega I$   
**assumes** *sound-items*  $\mathcal{G} \omega I$   
**assumes** *nonempty-derives*  $\mathcal{G}$   
**assumes** *wf-item*  $\mathcal{G} \omega z$   
**assumes** *item-end*  $z = k$   
**assumes** *next-symbol*  $z = \text{Some } A$   
**shows** *Complete*  $k (I \cup \{z\}) = \text{Complete } k I$

*Proof.* Assume for the sake of contradiction that *Complete*  $k (I \cup \{z\}) \neq \text{Complete } k I$ . Then we know that *Complete*  $k I \subset \text{Complete } k (I \cup \{z\})$  since the *Complete* operation is monotonic in  $I$ . Hence, there exist by definition of *Complete* items  $x$ ,  $x'$ , and  $y$  such that

$$\begin{aligned}
 x &\in \text{Complete } k (I \cup \{z\}) & (1) & \quad x \notin \text{Complete } k I & (2) \\
 x' &\in \text{bin } (I \cup \{z\}) \text{ (item-origin } y) & (3) & \quad \text{next-symbol } x' = \text{Some (item-rule-head } y) & (4) \\
 y &\in \text{bin } (I \cup \{z\}) \text{ } k & (5) & \quad \text{is-complete } y & (6) \\
 x &= \text{inc-item } x' k & (7)
 \end{aligned}$$

From assumptions (2-7) and the definition of *Complete* we need to consider two cases:

- $z = x'$ : Due to assumption *item-end*  $z = k$  and (3,5) we know that the item origin and end of item  $y$  is  $k$ . Additionally, the item is sound and well-formed due to assumptions (5,6) and *wf-items*  $\mathcal{G} \omega I$ , *wf-item*  $\mathcal{G} \omega z$ , and *sound-items*  $\mathcal{G} \omega I$ , *next-symbol*  $z = \text{Some } A$ . Moreover, using assumptions *wf- $\mathcal{G}$*   $\mathcal{G}$  and *nonempty-derives*  $\mathcal{G}$  and the fact that  $y$  is complete (6), we can discharge the assumptions of lemma *impossible-complete-item* and arrive at a contradiction.
- $z = y$ : Thus we know that  $z$  must be complete since  $y$  is complete by (6). But we also know that *next-symbol*  $z = \text{Some } A$ , a contradiction.

□

Next we prove that the items generated by function *Earley-bin-list'* cover the items generated by a single *Earley-step*. Note the assumption *Earley-step*  $k \mathcal{G} \omega (\text{bins-upto } bs \ k \ i) \subseteq \text{bins } bs$  stating that all items up to index  $i$  can already be considered to be 'done' or applying the function *Earley-step* to any of those items does not change the bins  $bs$ . This assumption is necessary since a call of the form *Earley-bin-list'*  $k \mathcal{G} \omega bs \ i$  intuitively skips the first  $i$  items. The proof is by *earley induction* and we only highlight the *Predict* case where we need lemma *Complete-Un-absorb*. The other cases are similar in overall structure. Lemma *Earley-step-sub-Earley-bin-list* then follows once more by definition.

**lemma** *Earley-step-sub-Earley-bin-list'*:

**assumes**  $(k, \mathcal{G}, \omega, bs) \in \text{wf-earley-input}$

**assumes** *sound-items*  $\mathcal{G} \omega (\text{bins } bs)$

**assumes** *is-sentence*  $\mathcal{G} \omega$

**assumes** *nonempty-derives*  $\mathcal{G}$

**assumes** *Earley-step*  $k \mathcal{G} \omega (\text{bins-upto } bs \ k \ i) \subseteq \text{bins } bs$

**shows** *Earley-step*  $k \mathcal{G} \omega (\text{bins } bs) \subseteq \text{bins } (\text{Earley-bin-list}' \ k \ \mathcal{G} \omega \ bs \ i)$

*Proof.* We are only highlighting the *Predict* case. Hence, we are currently considering an item  $x$  in the  $k$ -th bin at index  $i$  whose next symbol is some non-terminal  $N$ . Let  $bs'$  denote the updated bins or *bins-upd*  $bs \ k \ (\text{Predict-list } k \ \mathcal{G} \ N)$ . We know that the function *bins-upd* maintains well-formedness and soundness of the items, but to apply our induction hypothesis we need to prove one additional statement:

$$\text{Earley-step } k \ \mathcal{G} \omega (\text{bins-upto } bs' \ k \ (i + 1)) \subseteq \text{bins } bs'$$

Since *Earley-step* is defined as the union of the basic three operations we split this proof into these three cases:

- $Scan\ k\ \omega\ (bins\text{-}upto\ bs'\ k\ (i + 1)) \subseteq bins\ bs'$ :

$$\begin{aligned}
& Scan\ k\ \omega\ (bins\text{-}upto\ bs'\ k\ (i + 1)) \\
&= Scan\ k\ \omega\ (bins\text{-}upto\ bs'\ k\ i \cup \{items\ (bs'\ !\ k)\ !\ i\}) \quad (1) \\
&= Scan\ k\ \omega\ (bins\text{-}upto\ bs\ k\ i \cup \{x\}) \quad (2) \\
&\subseteq bins\ bs \cup Scan\ k\ \omega\ \{x\} \quad (3) \\
&= bins\ bs \quad (4) \\
&\subseteq bins\ bs' \quad (5)
\end{aligned}$$

(1) by definition of *bins-upto*. (2) function *bins-upd* does not change the order of the items of bin *k* upto and including index *i*. (3) function *Scan* distributes over set union, assumption *Earley-step*  $k\ \mathcal{G}\ \omega\ (bins\text{-}upto\ bs\ k\ i) \subseteq bins\ bs$  and the definition of *Earley-step*. (4) the next symbol of *x* is the non-terminal *N* and thus the *Scan* operation yield an empty set. (5) the set semantics of function *bins-upd*.

- $Predict\ k\ \mathcal{G}\ (bins\text{-}upto\ bs'\ k\ (i + 1)) \subseteq bins\ bs'$

$$\begin{aligned}
& Predict\ k\ \mathcal{G}\ (bins\text{-}upto\ bs'\ k\ (i + 1)) \\
&= Predict\ k\ \mathcal{G}\ (bins\text{-}upto\ bs'\ k\ i \cup \{items\ (bs'\ !\ k)\ !\ i\}) \quad (1) \\
&= Predict\ k\ \mathcal{G}\ (bins\text{-}upto\ bs\ k\ i \cup \{x\}) \quad (2) \\
&\subseteq bins\ bs \cup Predict\ k\ \mathcal{G}\ \{x\} \quad (3) \\
&= bins\ bs \cup set\ (items\ (Predict\text{-}list\ k\ \mathcal{G}\ N)) \quad (4) \\
&\subseteq bins\ bs' \quad (5)
\end{aligned}$$

(1-3,5) are identical to the first case. (4) the next symbol of *x* is the non-terminal *N* and thus the list-based implementation yields the same items as the set-based implementation.

- $Complete\ k\ (bins\text{-}upto\ bs'\ k\ (i + 1)) \subseteq bins\ bs'$

$$\begin{aligned}
& Complete\ k\ (bins\text{-}upto\ bs'\ k\ (i + 1)) \\
&= Complete\ k\ (bins\text{-}upto\ bs'\ k\ i \cup \{items\ (bs'\ !\ k)\ !\ i\}) \quad (1) \\
&= Complete\ k\ (bins\text{-}upto\ bs\ k\ i \cup \{x\}) \quad (2) \\
&= Complete\ k\ (bins\text{-}upto\ bs\ k\ i) \quad (3) \\
&\subseteq bins\ bs \quad (4) \\
&\subseteq bins\ bs' \quad (5)
\end{aligned}$$

(1-2,5) are identical to the first case. (3) by lemma *Complete-Un-absorb* using the well-formedness, soundness, non-empty derivation assumptions and the fact that

the item  $x$  is in the  $k$ -th bin and its next symbol is the non-terminal  $N$ . (4) by assumption  $\text{Earley-step } k \mathcal{G} \omega (\text{bins-upto } bs \ k \ i) \subseteq \text{bins } bs$  and the definition of *Earley-step*.

□

**lemma** *Earley-step-sub-Earley-bin-list*:

**assumes**  $(k, \mathcal{G}, \omega, bs) \in \text{wf-earley-input}$   
**assumes**  $\text{sound-items } \mathcal{G} \ \omega \ (\text{bins } bs)$   
**assumes**  $\text{is-sentence } \mathcal{G} \ \omega$   
**assumes**  $\text{nonempty-derives } \mathcal{G}$   
**assumes**  $\text{Earley-step } k \ \mathcal{G} \ \omega \ (\text{bins-upto } bs \ k \ 0) \subseteq \text{bins } bs$   
**shows**  $\text{Earley-step } k \ \mathcal{G} \ \omega \ (\text{bins } bs) \subseteq \text{bins } (\text{Earley-bin-list } k \ \mathcal{G} \ \omega \ bs)$

We have proven that the items generated by the execution of the list-based approach covers *one* single step of the set-based approach. Our next objective is to generalize this statement to the whole fixpoint computation, or an arbitrary number of steps. We need two, albeit small, quite technical lemmas, proving that the function *Earley-bin-list* is idempotent. This follows from the next lemma which states that when we execute the function *Earley-bin-list'* two times, passing as the argument for the bins of the second round the result of the first round, and are starting the execution from possibly different initial indices the result of the smaller index prevails. The intuition is clear: if we run through the worklist starting from index  $i \leq j$ , starting a second time from index  $j$  does not yield any new items, since we already covered all items of the second execution in the first turn and order does not matter due to the assumption of non-empty derivations. The proof is by *earley induction* for arbitrary  $j$  and once more utilizes lemma *impossible-complete-item*, we omit showing any details.

**lemma** *Earley-bin-list'-idem*:

**assumes**  $(k, \mathcal{G}, \omega, bs) \in \text{wf-earley-input}$   
**assumes**  $\text{sound-items } \mathcal{G} \ \omega \ (\text{bins } bs)$   
**assumes**  $\text{nonempty-derives } \mathcal{G}$   
**assumes**  $i \leq j$   
**shows**  $\text{bins } (\text{Earley-bin-list}' k \ \mathcal{G} \ \omega \ (\text{Earley-bin-list}' k \ \mathcal{G} \ \omega \ bs \ i) \ j) = \text{bins } (\text{Earley-bin-list}' k \ \mathcal{G} \ \omega \ bs \ i)$

**lemma** *Earley-bin-list-idem*:

**assumes**  $(k, \mathcal{G}, \omega, bs) \in \text{wf-earley-input}$   
**assumes**  $\text{sound-items } \mathcal{G} \ \omega \ (\text{bins } bs)$   
**assumes**  $\text{nonempty-derives } \mathcal{G}$   
**shows**  $\text{bins } (\text{Earley-bin-list } k \ \mathcal{G} \ \omega \ (\text{Earley-bin-list } k \ \mathcal{G} \ \omega \ bs)) = \text{bins } (\text{Earley-bin-list } k \ \mathcal{G} \ \omega \ bs)$

Lemma *Earley-bin-sub-Earley-bin-list* concludes the subsumption proof for a single bin. Since the function *Earley-bin* is defined as the fixpoint of the function *Earley-step* and the fact that  $x \in \text{limit } f \ X \equiv \exists n. x \in \text{funpower } f \ n \ X$  the core proof is by induction on the computation of *funpower*

**lemma** *Earley-bin-sub-Earley-bin-list*:

**assumes**  $(k, \mathcal{G}, \omega, bs) \in \text{wf-earley-input}$

**assumes** *sound-items*  $\mathcal{G} \ \omega \ (bins \ bs)$

**assumes** *is-sentence*  $\mathcal{G} \ \omega$

**assumes** *nonempty-derives*  $\mathcal{G}$

**assumes** *Earley-step*  $k \ \mathcal{G} \ \omega \ (bins\text{-}upto \ bs \ k \ 0) \subseteq bins \ bs$

**shows** *Earley-bin*  $k \ \mathcal{G} \ \omega \ (bins \ bs) \subseteq bins \ (Earley\text{-}bin\text{-}list \ k \ \mathcal{G} \ \omega \ bs)$

*Proof.* The goal is  $\text{funpower} \ (Earley\text{-}step \ k \ \mathcal{G} \ \omega) \ (Suc \ n) \ (bins \ bs) \subseteq bins \ (Earley\text{-}bin\text{-}list \ k \ \mathcal{G} \ \omega \ bs)$ .

For the base case we have  $\text{funpower} \ (Earley\text{-}step \ k \ \mathcal{G} \ \omega) \ 0 \ (bins \ bs) = bins \ bs$ . And we conclude the proof due to the fact that the function *Earley-bin-list* is monotonic in the bins.

For the induction step we first need to proof a necessary precondition for our induction hypothesis:

$$\begin{aligned} & Earley\text{-}step \ k \ \mathcal{G} \ \omega \ (bins\text{-}upto \ (Earley\text{-}bin\text{-}list \ k \ \mathcal{G} \ \omega \ bs) \ k \ 0) \\ &= Earley\text{-}step \ k \ \mathcal{G} \ \omega \ (bins\text{-}upto \ bs \ k \ 0) & (1) \\ &\subseteq bins \ bs & (2) \\ &\subseteq bins \ (Earley\text{-}bin\text{-}list \ k \ \mathcal{G} \ \omega \ bs) & (3) \end{aligned}$$

(1) *Earley-bin-list*  $k \ \mathcal{G} \ \omega \ bs$  does not change the contents of any bins  $B_j$  where  $j < k$  by definition of *bins-upto*. (2) by assumption. (3) function *Earley-bin-list* only adds to the bins.

$$\begin{aligned} & \text{funpower} \ (Earley\text{-}step \ k \ \mathcal{G} \ \omega) \ (Suc \ n) \ (bins \ bs) \\ &= Earley\text{-}step \ k \ \mathcal{G} \ \omega \ (\text{funpower} \ (Earley\text{-}step \ k \ \mathcal{G} \ \omega) \ n \ (bins \ bs)) & (1) \\ &\subseteq Earley\text{-}step \ k \ \mathcal{G} \ \omega \ (bins \ (Earley\text{-}bin\text{-}list \ k \ \mathcal{G} \ \omega \ bs)) & (2) \\ &\subseteq bins \ (Earley\text{-}bin\text{-}list \ k \ \mathcal{G} \ \omega \ (Earley\text{-}bin\text{-}list \ k \ \mathcal{G} \ \omega \ bs)) & (3) \\ &\subseteq bins \ (Earley\text{-}bin\text{-}list \ k \ \mathcal{G} \ \omega \ bs) & (4) \end{aligned}$$

(1) by definition of *funpower*. (2) by induction hypothesis and fact that the function *Earley-step* is monotonic in the set of items. (3) by lemma *Earley-step-sub-Earley-bin-list* using well-formedness, soundness, non-empty derivations assumptions. (4) by lemma *Earley-bin-list-idem* using once more the soundness and non-empty derivation assumptions.  $\square$

We finish the subsumption proof with lemmas *Earley-sub-Earley-list* and *Earley-sub-Earley-list*. The proofs are respectively by induction on  $k$  using lemmas *Init-list-eq-Init* and *Earley-bin-sub-Earley-bin-list*, and once more by definition using the previous lemma.

**lemma** *Earley-sub-Earley-list*:

**assumes**  $wf\text{-}\mathcal{G} \ \mathcal{G}$

**assumes**  $is\text{-sentence} \ \mathcal{G} \ \omega$

**assumes**  $nonempty\text{-derives} \ \mathcal{G}$

**assumes**  $k \leq |\omega|$

**shows**  $Earley \ k \ \mathcal{G} \ \omega \subseteq bins \ (Earley\text{-list} \ k \ \mathcal{G} \ \omega)$

**lemma** *Earley-sub-Earley-list*:

**assumes**  $wf\text{-}\mathcal{G} \ \mathcal{G}$

**assumes**  $is\text{-sentence} \ \mathcal{G} \ \omega$

**assumes**  $nonempty\text{-derives} \ \mathcal{G}$

**shows**  $Earley \ \mathcal{G} \ \omega \subseteq bins \ (Earley\text{-list} \ \mathcal{G} \ \omega)$

## 4.7 Correctness

We conclude the chapter presenting the final correctness theorem stating that there exists a finished item in the bins generated by the list-based implementation if and only if there exists a derivation of the input from the start symbol of the grammar. The proof is by lemmas *correctness-Earley*, *Earley-list-sub-Earley*, and *Earley-sub-Earley-list*.

**theorem** *correctness-Earley-list*:

**assumes**  $wf\text{-}\mathcal{G} \ \mathcal{G}$

**assumes**  $is\text{-sentence} \ \mathcal{G} \ \omega$

**assumes**  $nonempty\text{-derives} \ \mathcal{G}$

**shows**  $recognizing \ (bins \ (Earley\text{-list} \ \mathcal{G} \ \omega)) \ \mathcal{G} \ \omega \longleftrightarrow \mathcal{G} \vdash [\mathcal{G} \ \mathcal{G}] \Rightarrow^* \omega$



## 5 Earley Parser Implementation

Although a recognizer is a useful tool, for most practical applications we would like to not only know if the language specified by the grammar accepts the input, but we also want to obtain additional information of how the input can be derived in the form of parse trees. In particular, for our running example, the grammar  $S ::= S + S \mid x$  and the input  $\omega = x + x + x$  we want to obtain the two possible parse trees illustrated in Figures 5.1 5.2. But constructing all possible parse trees at once is no trivial task.

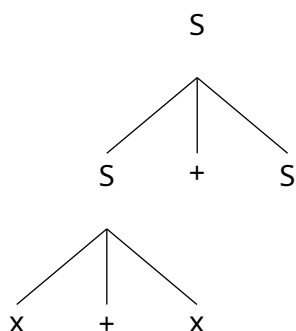


Figure 5.1: Parse Tree:  $\omega = (x + x) + x$

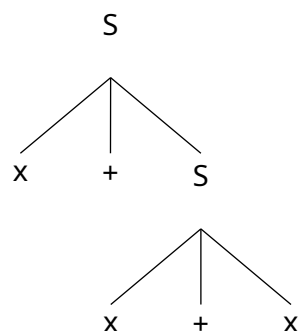


Figure 5.2: Parse Tree:  $\omega = x + (x + x)$

Earley [Earley:1970] turns his recognizer into a parser by adding the following pointers. If the algorithm performs a completion and constructs an item  $B \rightarrow \alpha A \bullet \beta, i, k$ , it adds a pointer from the *instance of the non-terminal A* to the complete item  $A \rightarrow \gamma \bullet, j, k$ . If there exists more than one possible way to complete the non-terminal  $A$  and obtain the item  $B \rightarrow \alpha A \bullet \beta, i, k$ , then multiple pointers originate from the instance of the non-terminal  $A$ . Annotating every non-terminal of the right-hand side of the item  $A \rightarrow \gamma \bullet, j, k$  recursively with pointers thus represents the derivation trees for the non-terminal  $A$ . After termination of the algorithm, the non-terminal that represents the start symbol contains pointers representing all possible derivation trees.

Note that Earley's pointers connect instances of non-terminals, but Tomita [Tomita:1985] showed that this approach is incorrect and may lead to spurious derivations in certain cases. Scott [Scott:2008] presents an example for the grammar  $S ::= SS \mid x$  and the input  $\omega = xxx$ . Earley's parser correctly constructs the parse trees for the input but

additionally returns erroneous parse trees representing derivations of  $xx$  and  $xxxx$ . The problem lies in the fact that left- and rightmost derivations are intertwined when they should not be, since pointers originate from instances of non-terminals and don't connect Earley items.

The most well-known data structure for representing all possible derivations, a shared packed parse forest (SPPF), was introduced by Tomita [Tomita:1985]. But Johnson [Johnson:1991] has shown that Tomita's representation of SPPFs are of worst case unbounded polynomial size and thus would turn our  $\mathcal{O}(n^4)$  recognizer into an unbounded polynomial parser. Scott [Scott:2008] adjust the SPPF data structure slightly and presents two algorithms based on Earley's recognizer that are of worst case cubic space and time. Unfortunately, these algorithms are highly non-trivial and very much imperative in nature, and thus not only exceed the scope of this thesis but are also very difficult to map to a functional approach.

In this chapter we develop an efficient functional algorithm constructing a single parse tree in Section 5.1 and prove its correctness. In Section 5.2 we generalize this approach, introducing a data structure representing all possible parse trees as a parse forest, adjusting the parse tree algorithm to compute such a forest and prove termination and soundness of the algorithm. Finally, in Section 5.3 we discuss the performance, the missing completeness proof of the algorithm and compare our approach to the algorithm of Scott in greater detail.

## 5.1 A Single Parse Tree

The data structure *tree* represents parse trees as shown in Figures 5.1 5.2. A *Leaf* always contains a single symbol (either terminal or non-terminal for partial derivation trees), a *Branch* consists of one non-terminal symbol and a list of subtrees. The function *root-tree* returns the symbol of the root of the parse tree. The yield of a leaf is the single symbol; to compute the yield for a branch with subtrees *ts* we apply the function *yield-tree* recursively and concatenate the results.

```
datatype 'a tree =
  Leaf 'a
| Branch 'a 'a tree list

fun root-tree :: 'a tree  $\Rightarrow$  'a where
  root-tree (Leaf a) = a
| root-tree (Branch N _) = N

fun yield-tree :: 'a tree  $\Rightarrow$  'a sentential where
  yield-tree (Leaf a) = [a]
```

|  $yield-tree (Branch - ts) = concat (map yield-tree ts)$

We introduce three notions of well-formedness for parse trees:

- *wf-rule-tree*: A parse tree must represent a valid (partial) derivation tree according to the grammar  $\mathcal{G}$ . A leaf of a parse tree is always well-formed by construction, for each branch  $Branch\ N\ ts$  there has to exist a corresponding rule of the grammar  $\mathcal{G}$  such that  $N \rightarrow map\ root-tree\ ts$  and each subtree  $t \in set\ ts$  is well-formed.
- *wf-item-tree*: Each branch  $Branch\ N\ ts$  corresponds to an Earley item  $N \rightarrow \alpha \bullet \beta, i, j$  such that the roots of the subtrees  $ts$  and  $\alpha$  coincide. Note that a branch is only well-formed according to the grammar if the roots of the subtrees are a *complete* right-hand side of a production rule of the grammar. In contrast, a branch is well-formed according to an item if the roots of the subtrees are equal to  $\alpha$ , or, since we assume that Earley items are themselves well-formed, a *prefix* of a right-hand side of a production rule.
- *wf-yield-tree*: For an item  $N \rightarrow \alpha \bullet \beta, i, j$ , the yield of a parse tree has to match the substring  $\omega[i..j]$  of the input.

```
fun wf-rule-tree :: 'a cfg  $\Rightarrow$  'a tree  $\Rightarrow$  bool where
  wf-rule-tree - (Leaf a)  $\longleftrightarrow$  True
| wf-rule-tree  $\mathcal{G}$  (Branch N ts)  $\longleftrightarrow$  (
  ( $\exists r \in set\ (\mathfrak{R}\ \mathcal{G}).\ N = rule-head\ r \wedge map\ root-tree\ ts = rule-body\ r$ )  $\wedge$ 
  ( $\forall t \in set\ ts.\ wf-rule-tree\ \mathcal{G}\ t$ ))
```

```
fun wf-item-tree :: 'a cfg  $\Rightarrow$  'a item  $\Rightarrow$  'a tree  $\Rightarrow$  bool where
  wf-item-tree  $\mathcal{G}$  - (Leaf a)  $\longleftrightarrow$  True
| wf-item-tree  $\mathcal{G}\ x$  (Branch N ts)  $\longleftrightarrow$  (
   $N = item-rule-head\ x \wedge$ 
   $map\ root-tree\ ts = take\ (item-bullet\ x)\ (item-rule-body\ x) \wedge$ 
  ( $\forall t \in set\ ts.\ wf-rule-tree\ \mathcal{G}\ t$ ))
```

```
definition wf-yield-tree :: 'a sentential  $\Rightarrow$  'a item  $\Rightarrow$  'a tree  $\Rightarrow$  bool where
  wf-yield-tree  $\omega\ x\ t \equiv yield-tree\ t = \omega[item-origin\ x..item-end\ x]$ 
```

### 5.1.1 Pointer Lemmas

In Chapter 4 we extended the algorithm of Chapter 3 in two orthogonal ways: implementing sets as lists and adding the additional information to construct parse trees in the form null, predecessor, and predecessor/reduction pointers. But we did not formally define the semantics of these pointers nor prove anything about their construction. In the following, we define and proof soundness of the pointers.

- A null pointer *Null* of an entry is sound if it *predicts* the item  $x$  of the entry, or the bullet of  $x$  is at the beginning of the right-hand side of its production rule and we have not yet scanned any substring of the input.
- A predecessor pointer *Pre pre* of an entry  $e$  is sound for the input  $\omega$ , bins  $bs$ , and the index of the current bin  $k$  if  $k > 0$ , the predecessor index does not exceed the length of the predecessor bin at index  $k - 1$ , and the predecessor item in bin  $k - 1$  at index *pre scans* then item of the entry  $e$ . Item  $x'$  *scans* item  $x$  for index  $k$  if the next symbol of  $x'$  coincides with the terminal symbol at index  $k - 1$  in the input  $\omega$  and the item  $x$  can be obtained by *inc-item*  $x' k$ .
- We define the soundness a pointer *PreRed p ps* of an entry  $e$  for each predecessor/reduction triple  $(k', pre, red) \in set (p \# ps)$ . The index  $k'$  of the predecessor bin must be strictly smaller than  $k$ , and both the predecessor and the reduction index must be within the bounds of their respective bins, or bin  $k'$  and  $k$ . Additionally, predicate *completes* holds for  $k$ , the predecessor item  $x'$ , the item  $x$  of entry  $e$  and the reduction item  $y$ , capturing the semantics of the *Complete* operation: The next symbol of  $x'$  is the non-terminal  $N$  which coincides with the item rule head of  $y$ . Furthermore, the item  $y$  is complete and the origin index of  $y$  aligns with the end index of  $x'$ . Finally, item  $x$  can be obtained once more by *inc-item*  $x' k$ .

**definition** *predicts* :: 'a item  $\Rightarrow$  bool **where**  
*predicts*  $x \equiv item\_bullet\ x = 0 \wedge item\_origin\ x = item\_end\ x$

**definition** *sound-null-ptr* :: 'a entry  $\Rightarrow$  bool **where**  
*sound-null-ptr*  $e \equiv pointer\ e = Null \longrightarrow predicts\ (item\ e)$

**definition** *scans* :: 'a sentential  $\Rightarrow$  nat  $\Rightarrow$  'a item  $\Rightarrow$  'a item  $\Rightarrow$  bool **where**  
*scans*  $\omega\ k\ x'\ x \equiv x = inc\_item\ x'\ k \wedge (\exists a. next\_symbol\ x' = Some\ a \wedge \omega!(k-1) = a)$

**definition** *sound-pre-ptr* :: 'a sentential  $\Rightarrow$  'a bins  $\Rightarrow$  nat  $\Rightarrow$  'a entry  $\Rightarrow$  bool **where**  
*sound-pre-ptr*  $\omega\ bs\ k\ e \equiv \forall pre. pointer\ e = Pre\ pre \longrightarrow$   
 $k > 0 \wedge pre < |bs!(k-1)| \wedge$   
*scans*  $\omega\ k\ (item\ (bs!(k-1)!pre))\ (item\ e)$

**definition** *completes* :: nat  $\Rightarrow$  'a item  $\Rightarrow$  'a item  $\Rightarrow$  'a item  $\Rightarrow$  bool **where**  
*completes*  $k\ x'\ x\ y \equiv x = inc\_item\ x'\ k \wedge is\_complete\ y \wedge item\_origin\ y = item\_end\ x' \wedge$   
 $(\exists N. next\_symbol\ x' = Some\ N \wedge N = item\_rule\_head\ y)$

**definition** *sound-prered-ptr* :: 'a bins  $\Rightarrow$  nat  $\Rightarrow$  'a entry  $\Rightarrow$  bool **where**  
*sound-prered-ptr*  $bs\ k\ e \equiv \forall p\ ps\ k'\ pre\ red. pointer\ e = PreRed\ p\ ps \wedge$   
 $(k', pre, red) \in set\ (p\#ps) \longrightarrow k' < k \wedge pre < |bs!k'| \wedge red < |bs!k| \wedge$

*completes*  $k$  (*item* ( $bs!k!pre$ )) (*item*  $e$ ) (*item* ( $bs!k!red$ ))

**definition** *sound-ptrs* :: 'a sentential  $\Rightarrow$  'a bins  $\Rightarrow$  bool **where**

*sound-ptrs*  $\omega$   $bs \equiv \forall k < |bs|. \forall e \in \text{set } (bs!k).$

*sound-null-ptr*  $e \wedge \text{sound-pre-ptr } \omega$   $bs$   $k$   $e \wedge \text{sound-prered-ptr } bs$   $k$   $e$

We then prove the semantics of the pointers. The structure of the proofs is as always: we first proof pointer soundness for the basic operations *bin-upd*, *bin-upds*, and *bins-upd*. Followed by the corresponding proofs for the computation of a single bin or functions *Earley-bin-list'* and *Earley-bin-list*. Finally, we prove that the initial bins are sound, and functions *Earley-list* and *Earley-list* maintain this property. Although it should be intuitively clear that the semantics of pointers hold, the proofs are surprisingly not trivial at all, especially the soundness proofs for functions *bin-upd* and *Earley-bin-list'*. The complexity mostly stems from the predecessor/reduction case that requires a quite significant amount of case splitting due to the indexing and depending on the type of the newly inserted items. Nonetheless, since the proofs are very technical but do not reveal anything new in structure, we only state them and omit going into detail.

**lemma** *sound-ptrs-bin-upd*:

**assumes**  $k < |bs|$

**assumes** *distinct* (*items* ( $bs!k$ ))

**assumes** *sound-ptrs*  $\omega$   $bs$

**assumes** *sound-null-ptr*  $e$

**assumes** *sound-pre-ptr*  $\omega$   $bs$   $k$   $e$

**assumes** *sound-prered-ptr*  $bs$   $k$   $e$

**shows** *sound-ptrs*  $\omega$  ( $bs[k := \text{bin-upd } e (bs!k)]$ )

**lemma** *sound-mono-ptrs-bin-upds*:

**assumes**  $k < |bs|$

**assumes** *distinct* (*items* ( $bs!k$ ))

**assumes** *distinct* (*items*  $es$ )

**assumes** *sound-ptrs*  $\omega$   $bs$

**assumes**  $\forall e \in \text{set } es. \text{sound-null-ptr } e \wedge \text{sound-pre-ptr } \omega$   $bs$   $k$   $e \wedge \text{sound-prered-ptr } bs$   $k$   $e$

**shows** *sound-ptrs*  $\omega$  ( $bs[k := \text{bin-upds } es (bs!k)]$ )

**lemma** *sound-mono-ptrs-Earley-bin-list'*:

**assumes**  $(k, \mathcal{G}, \omega, bs) \in \text{wf-earley-input}$

**assumes** *nonempty-derives*  $\mathcal{G}$

**assumes** *sound-items*  $\mathcal{G}$   $\omega$  (*bins*  $bs$ )

**assumes** *sound-ptrs*  $\omega$   $bs$

**shows** *sound-ptrs*  $\omega$  (*Earley-bin-list'*  $k$   $\mathcal{G}$   $\omega$   $bs$   $i$ )

**lemma** *sound-mono-ptrs-Earley-bin-list*:

**assumes**  $(k, \mathcal{G}, \omega, bs) \in \text{wf-earley-input}$

**assumes** *nonempty-derives*  $\mathcal{G}$   
**assumes** *sound-items*  $\mathcal{G} \ \omega \ (\text{bins } bs)$   
**assumes** *sound-ptrs*  $\omega \ bs$   
**shows** *sound-ptrs*  $\omega \ (\text{Earley-bin-list } k \ \mathcal{G} \ \omega \ bs)$

**lemma** *sound-mono-ptrs-Init-list:*  
**shows** *sound-ptrs*  $\omega \ (\text{Init-list } \mathcal{G} \ \omega)$

**lemma** *sound-mono-ptrs-Earley-list:*  
**assumes** *wf- $\mathcal{G}$*   $\mathcal{G}$   
**assumes** *nonempty-derives*  $\mathcal{G}$   
**assumes**  $k \leq |\omega|$   
**shows** *sound-ptrs*  $\omega \ (\text{Earley-list } k \ \mathcal{G} \ \omega)$

**lemma** *sound-mono-ptrs- $\mathcal{E}$ arley-list:*  
**assumes** *wf- $\mathcal{G}$*   $\mathcal{G}$   
**assumes** *nonempty-derives*  $\mathcal{G}$   
**shows** *sound-ptrs*  $\omega \ (\mathcal{E}\text{arley-list } \mathcal{G} \ \omega)$

### 5.1.2 The Parse Tree Algorithm

After execution of the  *$\mathcal{E}$ arley-list* algorithm we obtain bins representing the complete set of Earley items. The null, predecessor, and predecessor/reduction pointers provide a way to navigate between items or through these bins, and, since they are sound, a way to construct derivation trees. The function *build-tree'* constructs the *single* parse tree corresponding to the derivation tree represented by the item  $x$  of entry  $e$  at index  $i$  at the  $k$ -th bin according to the well-formedness definitions of the beginning of this section.

If the pointer of entry  $e$  is a null pointer, the algorithm starts building the tree rooted at the left-hand side non-terminal  $N$  of the production rule of the item  $x$  by constructing an initially empty branch containing the non-terminal  $N$  and an empty list of subtrees. If the algorithm encounters a predecessor pointer *Pre*  $pre$ , it first recursively calls itself, for bin  $B_{k-1}$  and the predecessor index  $pre$ , obtaining a partial parse tree *Branch*  $N \ ts$ . Since the predecessor pointer is sound, in particular the *scans* predicate holds, we append a Leaf containing the terminal symbol at index  $k - 1$  of the input  $\omega$  to the list of subtrees  $ts$ . In the case that the pointer contains predecessor/reduction triples the algorithm only considers the first triple  $(k', pre, red)$  since we are only constructing a single derivation tree. As for the predecessor case, it recursively calls itself obtaining a partial derivation tree *Branch*  $N \ ts$  for the predecessor index  $pre$  and bin  $k'$ , followed by yet another recursive call for the reduction item at the reduction index  $red$  in the current bin  $k$ , constructing a complete derivation tree  $t$ . This time the *completes* predicate holds, thus the next symbol of the predecessor item coincides with the item rule head of the reduction item, or we are allowed to append the complete tree  $t$  to the list of subtrees  $ts$ .

Some minor implementation details to note are: the function *build-tree'* is a partial function, and not tail recursive, hence we it has to return an optional value, as explained in Section 4.4. We are using the monadic do-notation commonly found in functional programming languages for the option monad. An alternative but equivalent implementation would use explicit case distinctions. Finally, if the function computes some value it is always a branch, never a single leaf.

**partial-function** (*option*) *build-tree'* :: 'a bins  $\Rightarrow$  'a sentential  $\Rightarrow$  nat  $\Rightarrow$  nat  $\Rightarrow$  'a tree option **where**

```

build-tree' bs  $\omega$  k i = (
  let e = bs!k!i in (
    case pointer e of
      Null  $\Rightarrow$  Some (Branch (item-rule-head (item e)) [])
    | Pre pre  $\Rightarrow$  (
      do {
        t  $\leftarrow$  build-tree' bs  $\omega$  (k-1) pre;
        case t of
          Branch N ts  $\Rightarrow$  Some (Branch N (ts @ [Leaf ( $\omega$ !(k-1))]))
        | -  $\Rightarrow$  None }
      | PreRed (k', pre, red) -  $\Rightarrow$  (
        do {
          t  $\leftarrow$  build-tree' bs  $\omega$  k' pre;
          case t of
            Branch N ts  $\Rightarrow$ 
              do {
                t  $\leftarrow$  build-tree' bs  $\omega$  k red;
                Some (Branch N (ts @ [t]))
              }
          | -  $\Rightarrow$  None })))
  )

```

Finally, function *build-tree* computes a complete derivation tree if there exists one. It searches the last bin for any finished items or items of the form  $S \rightarrow \gamma, 0, n$  where  $S$  is the start symbol of the grammar  $\mathcal{G}$  and  $n$  denotes the length of the input  $\omega$ . If there exists such an item it calls function *build-tree'* obtaining some parse tree representing the derivation  $\mathcal{G} \vdash S \Rightarrow^* \omega$  (we will have to proof that it never returns *None*), otherwise it returns *None* since there cannot exist a valid parse tree due to the correctness proof of Chapter 3.

**definition** *build-tree* :: 'a cfg  $\Rightarrow$  'a sentential  $\Rightarrow$  'a bins  $\Rightarrow$  'a tree option **where**

```

build-tree  $\mathcal{G}$   $\omega$  bs  $\equiv$ 
  let k = |bs| - 1 in (
    case filter-with-index ( $\lambda x$ . is-finished  $\mathcal{G}$   $\omega$  x) (items (bs!k)) of
      []  $\Rightarrow$  None
    | (_, i)#-  $\Rightarrow$  build-tree' bs  $\omega$  k i
  )

```

### 5.1.3 Termination

The function *build-tree'* uses the null, predecessor and predecessor/reduction pointers to navigate through the given bins by calling itself recursively. Sound pointers ensure that we are not indexing outside of the bins, but this does not imply that the algorithm terminates. In the following we outline when it always terminates with some parse tree. Lets assume the function starts its computation at index  $i$  of the  $k$ -th bin. If it encounters a null pointer, it terminates immediately. If the pointer is a simple predecessor pointer, it calls itself recursively for the previous bin. Due to the soundness of the predecessor pointer the index  $k - 1$  of this bin is strictly smaller than  $k$ . A similar argument holds for the first recursive call if the pointer is a predecessor/reduction pointer for the predecessor case ( $k' < k$ ). Or, we are following the pointers *strictly* back to the origin bin and thus must terminate at some point. But for the reduction pointer we run into a problem: the recursive call is for item at index  $i$  is in the same bin  $k$  for a different index  $red$ , which in turn might contain again reduction triples and so on. Hence, it is possible that we end up in a cycle of reductions and never terminate. Take for example the grammer  $A ::= x \mid B \quad B ::= A$  and the input  $\omega = x$ . Table 5.1 illustrates the bins computed by the algorithm of Chapter 3. Bin  $B_1$  contains the entry  $B \rightarrow A\bullet, 0, 1; (0, 2, 0), (0, 2, 2)$  at index 1 and its second reduction triple  $(0, 2, 2)$  a reduction pointer to index 2 of the same bin. There we find the entry  $A \rightarrow B\bullet, 0, 1; (0, 0, 1)$  with a reduction pointer to index 1 completing the cycle. This is indeed valid since the grammar itself is cyclic, allowing for derivations of the form  $A \rightarrow B \rightarrow A \rightarrow \dots \rightarrow A \rightarrow x$ .

Table 5.1: Cyclic reduction pointers

	$B_0$	$B_1$
0	$A \rightarrow \bullet B, 0, 0; \perp$	$A \rightarrow x\bullet, 0, 1; 1$
1	$A \rightarrow \bullet x, 0, 0; \perp$	$B \rightarrow A\bullet, 0, 1; (0, 2, 0), (0, 2, 2)$
2	$B \rightarrow \bullet A, 0, 0; \perp$	$A \rightarrow B\bullet, 0, 1; (0, 0, 1)$

We need to address this problem when constructing all possible parse trees in Section 5.2, but for now we are lucky. While constructing a single parse tree the algorithm always follows the first reduction triple that is created when the entry is constructed initially. Since we only append new entries to bins, the complete reduction item necessarily appears before the new entry with the reduction triple. Furthermore, the implementation of function *bin-upd* also makes sure to not change this triple. Thus, we know for each item in the  $k$ -th bin at index  $i$  that the reduction pointer  $red$ , that we follow while constructing a single parse tree, is strictly smaller than  $i$ . To summarize: if the algorithm encounters a null pointer it terminates immediately, for predecessor pointers it calls itself recursively in a bin with a strictly smaller index, and for reduction pointers it calls



itself in the same bin but for a strictly smaller index. The proofs for the monotonicity of the first reduction pointer for functions *bin-upd*, *bin-upds*, *bins-upd*, *Earley-bin-list'*, *Earley-bin-list*, *Earley-list*, and *Earley-list* are completely analogous to the soundness proof of the pointers. We omit them.

**definition** *mono-red-ptr* :: 'a bins  $\Rightarrow$  bool **where**

*mono-red-ptr* bs  $\equiv \forall k < |bs|. \forall i < |bs!k|.$

$\forall k' \text{ pre red ps. pointer } (bs!k!i) = \text{PreRed } (k', \text{pre}, \text{red}) \text{ ps} \longrightarrow \text{red} < i$

Similarly to Chapter 3 we define a suitable measure and a notion of well-formedness for the input of the function *build-tree'* and proof an induction schema, in the following referred to as *tree induction*, by complete induction on the measure. For the input quadruple  $(bs, \omega, k, i)$  the measure corresponds to the number of entries in the first  $k - 1$  bins plus  $i$ . We call the input well-formed if it must satisfy the following conditions: sound and monotonic pointers, index  $k$  does not exceed the length of the bins, and index  $i$  is within the bounds of the  $k$ -th bin.

**fun** *build-tree'-measure* :: ('a bins  $\times$  'a sentential  $\times$  nat  $\times$  nat)  $\Rightarrow$  nat **where**

*build-tree'-measure* (bs,  $\omega$ , k, i) = foldl (+) 0 (map length (take k bs)) + i

**definition** *wf-tree-input* :: ('a bins  $\times$  'a sentential  $\times$  nat  $\times$  nat) set **where**

*wf-tree-input* = { (bs,  $\omega$ , k, i) | bs  $\omega$  k i.

*sound-ptrs*  $\omega$  bs  $\wedge$  *mono-red-ptr* bs  $\wedge$  k < |bs|  $\wedge$  i < |bs!k| }

To conclude this subsection, we then prove termination of the function *build-tree'*, or for well-formed input it always terminates with some branch, by *tree induction*.

**lemma** *build-tree'-termination*:

**assumes** (bs,  $\omega$ , k, i)  $\in$  *wf-tree-input*

**shows**  $\exists N \text{ ts. build-tree' bs } \omega \text{ k i} = \text{Some (Branch N ts)}$

#### 5.1.4 Correctness

**lemma** *wf-item-tree-build-tree'*: — Detailed

**assumes** (bs,  $\omega$ , k, i)  $\in$  *wf-tree-input*

**assumes** *wf-bins*  $\mathcal{G}$   $\omega$  bs

**assumes** k < |bs|

**assumes** i < |bs!k|

**assumes** *build-tree'* bs  $\omega$  k i = Some t

**shows** *wf-item-tree*  $\mathcal{G}$  (item (bs!k!i)) t

**lemma** *wf-ylid-tree-build-tree'*: — Detailed

**assumes** (bs,  $\omega$ , k, i)  $\in$  *wf-tree-input*

**assumes** *wf-bins*  $\mathcal{G}$   $\omega$  bs

**assumes** k < |bs|

```

assumes  $k \leq |\omega|$ 
assumes  $i < |bs!k|$ 
assumes  $build-tree' \ bs \ \omega \ k \ i = Some \ t$ 
shows  $wf-yield-tree \ \omega \ (item \ (bs!k!i)) \ t$ 

```

**theorem** *wf-rule-root-yield-tree-build-tree:*

```

assumes  $wf-bins \ \mathcal{G} \ \omega \ bs$ 
assumes  $sound-ptrs \ \omega \ bs$ 
assumes  $mono-red-ptr \ bs$ 
assumes  $|bs| = |\omega| + 1$ 
assumes  $build-tree \ \mathcal{G} \ \omega \ bs = Some \ t$ 
shows  $wf-rule-tree \ \mathcal{G} \ t \wedge root-tree \ t = \mathfrak{S} \ \mathcal{G} \wedge yield-tree \ t = \omega$ 

```

**corollary** *wf-rule-root-yield-tree-build-tree- $\mathcal{E}$ arley-list:*

```

assumes  $wf-\mathcal{G} \ \mathcal{G}$ 
assumes  $nonempty-derives \ \mathcal{G}$ 
assumes  $build-tree \ \mathcal{G} \ \omega \ (\mathcal{E}arley-list \ \mathcal{G} \ \omega) = Some \ t$ 
shows  $wf-rule-tree \ \mathcal{G} \ t \wedge root-tree \ t = \mathfrak{S} \ \mathcal{G} \wedge yield-tree \ t = \omega$ 

```

**theorem** *correctness-build-tree- $\mathcal{E}$ arley-list:*

```

assumes  $wf-\mathcal{G} \ \mathcal{G}$ 
assumes  $is-sentence \ \mathcal{G} \ \omega$ 
assumes  $nonempty-derives \ \mathcal{G}$ 
shows  $(\exists t. build-tree \ \mathcal{G} \ \omega \ (\mathcal{E}arley-list \ \mathcal{G} \ \omega) = Some \ t) \longleftrightarrow \mathcal{G} \vdash [\mathfrak{S} \ \mathcal{G}] \Rightarrow^* \omega$ 

```

To conclude this section, We give an informal argument for the running time of  $\mathcal{O}(n^4)$ . We assume that the bins (or argument  $bs$ ) are valid according to Chapter 3 and contain only sound and monotonic pointers. The algorithm  $build-tree'$  calls itself at most once for each entry in the bins.

## 5.2 A Parse Forest

why not simply generate all parse trees integrated top down? yes for single parse tree, no for all since exponential blow up

**datatype**  $'a \ forest =$

```

  FLeaf  $'a$ 
| FBranch  $'a \ 'a \ forest \ list \ list$ 

```

**fun**  $combinations :: 'a \ list \ list \Rightarrow 'a \ list \ list$  **where**

```

  combinations [] = [[]]
| combinations (xs#xss) = [ x#cs . x <- xs, cs <- combinations xss ]

```

**fun**  $trees :: 'a \ forest \Rightarrow 'a \ tree \ list$  **where**

```

trees (FLeaf a) = [Leaf a]
| trees (FBranch N fss) = (
  let tss = (map (λf. concat (map (λf. trees f) fss)) fss) in
  map (λts. Branch N ts) (combinations tss)
)

```

### 5.2.1 The Parse Forest Algorithm

```

fun insert-group :: ('a ⇒ 'k) ⇒ ('a ⇒ 'v) ⇒ 'a ⇒ ('k × 'v list) list ⇒ ('k × 'v list) list where
  insert-group K V a [] = [(K a, [V a])]
| insert-group K V a ((k, vs)#xs) = (
  if K a = k then (k, V a # vs) # xs
  else (k, vs) # insert-group K V a xs
)

```

```

fun group-by :: ('a ⇒ 'k) ⇒ ('a ⇒ 'v) ⇒ 'a list ⇒ ('k × 'v list) list where
  group-by K V [] = []
| group-by K V (x#xs) = insert-group K V x (group-by K V xs)

```

```

partial-function (option) build-trees' :: 'a bins ⇒ 'a sentential ⇒ nat ⇒ nat ⇒ nat set ⇒ 'a
forest list option where

```

```

  build-trees' bs ω k i I = (
    let e = bs!k!i in (
      case pointer e of
        Null ⇒ Some ([FBranch (item-rule-head (item e)) []])
      | Pre pre ⇒ (
        do {
          pres ← build-trees' bs ω (k-1) pre {pre};
          those (map (λf.
            case f of
              FBranch N fss ⇒ Some (FBranch N (fss @ [[FLeaf (ω!(k-1))]]))
            | - ⇒ None
          ) pres)
        })
      | PreRed p ps ⇒ (
        let ps' = filter (λ(k', pre, red). red ∉ I) (p#ps) in
        let gs = group-by (λ(k', pre, red). (k', pre)) (λ(k', pre, red). red) ps' in
        map-option concat (those (map (λ((k', pre), reds).
          do {
            pres ← build-trees' bs ω k' pre {pre};
            rss ← those (map (λred. build-trees' bs ω k red (I ∪ {red})) reds);
            those (map (λf.
              case f of
                FBranch N fss ⇒ Some (FBranch N (fss @ [concat rss]))
            ) pres)
          })
        ) reds)
      )
    )
  )

```

```

      | - ⇒ None
    ) pres)
  }
) gs))
)
))

```

**definition** *build-trees* :: 'a cfg ⇒ 'a sentential ⇒ 'a bins ⇒ 'a forest list option **where**  
*build-trees*  $\mathcal{G} \ \omega \ bs \equiv$   
 let  $k = |bs| - 1$  in  
 let *finished* = filter-with-index ( $\lambda x. \text{is-finished } \mathcal{G} \ \omega \ x$ ) (*items* ( $bs!k$ )) in  
 map-option concat (those (map ( $\lambda(-, i). \text{build-trees}' \ bs \ \omega \ k \ i \ \{i\}$ ) *finished*))

### 5.2.2 Termination

**fun** *build-forest'-measure* :: ('a bins × 'a sentential × nat × nat × nat set) ⇒ nat **where**  
*build-forest'-measure* ( $bs, \omega, k, i, I$ ) = foldl (+) 0 (map length (take (k+1)  $bs$ )) - card  $I$

**definition** *wf-trees-input* :: ('a bins × 'a sentential × nat × nat × nat set) set **where**  
*wf-trees-input* = { ( $bs, \omega, k, i, I$ ) |  $bs \ \omega \ k \ i \ I$ .  
 $\text{sound-ptrs } \omega \ bs \wedge k < |bs| \wedge i < |bs!k| \wedge I \subseteq \{0..<|bs!k|\} \wedge i \in I$  }

**lemma** *build-trees'-termination*:

**assumes** ( $bs, \omega, k, i, I$ ) ∈ *wf-trees-input*  
**shows** ∃  $fs. \text{build-trees}' \ bs \ \omega \ k \ i \ I = \text{Some } fs \wedge (\forall f \in \text{set } fs. \exists N \ fss. f = \text{FBranch } N \ fss)$

**theorem** *termination-build-tree-ℰarley-list*:

**assumes** *wf- $\mathcal{G}$*   $\mathcal{G}$   
**assumes** *nonempty-derives*  $\mathcal{G}$   
**assumes**  $\mathcal{G} \vdash [\mathfrak{S} \ \mathcal{G}] \Rightarrow^* \omega$   
**shows** ∃  $fs. \text{build-trees } \mathcal{G} \ \omega \ (\mathcal{E}arley\text{-list } \mathcal{G} \ \omega) = \text{Some } fs$

### 5.2.3 Soundness

**lemma** *wf-item-tree-build-trees'*:

**assumes** ( $bs, \omega, k, i, I$ ) ∈ *wf-trees-input*  
**assumes** *wf-bins*  $\mathcal{G} \ \omega \ bs$   
**assumes**  $k < |bs|$   
**assumes**  $i < |bs!k|$   
**assumes**  $\text{build-trees}' \ bs \ \omega \ k \ i \ I = \text{Some } fs$   
**assumes**  $f \in \text{set } fs$   
**assumes**  $t \in \text{set } (\text{trees } f)$   
**shows** *wf-item-tree*  $\mathcal{G} \ (\text{item } (bs!k!i)) \ t$

**lemma** *wf- $\text{yield-tree-build-trees}'$* :

**assumes**  $(bs, \omega, k, i, I) \in \text{wf-trees-input}$   
**assumes**  $\text{wf-bins } \mathcal{G} \ \omega \ bs$   
**assumes**  $k < |bs|$   
**assumes**  $k \leq |\omega|$   
**assumes**  $i < |bs!k|$   
**assumes**  $\text{build-trees}' \ bs \ \omega \ k \ i \ I = \text{Some } fs$   
**assumes**  $f \in \text{set } fs$   
**assumes**  $t \in \text{set } (\text{trees } f)$   
**shows**  $\text{wf-yield-tree } \omega \ (\text{item } (bs!k!i)) \ t$

**theorem** *wf-rule-root-yield-tree-build-trees:*

**assumes**  $\text{wf-bins } \mathcal{G} \ \omega \ bs$   
**assumes**  $\text{sound-ptrs } \omega \ bs$   
**assumes**  $|bs| = |\omega| + 1$   
**assumes**  $\text{build-trees } \mathcal{G} \ \omega \ bs = \text{Some } fs$   
**assumes**  $f \in \text{set } fs$   
**assumes**  $t \in \text{set } (\text{trees } f)$   
**shows**  $\text{wf-rule-tree } \mathcal{G} \ t \wedge \text{root-tree } t = \mathfrak{S} \ \mathcal{G} \wedge \text{yield-tree } t = \omega$

**corollary** *wf-rule-root-yield-tree-build-trees-Earley-list:*

**assumes**  $\text{wf-}\mathcal{G} \ \mathcal{G}$   
**assumes**  $\text{nonempty-derives } \mathcal{G}$   
**assumes**  $\text{build-trees } \mathcal{G} \ \omega \ (\text{Earley-list } \mathcal{G} \ \omega) = \text{Some } fs$   
**assumes**  $f \in \text{set } fs$   
**assumes**  $t \in \text{set } (\text{trees } f)$   
**shows**  $\text{wf-rule-tree } \mathcal{G} \ t \wedge \text{root-tree } t = \mathfrak{S} \ \mathcal{G} \wedge \text{yield-tree } t = \omega$

**theorem** *soundness-build-trees-Earley-list:*

**assumes**  $\text{wf-}\mathcal{G} \ \mathcal{G}$   
**assumes**  $\text{is-sentence } \mathcal{G} \ \omega$   
**assumes**  $\text{nonempty-derives } \mathcal{G}$   
**assumes**  $\text{build-trees } \mathcal{G} \ \omega \ (\text{Earley-list } \mathcal{G} \ \omega) = \text{Some } fs$   
**assumes**  $f \in \text{set } fs$   
**assumes**  $t \in \text{set } (\text{trees } f)$   
**shows**  $\text{derives } \mathcal{G} \ [\mathfrak{S} \ \mathcal{G}] \ \omega$

### 5.3 A Word on Performance and Completeness

SNIPPET:

A shared packed parse forest SPPF is a representation designed to reduce the space required to represent multiple derivation trees for an ambiguous sentence. In an SPPF, nodes which have the same tree below them are shared and nodes which correspond to

different derivations of the same substring from the same non-terminal are combined by creating a packed node for each family of children. Nodes can be packed only if their yields correspond to the same portion of the input string. Thus, to make it easier to determine whether two alternates can be packed under a given node, SPPF nodes are labelled with a triple  $(x, i, j)$  where  $a_{j+1} \dots a_i$  is a substring matched by  $x$ . To obtain a cubic algorithm we use binarised SPPFs which contain intermediate additional nodes but which are of worst case cubic size. (EXAMPLE SPPF running example???)

We can turn earley's algorithm into a correct parser by adding pointers between items rather than instances of non-terminals, and labelling the pointers in a way which allows a binarised SPPF to be constructed by walking the resulting structure. However, in order to construct a binarised SPPF we also have to introduce additional nodes for grammar rules of length greater than two, complicating the final algorithm.

## 6 The Running Example

**definition**  $\varepsilon\text{-free} :: 'a \text{ cfg} \Rightarrow \text{bool}$  **where**  
 $\varepsilon\text{-free } \mathcal{G} \longleftrightarrow (\forall r \in \text{set } (\mathfrak{N} \mathcal{G}). \text{rule-body } r \neq [])$

**lemma**  $\varepsilon\text{-free-impl-non-empty-deriv}$ :  
 $\varepsilon\text{-free } \mathcal{G} \Longrightarrow N \in \text{set } (\mathfrak{N} \mathcal{G}) \Longrightarrow \neg (\mathcal{G} \vdash [N] \Rightarrow^* [])$   
**datatype**  $t = x \mid \text{plus}$   
**datatype**  $n = S$   
**datatype**  $s = \text{Terminal } t \mid \text{Nonterminal } n$

**definition**  $\text{nonterminals} :: s \text{ list}$  **where**  
 $\text{nonterminals} = [\text{Nonterminal } S]$

**definition**  $\text{terminals} :: s \text{ list}$  **where**  
 $\text{terminals} = [\text{Terminal } x, \text{Terminal plus}]$

**definition**  $\text{rules} :: s \text{ rule list}$  **where**  
 $\text{rules} = [$   
 $\quad (\text{Nonterminal } S, [\text{Terminal } x]),$   
 $\quad (\text{Nonterminal } S, [\text{Nonterminal } S, \text{Terminal plus}, \text{Nonterminal } S])$   
 $]$

**definition**  $\text{start-symbol} :: s$  **where**  
 $\text{start-symbol} = \text{Nonterminal } S$

**definition**  $\mathcal{G} :: s \text{ cfg}$  **where**  
 $\mathcal{G} = \text{CFG nonterminals terminals rules start-symbol}$

**definition**  $\omega :: s \text{ list}$  **where**  
 $\omega = [\text{Terminal } x, \text{Terminal plus}, \text{Terminal } x, \text{Terminal plus}, \text{Terminal } x]$

**lemma**  $\text{wf-}\mathcal{G}$ :  
**shows**  $\text{wf-}\mathcal{G} \mathcal{G}$

**lemma**  $\text{is-sentence-}\omega$ :  
**shows**  $\text{is-sentence } \mathcal{G} \omega$

**lemma**  $\text{nonempty-derives}$ :

**shows** *nonempty-derives*  $\mathcal{G}$

**lemma** *correctness*:

**shows** *recognizing-list* (*Earley-list*  $\mathcal{G}$   $\omega$ )  $\mathcal{G} \omega \longleftrightarrow \mathcal{G} \vdash [\mathfrak{S} \mathcal{G}] \Rightarrow^* \omega$



# 7 Conclusion

## 7.1 Summary

## 7.2 Future Work

Different approaches:

(1) SPPF style parse trees as in Scott et al -> need Imperative/HOL for this

Performance improvements:

(1) Look-ahead of  $k$  or at least 1 like in the original Earley paper. (2) Optimize the representation of the grammar instead of single list, group by production, ... (3) Complete faster by keeping a map from nonterminal which are next in the items to the actual items (4) Predict faster by organizing the grammar in an efficient manner by nonterminal (5) Refine the algorithm to an imperative version using a single linked list and actual pointers instead of natural numbers.

Parse tree disambiguation:

Parser generators like YACC resolve ambiguities in context-free grammars by allowing the user to specify precedence and associativity declarations restricting the set of allowed parses. But they do not handle all grammatical restrictions, like 'dangling else' or interactions between binary operators and functional 'if'-expressions.

Grammar rewriting:

Adams *et al* [Adams:2017] describe a grammar rewriting approach reinterpreting CFGs as the tree automata, intersectioning them with tree automata encoding desired restrictions and reinterpreting the results back into CFGs.

Afroozeh *et al* [Afroozeh:2013] present an approach to specifying operator precedence based on declarative disambiguation rules basing their implementation on grammar rewriting.

Thorup [Thorup:1996] develops two concrete algorithms for disambiguation of grammars based on the idea of excluding a certain set of forbidden sub-parse trees.

Parse tree filtering:

Klint *et al* [Klint:1997] propose a framework of filters to describe and compare a wide range of disambiguation problems in a parser-independent way. A filter is a function that selects from a set of parse trees the intended trees.