

# On the Intersection of Context-Free and Regular Languages

Clemente Pasti<sup>¶,§</sup> Andreas Opedal<sup>§</sup> Tiago Pimentel<sup>¶</sup>

Tim Vieira<sup>¶</sup> Jason Eisner<sup>¶</sup> Ryan Cotterell<sup>§</sup>


<sup>¶</sup>Università della Svizzera Italiana <sup>§</sup>ETH Zürich

<sup>¶</sup>University of Cambridge <sup>¶</sup>Johns Hopkins University

clemente.pasti@usi.ch andreas.opedal@inf.ethz.ch tp472@cam.ac.uk  
tim.f.vieira@gmail.com jason@cs.jhu.edu ryan.cotterell@inf.ethz.ch

## Abstract

The Bar-Hillel construction is a classic result in formal language theory. It shows, by a simple construction, that the intersection of a context-free language and a regular language is itself context-free. In the construction, the regular language is specified by a finite-state automaton. However, neither the original construction (Bar-Hillel et al., 1961) nor its weighted extension (Nederhof and Satta, 2003) can handle finite-state automata with  $\varepsilon$ -arcs. While it is possible to remove  $\varepsilon$ -arcs from a finite-state automaton efficiently without modifying the language, such an operation modifies the automaton’s set of paths. We give a construction that generalizes the Bar-Hillel in the case where the desired automaton has  $\varepsilon$ -arcs, and further prove that our generalized construction leads to a grammar that encodes the structure of both the input automaton and grammar while retaining the asymptotic size of the original construction.

 <https://github.com/rycolab/bar-hillel>

## 1 Introduction

Bar-Hillel et al.’s (1961) construction—together with its weighted generalization (Nederhof and Satta, 2003)—is a fundamental result in formal language theory. Given a weighted context-free grammar (WCFG)  $\mathcal{G}$  and a weighted finite-state automaton (WSFA)  $\mathcal{A}$ , the Bar-Hillel construction yields another WCFG  $\mathcal{G}_\cap$  whose language  $L(\mathcal{G}_\cap)$  is equal to the intersection of  $L(\mathcal{G})$  with  $L(\mathcal{A})$ . Importantly, the Bar-Hillel construction directly proves that weighted context-free languages are closed under intersection with weighted regular languages. The construction was later extended to other formalisms, e.g., tree automata (Maletti and Satta, 2009), synchronous tree substitution grammars (Maletti, 2010) and linear context-free re-writing systems (Seki et al., 1991; Nederhof and Satta, 2011b). Furthermore, the Bar-Hillel construction has seen applications in the computation of infix

probabilities (Nederhof and Satta, 2011a) and human sentence comprehension (Levy, 2008, 2011).

Unfortunately, Bar-Hillel et al.’s construction, as well as its weighted generalization by Nederhof and Satta, requires the input automaton to be  $\varepsilon$ -free.<sup>1</sup> Although any WFSFA can be converted to a weakly equivalent<sup>2</sup>  $\varepsilon$ -free WFSFA using well-known techniques (Mohri, 2001, 2002; Hanneforth and de la Higuera, 2010), such an approach adds an additional step of computation, typically increases the size of the output grammar  $\mathcal{G}_\cap$ , and does not, in general, maintain a bijection between derivations in  $\mathcal{G}_\cap$  and the Cartesian product of the derivations in  $\mathcal{G}$  and paths in  $\mathcal{A}$ . In other words,  $\mathcal{G}_\cap$  is *not* strongly equivalent to the product of  $\mathcal{G}$  and  $\mathcal{A}$ .<sup>3</sup>

In this note, we generalize the classical Bar-Hillel construction to the case where the automaton we seek to intersect with the grammar has  $\varepsilon$ -arcs. Our new construction produces a WCFG  $\mathcal{G}_\cap$  that is strongly equivalent to the product of  $\mathcal{G}$  and  $\mathcal{A}$ . We further generalize the Bar-Hillel construction to work with arbitrary commutative semirings. Finally, we give an asymptotic bound on the size of the resulting grammar and a detailed proof of correctness in the appendix.

## 2 Languages, Automata, and Grammars

As background, we now give formal definitions of semirings, weighted formal languages, finite-state automata, and context-free grammars.

### 2.1 Semirings

Semirings are useful algebraic structures for describing weighted languages (Droste et al., 2009, Chapter 1). In order to define semirings we must first give the definition of a monoid. A **monoid** is a 3-tuple  $\mathcal{M} = (A, \bullet, \mathbf{1})$ , where  $A$  is a set,

<sup>1</sup>But they do not require the input grammar to be  $\varepsilon$ -free.

<sup>2</sup>Two WFSAs are said to be weakly equivalent if they represent the same weighted formal language.

<sup>3</sup>Strong equivalence is formally defined in Definition 6 and Theorem 1.

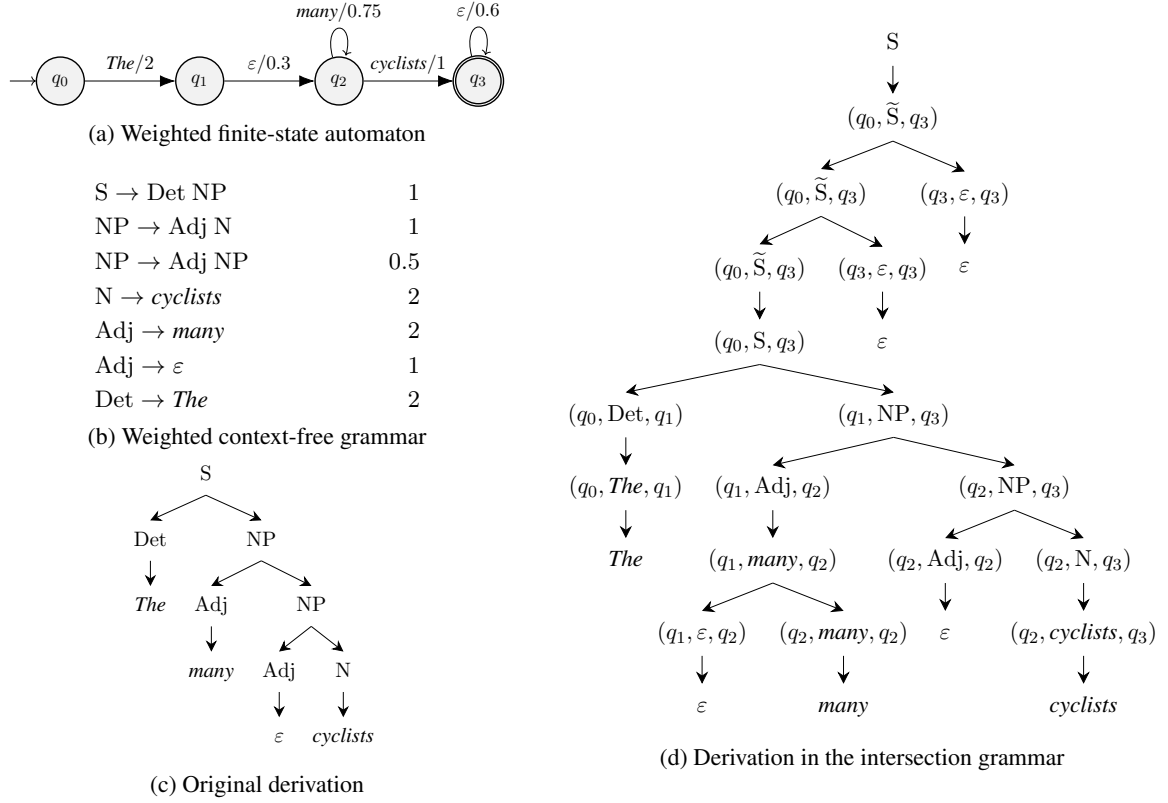


Figure 1: Example of a derivation in the grammar obtained as the intersection of the finite-state automaton (a) and the context-free grammar (b). The derivation tree (d) encodes the derivation tree (c) in the original grammar, and path  $q_0 \xrightarrow{\text{The}/2} q_1 \xrightarrow{\varepsilon/0.3} q_2 \xrightarrow{\text{many}/0.75} q_2 \xrightarrow{\text{cyclists}/1} q_3 \xrightarrow{\varepsilon/0.6} q_3 \xrightarrow{\varepsilon/0.6} q_3$ . We use rules from Eq. (5g) for  $\varepsilon$ -arcs appearing before an input symbol, and rules from Eq. (5b) for  $\varepsilon$ -arcs appearing at the end of the input.

$\bullet : A \times A \rightarrow A$  is an associative operator, and  $\mathbf{1} \in A$  is a distinguished identity element such that  $\mathbf{1} \bullet w = w \bullet \mathbf{1} = w$  for any  $w \in A$ . We say that a monoid is *commutative* if  $\bullet$  commutes, i.e.,  $w_1 \bullet w_2 = w_2 \bullet w_1$  for any  $w_1, w_2 \in A$ . We can now give the definition of a semiring.

**Definition 1.** A *semiring*  $\mathcal{W} = (A, \oplus, \otimes, \mathbf{0}, \mathbf{1})$  is a 5-tuple where  $(A, \oplus, \mathbf{0})$  is a commutative monoid,  $(A, \otimes, \mathbf{1})$  is a monoid,  $\otimes$  distributes over  $\oplus$ , and  $\mathbf{0}$  is an annihilator for  $\otimes$ , meaning that  $\mathbf{0} \otimes w = w \otimes \mathbf{0} = \mathbf{0}$  for any  $w \in A$ .

We say that  $\mathcal{W}$  is commutative if  $\otimes$  commutes. In this work, we assume commutative semirings.

## 2.2 Weighted Formal Languages

This paper concerns itself with transforms between devices that generate weighted formal languages.

**Definition 2.** Let  $\Sigma$  be an alphabet and  $\mathcal{W} = (A, \oplus, \otimes, \mathbf{0}, \mathbf{1})$  be a semiring. Then a *weighted formal language*  $L : \Sigma^* \rightarrow A$  is a mapping from the Kleene closure of  $\Sigma$  to the set of weights  $A$ . Furthermore, the set  $\text{supp}(L) = \{s \in \Sigma^* \mid L(s) \neq \mathbf{0}\}$  is called the language's *support*.

Unweighted formal languages (e.g., Sipser, 2006; Hopcroft et al., 2006) are simply the special case of Definition 2 where  $\mathcal{W}$  is the boolean semiring. In this note, we discuss algorithms for computing the intersection of two weighted formal languages.<sup>4</sup>

**Definition 3.** Let  $L_1$  and  $L_2$  be two weighted formal languages over the same alphabet  $\Sigma$  and the same semiring  $\mathcal{W}$ . The *intersection* of  $L_1$  with  $L_2$  is defined as the weighted language

$$(L_1 \cap L_2)(s) \stackrel{\text{def}}{=} L_1(s) \otimes L_2(s), \quad \forall s \in \Sigma^* \quad (1)$$

Specifically, this paper concerns itself with the special case of Definition 3 when  $L_1$  is a weighted context-free language (represented by a WCFG), and  $L_2$  is a weighted regular language (represented by a WFSA); we define these two formalisms in the subsequent sections.

In the following, the symbol  $\varepsilon$  always represents the empty string.

<sup>4</sup>The intersection of two weighted languages is also called their Hadamard product (Droste et al., 2009, Chapter 1).

## 2.3 Weighted Finite-State Automata

We now review the basics of weighted finite-state automata (WFSA), which provide a formalism to represent weighted regular languages.

**Definition 4.** A *weighted finite-state automaton*  $\mathcal{A}$  over a semiring  $\mathcal{W} = (A, \oplus, \otimes, \mathbf{0}, \mathbf{1})$  is a 6-tuple  $(\Sigma, Q, \delta, \lambda, \rho, \mathcal{W})$ . In this tuple,  $\Sigma$  is an alphabet,  $Q$  is a finite set of states, and  $\delta \subseteq Q \times Q \times (\Sigma \cup \{\varepsilon\}) \times A$  is a finite multi-set of weighted arcs. Further,  $\lambda : Q \rightarrow A$  and  $\rho : Q \rightarrow A$  are the initial and final weight functions, respectively. We also define the sets  $Q_I = \{q \mid q \in Q, \lambda(q) \neq \mathbf{0}\}$  and  $Q_F = \{q \mid q \in Q, \rho(q) \neq \mathbf{0}\}$  for convenience.

We will represent an arc in  $\delta$  with the notation  $q_0 \xrightarrow{a/w} q_1$  where  $a \in \Sigma \cup \{\varepsilon\}$  and  $w \in A$ . A **path**  $\pi$  (of length  $N > 0$ ) is a sequence of arcs in  $\delta^*$  where the states of adjacent arcs are matched, i.e.,

$$q_0 \xrightarrow{a_1/w_1} \cdots q_{n-1} \xrightarrow{a_n/w_n} q_n \cdots \xrightarrow{a_N/w_N} q_N \quad (2)$$

and where  $q_0 \in Q_I$  and  $q_N \in Q_F$ , i.e., the path starts at an initial state and ends at a final state. The path's **yield**, denoted  $\text{yield}(\pi)$ , is the concatenation  $a_1 a_2 \cdots a_N$  of all its arc labels (strings of length  $\leq 1$ ). The path's **weight**, denoted  $w(\pi)$ , is the product

$$w(\pi) = \lambda(q_0) \otimes \left( \bigotimes_{n=1}^N w_n \right) \otimes \rho(q_N) \quad (3)$$

We denote the set of all paths in  $\mathcal{A}$  as  $\mathcal{D}_{\mathcal{A}}$ , and the set of all paths with yield  $s$  as  $\mathcal{D}_{\mathcal{A}}(s)$ . Finally, we define the **language of an automaton** as the mapping  $L_{\mathcal{A}} : \Sigma^* \rightarrow A$  where we have<sup>5</sup>  $L_{\mathcal{A}}(s) = \bigoplus_{\pi \in \mathcal{D}_{\mathcal{A}}(s)} w(\pi)$ . The set of languages that can be encoded by a WFSA forms the class of **weighted regular languages**.

## 2.4 Weighted Context-Free Grammars

We now go over the necessary background on weighted context-free grammars (WCFGs).

**Definition 5.** A *weighted context-free grammar* is a tuple  $\mathcal{G} = (\mathcal{N}, \Sigma, \mathcal{W}, S, \mathcal{P})$ , where  $\mathcal{N}$  is a non-empty set of nonterminal symbols,  $\Sigma$  is an alphabet of terminal symbols,  $\mathcal{W} = (A, \oplus, \otimes, \mathbf{0}, \mathbf{1})$  is a semiring,  $S \in \mathcal{N}$  is a distinguished start symbol, and  $\mathcal{P}$  is a set of production rules. Each rule  $p \in \mathcal{P}$  is of the form  $X \xrightarrow{w} \alpha$ , with  $X \in \mathcal{N}$ ,  $w \in A$ , and  $\alpha \in (\Sigma \cup \mathcal{N})^*$ .

<sup>5</sup>In the main paper we gloss over the question of how  $\bigoplus$ -summations over infinite sets are to be defined (or left undefined), but we treat this issue in App. B.2.

Given two strings  $\alpha, \beta \in (\Sigma \cup \mathcal{N})^*$ , we write  $\alpha \xrightarrow{p}_L \beta$  if and only if we can express  $\alpha = z X \delta$  and  $\beta = z \gamma \delta$  where  $z \in \Sigma^*$  and  $p \in \mathcal{P}$  is the rule  $X \xrightarrow{w} \gamma$ . A **derivation**  $d$  (more precisely, a leftmost derivation) is a sequence  $\alpha_0, \dots, \alpha_N$  with  $N > 0$ ,  $\alpha_0 = S$ , and  $\alpha_N \in \Sigma^*$ , such that for all  $0 < n \leq N$ , we have  $\alpha_{n-1} \xrightarrow{p_n}_L \alpha_n$  for some (necessarily unique)  $p_n \in \mathcal{P}$ . The derivation's **yield**,  $\text{yield}(d)$ , is  $\alpha_N$ , and its weight,  $w(d)$ , is  $w(p_1) \otimes \cdots \otimes w(p_N)$ . We denote the set of derivations under a grammar  $\mathcal{G}$  as  $\mathcal{D}_{\mathcal{G}}$  and the set of all derivations with yield  $s$  as  $\mathcal{D}_{\mathcal{G}}(s)$ . Finally, we define the **language of a grammar** as  $L_{\mathcal{G}}$  where<sup>5</sup>  $L_{\mathcal{G}}(s) \stackrel{\text{def}}{=} \bigoplus_{d \in \mathcal{D}_{\mathcal{G}}(s)} w(d), \forall s \in \Sigma^*$ . The languages that can be encoded by a WCFG are known as **weighted context-free languages**.

## 3 Generalizing Bar-Hillel

Given any context-free grammar (CFG)  $\mathcal{G}$  and finite-state automaton (FSA)  $\mathcal{A}$ , Bar-Hillel et al. (1961) showed how to construct a CFG  $\mathcal{G}_{\cap}$  such that  $L_{\mathcal{G}_{\cap}} = L_{\mathcal{G}} \cap L_{\mathcal{A}}$ . Later, Nederhof and Satta (2003) generalized Bar-Hillel's construction to work on a *weighted* context-free grammar and a *weighted* finite-state automaton. While they focused on the real semiring, their construction actually works for any commutative semiring. However, neither of these versions correctly computes the intersection when the WFSA (or FSA) contains  $\varepsilon$ -arcs. Yet, in several applications—such as modeling noisy inputs for human sentence comprehension (Levy, 2008, 2011)—we may be interested in using a WFSA  $\mathcal{A}$  that contains  $\varepsilon$ -arcs. A naïve application of the construction would ignore paths in  $\mathcal{A}$  that contain  $\varepsilon$ -arcs. The problem may be sidestepped by transforming  $\mathcal{A}$  into a weakly equivalent  $\varepsilon$ -free WFSA<sup>6</sup> before applying the construction;<sup>7</sup> this, however, might increase the size of the WFSA and of the intersection grammar, and it would not allow us to identify the paths in the input WFSA that yield a target string in the

<sup>6</sup>See footnote 2 for the definition of weak equivalence.

<sup>7</sup>Levy (2008, 2011) uses WSAs to model the degree of uncertainty under which a human comprehends a particular sentence, in which  $\varepsilon$ -arcs are used to represent word deletion. He applies the Bar-Hillel construction to compute the intersection of the language represented by the WFSA and the language encoded by a WCFG that represents the comprehender's grammatical knowledge, in order to obtain a joint posterior distribution over parses and words. While he transforms  $\mathcal{A}$  to eliminate  $\varepsilon$ -arcs prior to applying the Bar-Hillel construction (Levy, p.c.), the solution we propose here is an alternative.

intersection grammar.<sup>8</sup>

### 3.1 The problem with $\varepsilon$ -arcs

Before proposing our solution, we explain how the original construction works, and how it fails in the case of  $\varepsilon$ -arcs. Given a WFSA  $\mathcal{A} = (\Sigma, Q, \delta, \lambda, \rho, \mathcal{W})$  and a WCFG  $\mathcal{G} = (\mathcal{N}, \Sigma, \mathcal{W}, \mathcal{S}, \mathcal{P})$  over the same alphabet  $\Sigma$  and commutative semiring  $\mathcal{W}$ , their intersection  $\mathcal{G}_\cap$  is defined by the tuple  $(\mathcal{N}_\cap, \Sigma, \mathcal{W}, \mathcal{S}, \mathcal{P}_\cap)$ , where:

- The set of nonterminal symbols  $\mathcal{N}_\cap = \{\mathcal{S}\} \cup Q \times (\mathcal{N} \cup \Sigma) \times Q$  contains the triplets  $(q_i, X, q_j)$  plus the start symbol  $\mathcal{S}$ .<sup>9</sup>
- The set of production rules  $\mathcal{P}_\cap$  is given by the equations in Construction 1 of Fig. 2.<sup>10</sup>
- $\Sigma, \mathcal{W}, \mathcal{S}$  are the same as in the input grammar.

The intuition behind this construction is that a derivation in the intersection grammar encodes both a path in the input WFSA and a derivation in the input WCFG with matching yield. Specifically, rules (4f) encode arcs in the WFSA and rules (4d) encode production rules in the WCFG. Rules (4e) handle the special case of  $\varepsilon$ -productions in the input WCFG and rules (4a) are designed to take into account the initial and final weight of a path. These rules may combine through matching nonterminals to permit derivations in the intersection grammar  $\mathcal{G}_\cap$ .

Unfortunately, this mechanism breaks in the presence of  $\varepsilon$ -arcs. Although the rules (4f) do construct nonterminals for  $\varepsilon$ -arcs (when  $a = \varepsilon$ ), the rules (4d) never generate those nonterminals (since the  $X_m$  on the right-hand side of a rule are never  $\varepsilon$ ). We show this with an example. Consider the automaton and the grammar in Fig. 1, both of which assign non-zero weight to the string *The many cyclists*. However, their intersection computed with the Bar-Hillel construction is empty. To see this, note that all the paths from  $q_0$  to  $q_3$  contain the arc  $q_1 \xrightarrow{\varepsilon/0.3} q_2$ . Eq. (4f) will create a rule

<sup>8</sup>In contrast, this is easy under our construction. Each derivation of the target string under  $\mathcal{G}_\cap$  uses a particular path in  $\mathcal{A}$ . To reconstruct that path,  $\varepsilon$ -arcs and all, simply traverse from left to right the leaves of the derivation tree (e.g., Fig. 1d) and list the states on the triplets where rule (5f) is applied.

<sup>9</sup>Many of the nonterminals will turn out to be **useless** in that they do not participate in any derivation in  $\mathcal{D}_{\mathcal{G}_\cap}$ . These can be pruned from the grammar along with all rules that mention them (Hopcroft et al., 2006).

<sup>10</sup>Note that this construction can handle multiple initial and final states, whereas Nederhof and Satta's (2003) construction assumes a WFSA with a single initial and a single final state. A path's initial and final weights are taken into account by the weight of rules (4a) of Construction 1 in Fig. 2.

$(q_1, \varepsilon, q_2) \xrightarrow{0.3} \varepsilon$ , but none of the rules produced by Eqs. (4d) and (4e) has the triplet  $(q_1, \varepsilon, q_2)$  on the right hand side. This misalignment results in an empty set of derivations in  $\mathcal{G}_\cap$ . In App. A we describe more failure cases in a detailed manner.

### 3.2 Our generalized construction

We now describe an improved version of the Bar-Hillel construction that handles  $\varepsilon$ -arcs in the WFSA. In comparison to the original construction, our version of  $\mathcal{G}_\cap = (\mathcal{N}_\cap, \Sigma, \mathcal{W}, \mathcal{S}, \mathcal{P}_\cap)$  has

- $\mathcal{N}_\cap = \{\mathcal{S}\} \cup Q \times (\mathcal{N} \cup \{\tilde{\mathcal{S}}\} \cup \Sigma) \times Q$  as the set of nonterminals, where  $\tilde{\mathcal{S}}$  is a new symbol;
- $\mathcal{P}_\cap$  as the augmented set of production rules given in Construction 2 of Fig. 2.

Our generalized construction adds additional production rules that traverse the  $\varepsilon$ -arcs. Rules (5g) can traverse a WFSA subpath labeled with  $\varepsilon^*a$  to yield a terminal symbol  $a \in \Sigma$ . At the end of the yielded string, rules (5b) can traverse a WFSA subpath labeled with  $\varepsilon^*$  that ends at a final state  $q_F$ . Our construction carefully avoids overcounting<sup>11</sup> by ensuring that each matching pair of an  $\mathcal{A}$ -path and a  $\mathcal{G}$ -derivation of its string corresponds to *exactly one*  $\mathcal{G}_\cap$ -derivation of that string, as illustrated in Fig. 1. Note that rules (5d) to (5f) are identical to their counterparts in the original construction. Rules (5a) are a modified version of rules (4a) with the special start symbol  $\tilde{\mathcal{S}}$ ; this allows our construction to handle  $\varepsilon$ -arcs immediately before the final state—by repeated applications of rule (5b)—before switching  $\tilde{\mathcal{S}}$  back to  $\mathcal{S}$  with rule (5c). In App. A we illustrate the mechanism with examples.

We now state the theorem of correctness.

**Definition 6.** *Let  $\Sigma$  be an alphabet and  $\mathcal{W}$  be a commutative semiring. Let  $\mathcal{G}$  be a WCFG and  $\mathcal{A}$  be a WFSA—both over  $\Sigma$  and  $\mathcal{W}$ . The **weighted join** of the derivations in  $\mathcal{D}_{\mathcal{G}}$  with the paths in  $\mathcal{D}_{\mathcal{A}}$  is defined as:*

$$(\mathcal{D}_{\mathcal{G}} \bowtie \mathcal{D}_{\mathcal{A}}) \stackrel{\text{def}}{=} \left\{ \langle \mathbf{d}, \boldsymbol{\pi} \rangle \mid \mathbf{d} \in \mathcal{D}_{\mathcal{G}}, \boldsymbol{\pi} \in \mathcal{D}_{\mathcal{A}} \right. \\ \left. \text{s.t. } \text{yield}(\mathbf{d}) = \text{yield}(\boldsymbol{\pi}) \right\} \quad (6)$$

with  $w(\langle \mathbf{d}, \boldsymbol{\pi} \rangle) = w(\mathbf{d}) \otimes w(\boldsymbol{\pi})$ .

<sup>11</sup>As Fig. 1d illustrates, we do this by introducing a single, right-branching subderivation for each  $\mathcal{A}$ -subpath  $\varepsilon^*a$  that matches an input symbol  $a$ . A nonterminal of the form  $(q_0, \varepsilon, q_1)$  is never used as a right child, nor does it ever combine with a nonterminal of the form  $(q_0, X, q_1)$ , except at the end of the input, which is specially handled by rules (5b). Similarly, Allauzen et al. (2010) avoid overcounting when intersecting or composing finite-state machines that have  $\varepsilon$ -arcs.

### Construction 1

$$S \xrightarrow{\lambda(q_I) \otimes \rho(q_F)} (q_I, S, q_F) \quad (4a)$$

$$\forall q_I \in I, \forall q_F \in F$$

$$(q_0, X, q_M) \xrightarrow{w} (q_0, X_1, q_1) \cdots (q_{M-1}, X_M, q_M)$$

$$\forall (X \xrightarrow{w} X_1 \cdots X_M) \in \mathcal{P}, M > 0 \quad (4d)$$

$$\forall q_0, \dots, q_M \in Q$$

$$(q_0, X, q_0) \xrightarrow{w} \varepsilon \quad (4e)$$

$$\forall (X \xrightarrow{w} \varepsilon) \in \mathcal{P}, \forall q_0 \in Q$$

$$(q_0, a, q_1) \xrightarrow{w} a \quad (4f)$$

$$\forall (q_0 \xrightarrow{a/w} q_1) \in \delta$$

### Construction 2

$$S \xrightarrow{\lambda(q_I) \otimes \rho(q_F)} (q_I, \tilde{S}, q_F) \quad (5a)$$

$$\forall q_I \in I, \forall q_F \in F$$

$$(q_I, \tilde{S}, q_1) \xrightarrow{\mathbf{1}} (q_I, \tilde{S}, q_0)(q_0, \varepsilon, q_1) \quad (5b)$$

$$\forall q_I \in I, \forall q_0, q_1 \in Q$$

$$(q_I, \tilde{S}, q_0) \xrightarrow{\mathbf{1}} (q_I, S, q_0) \quad (5c)$$

$$\forall q_I \in I, \forall q_0 \in Q$$

$$(q_0, X, q_M) \xrightarrow{w} (q_0, X_1, q_1) \cdots (q_{M-1}, X_M, q_M)$$

$$\forall (X \xrightarrow{w} X_1 \cdots X_M) \in \mathcal{P}, M > 0 \quad (5d)$$

$$\forall q_0, \dots, q_M \in Q$$

$$(q_0, X, q_0) \xrightarrow{w} \varepsilon \quad (5e)$$

$$\forall (X \xrightarrow{w} \varepsilon) \in \mathcal{P}, \forall q_0 \in Q$$

$$(q_0, a, q_1) \xrightarrow{w} a \quad (5f)$$

$$\forall (q_0 \xrightarrow{a/w} q_1) \in \delta, a \in \Sigma \cup \{\varepsilon\}$$

$$(q_0, a, q_2) \xrightarrow{\mathbf{1}} (q_0, \varepsilon, q_1)(q_1, a, q_2) \quad (5g)$$

$$\forall a \in \Sigma, \forall q_0, q_1, q_2 \in Q$$

Figure 2: The original Bar-Hillel construction (left) and our generalized version (right) that covers  $\varepsilon$ -arcs. We highlight the differences from the original construction in red. Note that the weights of rules (4a) and (4f) (respectively (5a) and (5f)) encode the weights of the WFSAs, while the weights of rules (4d) and (4e) (respectively (5d) and (5e)) encode weights of the WCFG. All other rules in the generalized construction ((5b), (5c) and (5g)) are assigned weight  $\mathbf{1}$ , and, thus, they do not change the weight of a derivation.

**Theorem 1.** *Let  $\mathcal{G}$  be a WCFG and  $\mathcal{A}$  a WFSAs over the same alphabet  $\Sigma$  and commutative semiring  $\mathcal{W}$ . Let  $\mathcal{G}_\cap$  be the grammar obtained with our generalized construction. Then we have strong equivalence between  $\mathcal{G}_\cap$  and  $\langle \mathcal{G}, \mathcal{A} \rangle$ ; meaning that there is a weight-preserving, yield-preserving bijection between  $\mathcal{D}_{\mathcal{G}_\cap}$  and  $(\mathcal{D}_{\mathcal{G}} \bowtie \mathcal{D}_{\mathcal{A}})$ .*

**Corollary 1.**  *$\mathcal{G}_\cap$  and  $\langle \mathcal{G}, \mathcal{A} \rangle$  are weakly equivalent, meaning that  $L_{\mathcal{G}_\cap}(s) = L_{\mathcal{G}}(s) \otimes L_{\mathcal{A}}(s)$  whenever the values on the right-hand side are defined.*

See App. B for proofs. Theorem 1 may be seen as a generalization of Theorem 8.1 by Bar-Hillel et al. (1961) and Theorem 12 by Nederhof and Satta (2003). Indeed, the set of derivations produced by Construction 1 is equivalent to the set of derivations produced by Construction 2, modulo an unfold transform (Tamaki and Sato, 1984) to remove rules containing  $\tilde{S}$ . Among the groups of rules listed in Fig. 2, the set of rules with maximum cardinality is the one defined by Eq. (5d). This set has cardinality  $\mathcal{O}(|\mathcal{P}||Q|^{M_\star})$ , where  $M_\star$  is 1 plus

the length of the longest right-hand side among all the rules  $\mathcal{P}$ . All other equations in this construction lead to smaller sets of added rules. Since Eq. (5d) is unchanged from Eq. (4d) in the original construction, the asymptotic bound on the number of rules in our output grammar remains unchanged.

## 4 Conclusion

We generalized the weighted Bar-Hillel intersection construction so that the given WFSAs may contain  $\varepsilon$ -arcs. Our construction is strongly equivalent to the product of the original WCFG and WFSAs, i.e., every derivation tree in the resulting grammar represents a pairing of a derivation tree in the input WCFG and a path in the WFSAs with the same yield. We gave a full proof of correctness for our construction. By adding output strings to the WFSAs arcs and having rule (5f) rewrite to the arc's output string, our method can also be used to compose a WCFG with a weighted finite-state transducer (WFST) that could usefully model morphological post-processing or speaker errors.

## 5 Acknowledgements

The authors acknowledge Roger Levy for correspondence about Levy (2008) and Levy (2011).

## 6 Limitations

In this note, we generalize a fundamental theoretical result in formal language theory, which has seen a variety of practical applications, including human sentence comprehension under uncertain input (Levy, 2008, 2011) and infix probability computation (Nederhof and Satta, 2003). Although we motivate our paper by discussing the necessity of performing intersections on automata with  $\varepsilon$ -arcs, we do not explore any such practical applications. Further, while we show that the asymptotic bound on the size of our intersection grammar matches the original Bar-Hillel construction's, we do not discuss multiplicative or added constants introduced in our grammar's size.

## Ethical Statement

We do not foresee any ethical issues with our work.

## References

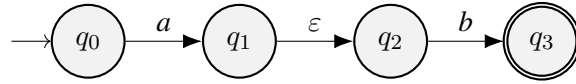
- Cyril Allauzen, Michael Riley, and Johan Schalkwyk. 2010. [Filters for efficient composition of weighted finite-state transducers](#). In *Proceedings of the 15th International Conference on Implementation and Application of Automata*, International Conference on Implementation and Application of Automata, page 28–38, Berlin, Heidelberg. Springer-Verlag.
- Yehoshua Bar-Hillel, M. Perles, and E. Shamir. 1961. [On formal properties of simple phrase structure grammars](#). *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 14:143–172. Reprinted in Y. Bar-Hillel. (1964). *Language and Information: Selected Essays on their Theory and Application*, Addison-Wesley 1964, 116–150.
- Manfred Droste, Werner Kuich, and Heiko Vogler. 2009. *Handbook of Weighted Automata*. Springer Berlin, Heidelberg.
- Thomas Hanneforth and Colin de la Higuera. 2010.  [\$\varepsilon\$ -removal by loop reduction for finite-state automata](#). In *Language and Logos*, pages 297–312, Berlin. Akademie Verlag.
- John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. 2006. *Introduction to Automata Theory, Languages, and Computation*, 3 edition. Addison-Wesley Longman Publishing Co., Inc., USA.
- Liang Huang. 2008. [Advanced dynamic programming in semiring and hypergraph frameworks](#). In *Coling 2008: Advanced Dynamic Programming in Computational Linguistics: Theory, Algorithms and Applications - Tutorial notes*, pages 1–18, Manchester, UK. Coling 2008 Organizing Committee.
- Roger Levy. 2008. [A noisy-channel model of human sentence comprehension under uncertain input](#). In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 234–243, Honolulu, Hawaii. Association for Computational Linguistics.
- Roger Levy. 2011. [Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1055–1065, Portland, Oregon, USA. Association for Computational Linguistics.
- Andreas Maletti. 2010. [Why synchronous tree substitution grammars?](#) In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 876–884, Los Angeles, California. Association for Computational Linguistics.
- Andreas Maletti and Giorgio Satta. 2009. [Parsing algorithms based on tree automata](#). In *Proceedings of the 11th International Conference on Parsing Technologies*, pages 1–12, Paris, France. Association for Computational Linguistics.
- Mehryar Mohri. 2001. [Generic  \$\varepsilon\$ -removal algorithm for weighted automata](#). In *Implementation and Application of Automata*, pages 230–242, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Mehryar Mohri. 2002. [Semiring frameworks and algorithms for shortest-distance problems](#). *Journal of Automata, Languages and Combinatorics*, 7(3):321–350.
- Mark-Jan Nederhof and Giorgio Satta. 2003. [Probabilistic parsing as intersection](#). In *Proceedings of the Eighth International Conference on Parsing Technologies*, pages 137–148, Nancy, France.
- Mark-Jan Nederhof and Giorgio Satta. 2011a. [Computation of infix probabilities for probabilistic context-free grammars](#). In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1213–1221, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Mark-Jan Nederhof and Giorgio Satta. 2011b. [Prefix probabilities for linear context-free rewriting systems](#). In *Proceedings of the 12th International Conference on Parsing Technologies*, pages 151–162, Dublin, Ireland. Association for Computational Linguistics.
- Hiroyuki Seki, Takashi Matsumura, Mamoru Fujii, and Tadao Kasami. 1991. [On multiple context-free grammars](#). *Theoretical Computer Science*, 88(2):191–229.

Michael Sipser. 2006. *Introduction to the Theory of Computation*, 2 edition. Thomson Course Technology.

Hisao Tamaki and Taisuke Sato. 1984. *Unfold/fold transformation of logic programs*. In *Proceedings of the Second International Logic Programming Conference*, pages 127–138, Uppsala, Sweden. Uppsala University.

## A Failure Cases of Original Construction

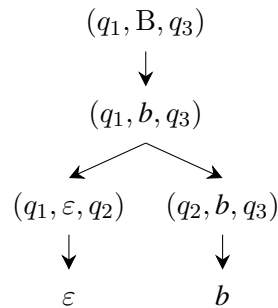
We distinguish two types of failure cases: (i)  $\text{supp}(L_{\mathcal{G}_\cap}) \neq \text{supp}(L_{\mathcal{A}}) \cap \text{supp}(L_{\mathcal{G}})$  and (ii)  $L_{\mathcal{G}_\cap} \neq L_{\mathcal{A}} \cap L_{\mathcal{G}}$ , both of which we will exemplify now. Notably, the case (ii) follows from (i), but—to be comprehensible—we will nonetheless give an example where (ii) fails without (i). For case (i), consider the following unweighted FSA:



and the following unweighted CFG:

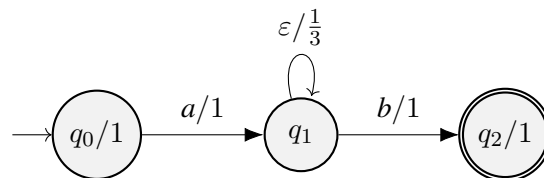
$$\begin{aligned} S &\rightarrow AB \\ A &\rightarrow a \\ B &\rightarrow b \end{aligned}$$

It is easy to see that the intersection of the language accepted by the FSA and the language generated by the CFG is  $\{ab\}$ . Construction 1, however, outputs an empty grammar (after pruning useless rules as in footnote 9) and, hence, an empty language. To see this, consider Eq. (4d) and Eq. (4f). First, Eq. (4f) will create a rule  $(q_1, \varepsilon, q_2) \rightarrow \varepsilon$ , but  $(q_1, \varepsilon, q_2)$  will be useless because it cannot be reached from any of the rules produced by Eq. (4d). Second, Eq. (4d) will produce reachable nonterminals  $(q_0, A, q_i)$  and  $(q_i, B, q_3)$ , with  $i \in \{1, 2\}$ . The case of  $i = 1$  will reach  $a$  but not  $b$ , and  $i = 2$  will reach  $b$  but not  $a$ . Let us now show how our generalized construction fixes this failure case. Eq. (5g) generates the rule  $(q_1, b, q_3) \rightarrow (q_1, \varepsilon, q_2)(q_2, b, q_3)$  which then combines with rule  $(q_1, B, q_3) \rightarrow (q_1, b, q_3)$  to form a subderivation<sup>12</sup> that covers the substring  $\varepsilon b$ , as shown in the picture below.



Note that rules generated by Eq. (5g) can only mention symbol  $\varepsilon$  in the left child, not in the right child, as discussed in footnote 11.

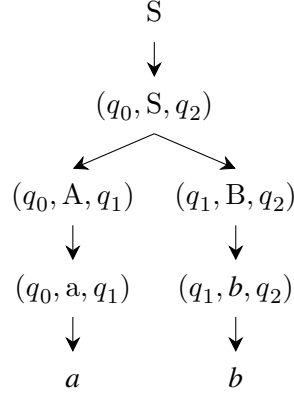
As stated above, to be comprehensive, we also show a case where only case (ii) fails, without (i). Take the following WFSA over the Inside semiring (Huang, 2008):



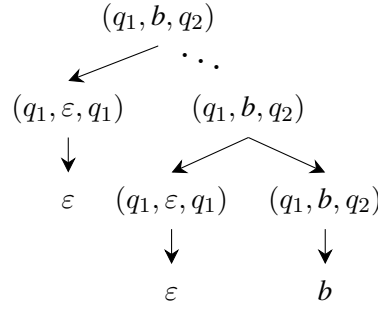
and the same grammar as above with weight 1 for all rules. It is easy to see that the language's weight for  $s = ab$  in the WFSA is a geometric series  $L_{\mathcal{A}}(s) = \sum_{i=0}^{\infty} \left(\frac{1}{3}\right)^i = \frac{3}{2}$ , while in the WCFG,  $L_{\mathcal{G}}(s) = 1$ . However, the output grammar  $\mathcal{G}_\cap$  of Construction 1 will contain one single derivation  $d$ :

<sup>12</sup>In App. B we give a formal definition of subderivation.





with  $w(\mathbf{d}) = 1$ , as all rules either stem from  $\mathcal{G}$  or from the arcs  $q_0 \xrightarrow{a/1} q_1$  and  $q_1 \xrightarrow{b/1} q_2$ . This will result in  $L_{\mathcal{G}_\cap} = 1$ , but  $L_{\mathcal{A}} \cap L_{\mathcal{G}} = \frac{3}{2}$ . This is because there are no derivations rooted at  $S$  in  $\mathcal{G}_\cap$  that match with the  $\varepsilon$ -arcs in  $\mathcal{A}$ : Similarly to the example above,  $(q_1, \varepsilon, q_1)$  will not be reachable. We will now briefly show how our construction fixes this failure case as well. Note that there are infinitely many paths in the WFSA with yield  $s = ab$ ; but there is also only a single derivation in  $\mathcal{D}_{\mathcal{G}}$  with this yield. Our construction thus ensures that there is exactly one derivation in  $\mathcal{D}_{\mathcal{G}_\cap}$  for every  $ab$  path in  $\mathcal{D}_{\mathcal{A}}$ . As the  $\varepsilon$ -loop allows unboundedly long subpaths from  $q_1$  to  $q_2$  that are labeled with  $\varepsilon^*b$ , the rules generated by Eq. (5g) will build corresponding unboundedly deep subderivations of the following form:



Finally we observe that a similar argument holds for rules generated by Eq. (5b), and  $\varepsilon$ -arcs that occur immediately before a final state.

## B Proofs

### B.1 Proof of Theorem 1

Theorem 1 gives a result for derivations (which are always rooted at  $S$ ) and paths (which always connect an initial state with a final state). However, in order to prove this theorem we must also consider subderivations and subpaths. We define subderivations as follows: a **subderivation**  $\tilde{\alpha}$  is a sequence  $\alpha_0, \dots, \alpha_N$  with  $N \geq 0$ , where (i) in the case of  $N > 0$ ,  $\alpha_0 = X$ ,  $X \in \mathcal{N}$ , and  $\alpha_N \in (\varepsilon \cup \Sigma^*)$ , such that for all  $0 < n \leq N$ , we have  $\alpha_{n-1} \xrightarrow{p_n} \alpha_n$  for some  $p_n \in \mathcal{P}$ , and (ii) in the case of  $N = 0$ ,  $\alpha_0 \in \Sigma \cup \{\varepsilon\}$ . The weight and yield of subderivations are defined analogously to that of derivations. In the extended case of  $N = 0$ , the yield is equal to  $\alpha_0$  and the weight is set to  $\mathbf{1}$ . We will say that a subderivation is **rooted** at  $X$  if  $\alpha_0 = X$ . We denote the set of subderivations rooted at  $X$  with  $\mathcal{D}_{\mathcal{G}}(X)$ . Moreover, a subpath is defined as follows: A **subpath**  $\tilde{\pi}$  (of length  $N \geq 0$ ), is (i) in the case of  $N > 0$ , a sequence of arcs in  $\delta^*$  where the states of adjacent arcs are matched, and (ii) in the case of  $N = 0$  a single state  $q \in Q$ .<sup>13</sup> The subpath's weight, denoted  $\tilde{w}(\tilde{\pi})$ , is the product  $\tilde{w}(\tilde{\pi}) = \bigotimes_{n=1}^N w_n$  of the weights of the arcs along the subpath. In the extended case  $N = 0$  we set the weight to  $\mathbf{1}$  and the yield to  $\varepsilon$ . Note that, in contrast to the

<sup>13</sup>We note the difference to paths defined in §2.3: a subpath does not need to start in an initial state and end in a final state.

weight of a path, the weight of a subpath does not account for initial and final weights. The yield is defined identically to that of paths. We denote the set of all paths starting at  $q_i$  and ending at  $q_j$  with  $\mathcal{D}_{\mathcal{A}}(\{q_i, q_j\})$ . Note that the definitions of subderivation and subpath encapsulate the definitions of derivation and path respectively. Furthermore, we will denote with  $p(\pi)$  and  $n(\pi)$ , respectively, the first and the last state encountered along a path.

We will now prove two lemmas that will be necessary for the proof of Theorem 1.

**Lemma 1.** *For any triplet  $(q_0, X, q_m) \in \mathcal{N}_{\cap}$ , with  $X \neq \tilde{S}$  and  $q_0, q_m \in Q$ , there is a bijection  $\psi(\tilde{\mathbf{d}}_{\cap}) = \langle \tilde{\mathbf{d}}, \tilde{\pi} \rangle$  from  $\mathcal{D}_{\mathcal{G}_{\cap}}((q_0, X, q_m))$  to the weighted join  $(\mathcal{D}_{\mathcal{G}}(X) \bowtie \mathcal{D}_{\mathcal{A}}(\{q_0, q_m\}))$ , restricted to tuples in which the path does not have an  $\varepsilon$ -arc immediately before a final state. Moreover, it holds that:*

$$w(\tilde{\mathbf{d}}_{\cap}) = w(\tilde{\mathbf{d}}) \otimes \tilde{w}(\tilde{\pi}) \quad (7)$$

$$\text{yield}(\tilde{\mathbf{d}}_{\cap}) = \text{yield}(\tilde{\mathbf{d}}) = \text{yield}(\tilde{\pi}) \quad (8)$$

*Proof.* We begin by showing that  $\psi$  is well defined, that it is injective and that it satisfies the properties in Eqs. (7) and (8). We prove this by induction on subderivations.

*Lemma 1's Base Case.* We begin by observing that the only terminal rules from  $\mathcal{P}_{\cap}$  are defined by Eq. (5f) and Eq. (5e).

*Lemma 1's Base Case, Part #1.*  $\tilde{\mathbf{d}}_{\cap}$  is obtained by the application of a single production rule  $(q_0, a, q_1) \xrightarrow{w} a$  from Eq. (5f). We define  $\psi(\tilde{\mathbf{d}}_{\cap}) = \langle \tilde{\mathbf{d}}, \tilde{\pi} \rangle$ , where  $\tilde{\pi} = q_0 \xrightarrow{a/w} q_1$  and  $\tilde{\mathbf{d}} = a$  is the subderivation that contains just the string  $a$  with weight  $\mathbf{1}$ . It is easy to see that the yield is preserved. Moreover:

$$w(\tilde{\mathbf{d}}_{\cap}) = w \quad (\text{by Eq. (5f)}) \quad (9a)$$

$$= w \otimes \mathbf{1} \quad (9b)$$

$$= \tilde{w}(\tilde{\pi}) \otimes w(\tilde{\mathbf{d}}) \quad (9c)$$

*Lemma 1's Base Case, Part #2.*  $\tilde{\mathbf{d}}_{\cap}$  is obtained by the application of a single production rule  $(q_0, X, q_0) \xrightarrow{w} \varepsilon$  from Eq. (5e). We construct  $\psi$  as follows:  $\psi(\tilde{\mathbf{d}}_{\cap}) = \langle \tilde{\mathbf{d}}, \tilde{\pi} \rangle$ , where  $\tilde{\mathbf{d}} = X \xrightarrow{p}_L \varepsilon$  with  $p = X \xrightarrow{w} \varepsilon$ , and  $\tilde{\pi}$  is the subpath  $q_0$  with weight  $\mathbf{1}$ . Clearly the yield is preserved and:

$$w(\tilde{\mathbf{d}}_{\cap}) = w \quad (\text{by Eq. (5e)}) \quad (10a)$$

$$= w \otimes \mathbf{1} \quad (10b)$$

$$= w(\tilde{\mathbf{d}}) \otimes \tilde{w}(\tilde{\pi}) \quad (10c)$$

*Lemma 1's Induction Step.* In the induction step, we show that the properties that we have shown for the base case propagate upwards along the derivation. In general, we will show that for any  $\tilde{\mathbf{d}}_{\cap} = (q_0, X, q_M) \xrightarrow{p}_L (q_0, X_1, q_1), \dots, (q_{M-1}, X_M, q_M) \Rightarrow_L \dots$ , we can construct  $\psi(\tilde{\mathbf{d}}_{\cap}) = \langle \tilde{\mathbf{d}}, \tilde{\pi} \rangle$  such that the mapping is injective and that the properties in Eqs. (7) and (8) hold. Additionally, as for the base case, we will show that  $\tilde{\pi}$  connects  $q_0$  with  $q_M$  and that  $\tilde{\mathbf{d}}$  is rooted at  $X$ . As our inductive hypothesis, we will assume that each of these hypotheses hold for the subderivations rooted at each of the child nonterminals  $(q_0, X_1, q_1), \dots, (q_{M-1}, X_M, q_M)$ . We note that the rules from  $\mathcal{P}_{\cap}$  which apply to a nonterminal of form  $(q_0, X, q_M)$  with  $X \in \Sigma$  are discussed in base case #1, if instead  $X \in \mathcal{N}$ , we either have base case #2 or one of the rules defined by Eq. (5d) and Eq. (5g); we discuss each now.

*Lemma 1's Induction Step, Part #1.* The topmost rule applied in  $\tilde{\mathbf{d}}_{\cap}$  is  $p = (q_0, a, q_2) \xrightarrow{\mathbf{1}} (q_0, \varepsilon, q_1)(q_1, a, q_2)$  defined by Eq. (5g). We denote with  $\tilde{\mathbf{d}}_{\cap,1}$  the subderivation rooted at  $(q_0, \varepsilon, q_1)$ , and we observe that the only possible form for this derivation is  $(q_0, \varepsilon, q_1) \xrightarrow{p}_L \varepsilon$  for

some  $p = (q_0, \varepsilon, q_1) \xrightarrow{w} \varepsilon$ . We denote with  $\tilde{\mathbf{d}}_{\Gamma,2}$  the subderivation rooted at  $(q_1, a, q_2)$ , then by inductive hypothesis, we know that there is a mapping  $\psi(\tilde{\mathbf{d}}_{\Gamma,2}) = \langle \tilde{\mathbf{d}}_2, \tilde{\pi}_2 \rangle$  such that Eqs. (7) and (8) are satisfied.

Then we construct  $\psi(\tilde{\mathbf{d}}_{\Gamma}) = \langle \tilde{\mathbf{d}}, \tilde{\pi} \rangle$ , so that  $\tilde{\mathbf{d}} = \tilde{\mathbf{d}}_2$  and  $\tilde{\pi} = q_0 \xrightarrow{\varepsilon/w} q_1 \circ \tilde{\pi}_2$ . As the yield of the subderivation rooted at  $(q_0, \varepsilon, q_1)$  is  $\varepsilon$ , the yield of  $\tilde{\mathbf{d}}_{\Gamma}$  is the same as that of  $\tilde{\mathbf{d}}_{\Gamma,2}$ . Further, the yield of  $\tilde{\pi}$  is the same as  $\tilde{\pi}_2$ . We thus have that:

$$\text{yield}(\tilde{\mathbf{d}}_{\Gamma}) = \text{yield}(\tilde{\mathbf{d}}_{\Gamma,2}), \quad \text{yield}(\tilde{\mathbf{d}}) = \text{yield}(\tilde{\mathbf{d}}_2), \quad \text{yield}(\tilde{\pi}) = \text{yield}(\tilde{\pi}_2) \quad (11)$$

By induction, we have that the yield is preserved. Similarly, we have that the weight is preserved:

$$w(\tilde{\mathbf{d}}_{\Gamma}) = \mathbf{1} \otimes w(\tilde{\mathbf{d}}_{\Gamma,1}) \otimes w(\tilde{\mathbf{d}}_{\Gamma,2}) \quad (12a)$$

$$= \mathbf{1} \otimes w \otimes w(\tilde{\mathbf{d}}_2) \otimes \tilde{w}(\tilde{\pi}_2) \quad (\text{inductive hypothesis}) \quad (12b)$$

$$= w(\tilde{\mathbf{d}}_2) \otimes (w \otimes \tilde{w}(\tilde{\pi}_2)) \quad (\text{commutativity}) \quad (12c)$$

$$= w(\tilde{\mathbf{d}}) \otimes \tilde{w}(\tilde{\pi}) \quad (12d)$$

Finally, by induction we assume that  $\tilde{\pi}_2$  connects state  $q_1$  with state  $q_2$ , which implies that  $\tilde{\pi}$  connects state  $q_0$  with state  $q_2$ .

*Lemma 1's Induction Step, Part # 2.* The topmost rule applied in  $\tilde{\mathbf{d}}_{\Gamma}$  is  $p = (q_0, X, q_M) \xrightarrow{w} (q_0, X_1, q_1), \dots, (q_{M-1}, X_M, q_M)$  defined by Eq. (5d). By induction we assume that the subderivation  $\tilde{\mathbf{d}}_{\Gamma,m}$  rooted at  $(q_{m-1}, X_m, q_m)$  is mapped by  $\psi$  into a subderivation  $\tilde{\mathbf{d}}_m$  rooted at  $X_m$  and a path  $\tilde{\pi}_m$ , so that  $\text{yield}(\tilde{\mathbf{d}}_{\Gamma,m}) = \text{yield}(\tilde{\mathbf{d}}_m) = \text{yield}(\tilde{\pi}_m)$  and that  $w(\tilde{\mathbf{d}}_{\Gamma,m}) = w(\tilde{\mathbf{d}}_m) \otimes \tilde{w}(\tilde{\pi}_m)$ . We then define  $\psi(\tilde{\mathbf{d}}_{\Gamma}) = \langle \tilde{\mathbf{d}}, \tilde{\pi} \rangle$  where  $\tilde{\mathbf{d}} = X \Rightarrow_L^p X_1, \dots, X_M \Rightarrow_L \dots$  with  $p = X \xrightarrow{w} X_1, \dots, X_M$  and  $\tilde{\pi} = \tilde{\pi}_1 \circ \dots \circ \tilde{\pi}_M$ . As the states of neighboring triplets are matched, and by induction we assume that  $\tilde{\pi}_m$  connects states  $q_{m-1}$  with state  $q_m$ , we have that  $\tilde{\pi}$  is a path from  $q_0$  to  $q_M$ . We note that the yield of  $\tilde{\mathbf{d}}$  is obtained by concatenation of  $\text{yield}(\tilde{\mathbf{d}}_m)$  from left to right, and that similarly the yield of  $\tilde{\pi}$  is obtained by concatenation of  $\text{yield}(\tilde{\pi}_m)$  from left to right. This, together with the inductive hypothesis proves Eq. (8) of the lemma—as the yield of  $\tilde{\mathbf{d}}_{\Gamma}$  will also be given by the concatenation of  $\text{yield}(\tilde{\mathbf{d}}_{\Gamma,m})$  from left to right. We now show that Eq. (7) on weights holds:

$$w(\tilde{\mathbf{d}}_{\Gamma}) = w \otimes \bigotimes_{m=1}^M w(\tilde{\mathbf{d}}_{\Gamma,m}) \quad (13a)$$

$$= w \otimes \bigotimes_{m=1}^M w(\tilde{\mathbf{d}}_m) \otimes \tilde{w}(\tilde{\pi}_m) \quad (\text{inductive hypothesis}) \quad (13b)$$

$$= \left( w \otimes \bigotimes_{m=1}^M w(\tilde{\mathbf{d}}_m) \right) \otimes \bigotimes_{m=1}^M \tilde{w}(\tilde{\pi}_m) \quad (\text{commutativity}) \quad (13c)$$

$$= w(\tilde{\mathbf{d}}) \otimes \tilde{w}(\tilde{\pi}) \quad (13d)$$

We have defined  $\psi$  in a bottom-up fashion. At each step changing the topmost rule would result either in a different tree  $\tilde{\mathbf{d}}$  or in a different path  $\tilde{\pi}$ , which proves injectivity. The proof that  $\psi$  is surjective is very similar, and consists in showing by induction, that for any  $\tilde{\mathbf{d}} \in \mathcal{D}_{\mathcal{G}}(X)$ , and for any path  $\tilde{\pi}$  that does not have a sequence of  $\varepsilon$ -arc before a final state, it is always possible to build a derivation in  $\mathcal{D}_{\mathcal{G}_{\Gamma}}((p(\tilde{\pi}), X, n(\tilde{\pi})))$ . We limit ourselves to noting that it is always possible to do so by using rules from Eqs. (5d) to (5f), as in the original Bar-Hillel construction, and by using rules defined by Eq. (5g) to cover  $\varepsilon$ -arcs in the WFSa. ■

**Lemma 2.** For any triplet  $(q_I, \tilde{S}, q) \in \mathcal{N}_\cap$ , with  $q_I \in Q_I, q \in Q$ , there is a bijection  $\xi(\tilde{\mathbf{d}}_\cap) = \langle \tilde{\mathbf{d}}, \tilde{\pi} \rangle$  from  $\mathcal{D}_{\mathcal{G}_\cap}((q_I, \tilde{S}, q))$  to the join  $(\mathcal{D}_{\mathcal{G}}(\mathcal{S}) \bowtie \mathcal{D}_{\mathcal{A}}(\{q_I, q\}))$ , and we have that:

$$w(\tilde{\mathbf{d}}_\cap) = w(\tilde{\mathbf{d}}) \otimes \tilde{w}(\tilde{\pi}) \quad (14)$$

$$\text{yield}(\tilde{\mathbf{d}}_\cap) = \text{yield}(\tilde{\mathbf{d}}) = \text{yield}(\tilde{\pi}) \quad (15)$$

*Proof.* We now present an inductive proof (similar to the above) for this lemma.

*Lemma 2's Base Case.* The topmost rule applied in  $\tilde{\mathbf{d}}_\cap$  is  $(q_I, \tilde{S}, q) \xrightarrow{1} (q_I, \mathcal{S}, q)$  from rules defined by Eq. (5c). We denote with  $\tilde{\mathbf{d}}_{\cap,1}$  the subderivation rooted at  $(q_I, \mathcal{S}, q)$ . Then by Lemma 1, we know that there is a mapping  $\psi(\tilde{\mathbf{d}}_{\cap,1}) = \langle \tilde{\mathbf{d}}_1, \tilde{\pi}_1 \rangle$  such that Eqs. (14) and (15) are satisfied. We then define  $\xi(\tilde{\mathbf{d}}_\cap) = \langle \tilde{\mathbf{d}}_1, \tilde{\pi}_1 \rangle$ , and one can easily see that the properties in Eqs. (14) and (15) are satisfied.

*Lemma 2's Induction Step.* The topmost rule applied in  $\tilde{\mathbf{d}}_\cap$  is  $(q_I, \tilde{S}, q_1) \xrightarrow{1} (q_I, \tilde{S}, q_0)(q_0, \varepsilon, q_1)$  from rules defined by Eq. (5b). We denote with  $\tilde{\mathbf{d}}_{\cap,1}$  the subderivation rooted at  $(q_I, \tilde{S}, q_0)$ , and we assume by induction that  $\xi(\tilde{\mathbf{d}}_{\cap,1}) = \langle \tilde{\mathbf{d}}_1, \tilde{\pi}_1 \rangle$  and that properties in Eqs. (14) and (15) hold. We denote with  $\tilde{\mathbf{d}}_{\cap,2}$  the subderivation rooted at  $(q_0, \varepsilon, q_1)$ , and we observe that the only possible form for this derivation is  $(q_0, \varepsilon, q_1) \xrightarrow{p}_L \varepsilon$  for some  $p = (q_0, \varepsilon, q_1) \xrightarrow{w} \varepsilon$ . Then we can construct  $\xi(\tilde{\mathbf{d}}_\cap) = \langle \tilde{\mathbf{d}}, \tilde{\pi} \rangle$ , where  $\tilde{\mathbf{d}} = \tilde{\mathbf{d}}_1$  and  $\tilde{\pi} = \tilde{\pi}_1 \circ q_0 \xrightarrow{\varepsilon/w} q_1$ . The property in Eq. (15) is clearly satisfied, for property Eq. (14), we have:

$$w(\tilde{\mathbf{d}}_\cap) = \mathbf{1} \otimes w(\tilde{\mathbf{d}}_{\cap,1}) \otimes w(\tilde{\mathbf{d}}_{\cap,2}) \quad (16a)$$

$$= w(\tilde{\mathbf{d}}_{\cap,1}) \otimes w \quad (\text{weight of } \tilde{\mathbf{d}}_{\cap,2}) \quad (16b)$$

$$= w(\tilde{\mathbf{d}}_1) \otimes \tilde{w}(\tilde{\pi}_1) \otimes w \quad (\text{inductive hypothesis}) \quad (16c)$$

$$= w(\tilde{\mathbf{d}}) \otimes \tilde{w}(\tilde{\pi}) \quad (\text{weight of } \tilde{\pi}) \quad (16d)$$

As for Lemma 1 we note that modifying the topmost rule in  $\tilde{\mathbf{d}}_\cap$ , would always result either in a different derivation  $\tilde{\mathbf{d}}$  or in a different path  $\tilde{\pi}$ , which proves injectivity. Surjectivity can be shown by induction, similarly to how we did for injectivity. We will simply note that given any derivation  $\tilde{\mathbf{d}}$  rooted at  $\mathcal{S}$ , and given any path  $\tilde{\pi}$  starting from an initial state, it is always possible to build a matching derivation  $\tilde{\mathbf{d}}_\cap$  in  $\mathcal{D}_{\mathcal{G}_\cap}((p(\tilde{\pi}), \tilde{S}, n(\tilde{\pi})))$ , by using the result from Lemma 1, and applying rules defined by Eqs. (5b) and (5c). ■

We can finally prove Theorem 1, which we restate here for convenience.

**Theorem 1.** Let  $\mathcal{G}$  be a WCFG and  $\mathcal{A}$  a WFSA over the same alphabet  $\Sigma$  and commutative semiring  $\mathcal{W}$ . Let  $\mathcal{G}_\cap$  be the grammar obtained with our generalized construction. Then we have strong equivalence between  $\mathcal{G}_\cap$  and  $\langle \mathcal{G}, \mathcal{A} \rangle$ ; meaning that there is a weight-preserving, yield-preserving bijection between  $\mathcal{D}_{\mathcal{G}_\cap}$  and  $(\mathcal{D}_{\mathcal{G}} \bowtie \mathcal{D}_{\mathcal{A}})$ .

*Proof.* Any derivation  $\mathbf{d}_\cap$  in  $\mathcal{D}_{\mathcal{G}_\cap}(\mathcal{S})$  takes the form  $\mathcal{S} \xrightarrow{p}_L (q_I, \tilde{S}, q_F) \Rightarrow_L \dots$  with  $p = \mathcal{S} \xrightarrow{\lambda(q_I) \otimes \rho(q_F)} (q_I, \tilde{S}, q_F)$ , for  $q_I \in Q_I$  and  $q_F \in Q_F$ . We denote with  $\tilde{\mathbf{d}}_\cap$  the subderivation rooted at  $(q_I, \tilde{S}, q_F)$ . We can thus define  $\phi(\mathbf{d}_\cap) = \langle \mathbf{d}, \pi \rangle = \langle \tilde{\mathbf{d}}, \tilde{\pi} \rangle$ , where  $\xi(\tilde{\mathbf{d}}_\cap) = \langle \tilde{\mathbf{d}}, \tilde{\pi} \rangle$ , and  $\xi$  is the bijection defined in Lemma 2. By Lemma 2 we have that  $\tilde{\mathbf{d}} = \mathbf{d}$  is rooted at  $\mathcal{S}$ , and that  $\tilde{\pi} = \pi$  has initial and final states:  $p(\tilde{\pi}) = q_I$  and  $n(\tilde{\pi}) = q_F$ . Clearly,  $\text{yield}(\mathbf{d}_\cap) = \text{yield}(\tilde{\mathbf{d}}_\cap)$  and, by Lemma 2,  $\text{yield}(\tilde{\mathbf{d}}_\cap) = \text{yield}(\tilde{\mathbf{d}}) = \text{yield}(\tilde{\pi})$ . Further, by definition  $\text{yield}(\mathbf{d}) = \text{yield}(\tilde{\mathbf{d}})$

and  $\text{yield}(\tilde{\pi}) = \text{yield}(\pi)$ . Moreover, we have that:

$$w(\mathbf{d}_\cap) = w(p) \otimes w(\tilde{\mathbf{d}}_\cap) \quad (\text{weight of a derivation}) \quad (17a)$$

$$= w(p) \otimes w(\tilde{\mathbf{d}}) \otimes \tilde{w}(\tilde{\pi}) \quad (\text{Lemma 2}) \quad (17b)$$

$$= \lambda(q_I) \otimes \rho(q_F) \otimes w(\tilde{\mathbf{d}}) \otimes \tilde{w}(\tilde{\pi}) \quad (\text{weight of } p) \quad (17c)$$

$$= w(\tilde{\mathbf{d}}) \otimes \lambda(q_I) \otimes \tilde{w}(\tilde{\pi}) \otimes \rho(q_F) \quad (\text{commutativity}) \quad (17d)$$

$$= w(\mathbf{d}) \otimes w(\pi) \quad (\text{definition of weight of a path}) \quad (17e)$$

which proves that  $\phi$  is weight and yield preserving. By Lemma 2 we know that  $\xi$  is a bijection, which implies that modifying the topmost rule  $p$  would result in a different tuple  $\langle \mathbf{d}, \pi \rangle$ . This proves the injectivity of  $\phi$ . Conversely, consider any path  $\pi$  connecting an initial state with a final one and any derivation  $\mathbf{d}$  rooted at  $S$ , such that  $\text{yield}(\mathbf{d}) = \text{yield}(\pi)$ . By Lemma 2 we know that it is always possible to construct a subderivation  $\tilde{\mathbf{d}}_\cap$ , rooted at  $(q_I, \tilde{S}, q_F)$ , that satisfies Eqs. (14) and (15). Thus we can construct  $\mathbf{d}_\cap = S \xrightarrow{p} (q_I, \tilde{S}, q_F) \Rightarrow_L \dots$  with  $p = S \xrightarrow{\lambda(q_I) \otimes \rho(q_F)} (q_I, \tilde{S}, q_F)$  a rule from Eq. (5a). This shows the surjectivity of  $\phi$ . ■

## B.2 Proof of Corollary 1

**Corollary 1.**  $\mathcal{G}_\cap$  and  $\langle \mathcal{G}, \mathcal{A} \rangle$  are weakly equivalent, meaning that  $L_{\mathcal{G}_\cap}(\mathbf{s}) = L_{\mathcal{G}}(\mathbf{s}) \otimes L_{\mathcal{A}}(\mathbf{s})$  whenever the values on the right-hand side are defined.

*Proof.* §2.1 defined both  $L_{\mathcal{A}}(\mathbf{s})$  and  $L_{\mathcal{G}}(\mathbf{s})$  as sums over derivations that yield  $\mathbf{s}$ . If there are only finitely many such derivations, then the sum is well-defined by applying the associative–commutative operator  $\oplus$  finitely many times. However, footnote 5 noted that countably infinite sums can arise. We treat this issue by augmenting the semiring with an operator  $\bigoplus$  that is applied to a countable (possibly infinite) multiset of weights and returns a value that is interpreted as the sum of those weights, or else returns a special “undefined” value  $\perp \notin A$  to indicate that the sum diverges.

We require  $\bigoplus$  to satisfy the following axioms for any two countable multisets  $I, J \subseteq A$  such that

$$\bigoplus I = W \in A \quad \bigoplus J = V \in A \quad (18)$$

- *Infinite distributivity:* Let  $I \otimes J$  denote the multiset  $\{\{i \otimes j : i \in I, j \in J\}\}$ . Then  $\bigoplus(I \otimes J) = W \otimes V \in A$ .
- *Infinite associativity:* for any partition<sup>14</sup>  $I = \bigcup_{k \in K} I_k$ , we have  $\bigoplus I_k \in A$  for each  $k \in K$  and furthermore  $\bigoplus_{k \in K} (\bigoplus I_k) = W$ .
- *Base cases:* For any  $w, w' \in A$ ,  $\bigoplus \{\{w, w'\}\} = w \oplus w'$ ,  $\bigoplus \{\{w\}\} = w$ , and  $\bigoplus \{\emptyset\} = \mathbf{0}$ . Together with the previous property, this ensures that  $\bigoplus$  agrees with the  $\oplus$ -based definition on finite multisets.

The first two axioms are adapted from part of Mohri (2002)’s definition of closed semirings. The proof of Corollary 1 uses only the first axiom, as follows. Given a string  $\mathbf{s}$  such that  $L_{\mathcal{A}}(\mathbf{s}), L_{\mathcal{G}}(\mathbf{s}) \in A$ . By definition (§§2.3–2.4),  $L_{\mathcal{A}}(\mathbf{s}) = \bigoplus I$  and  $L_{\mathcal{G}}(\mathbf{s}) = \bigoplus J$  if we define  $I = \{\{w(\pi) : \pi \in \mathcal{D}_{\mathcal{A}}(\mathbf{s})\}\}$  and  $J = \{\{w(\mathbf{d}) : \mathbf{d} \in \mathcal{D}_{\mathcal{G}}(\mathbf{s})\}\}$ . Then also  $L_{\mathcal{G}_\cap}(\mathbf{s}) = \bigoplus(I \otimes J)$  since  $I \otimes J = \{\{w(\mathbf{d}) : \mathbf{d} \in \mathcal{D}_{\mathcal{G}_\cap}(\mathbf{s})\}\}$  according to Theorem 1. By infinite distributivity, then,  $L_{\mathcal{G}_\cap}(\mathbf{s}) = (\bigoplus I) \otimes (\bigoplus J) = L_{\mathcal{A}}(\mathbf{s}) \otimes L_{\mathcal{G}}(\mathbf{s}) \in A$  as claimed. ■

<sup>14</sup>Recall that partitions are definitionally disjoint.