# A Bottom-Up Deterministic Weighted Tree Automaton for the $n$-Gram Yield Function

Matthias Büchse Tobias Denkinger Heiko Vogler

Institute of Theoretical Computer Science
Technische Universität Dresden

2013-09-26

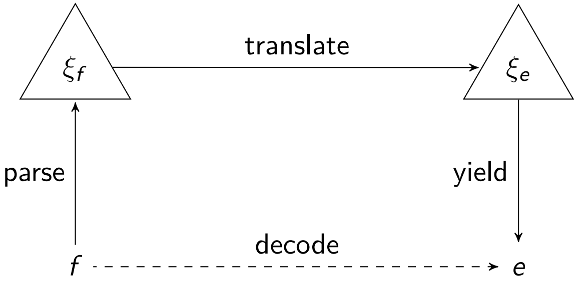
# Outline

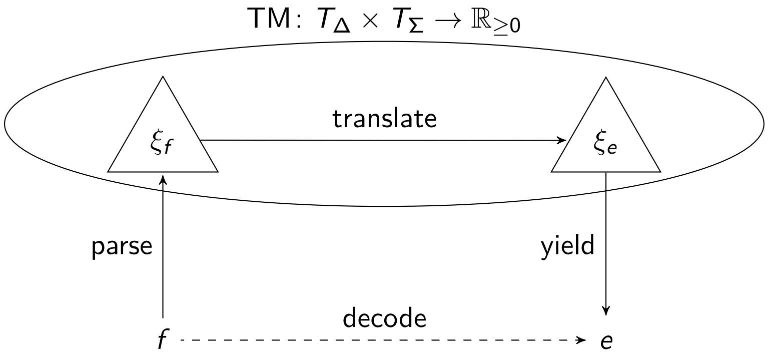
1. Motivation: Statistical Machine Translation
2. Generalized *n*-Gram Models
3. Lifting using Bar-Hillel, Perles, Shamir Algorithm
4. Generalized *n*-Gram Weighted Tree Automaton
5. Further Research

# Motivation: Statistical Machine Translation

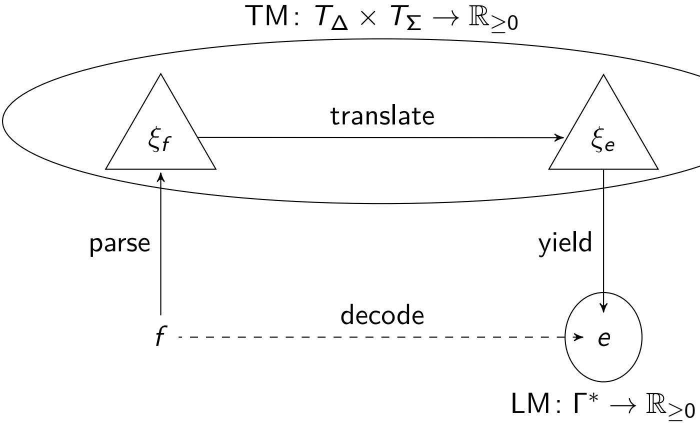
$$f \xrightarrow{\text{decode}} e$$

# Motivation: Statistical Machine Translation




TM: $T_\Delta \times T_\Sigma \to \mathbb{R}_{\geq 0}$
$\xi_f$
translate
$\xi_e$
parse
yield
decode
$f$
$e$

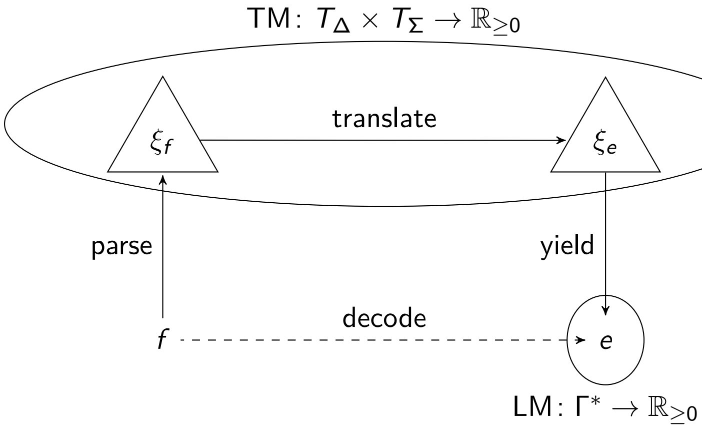
# Motivation: Statistical Machine Translation




---

$\mathbf{\Gamma} = \mathbf{\Sigma}^{(0)}$

# Motivation: Statistical Machine Translation




$$\hat{e} = \pi_{\text{out}}\Big(\text{argmax}_d\ (\ (f \triangleleft \text{TM})\ \triangleright\ \text{LM}\ )(d)\Big)$$

---

$\boldsymbol{\Gamma} = \boldsymbol{\Sigma}^{(0)}$

# Generalized $n$-Gram Models

Let $n \in \mathbb{N} \setminus \{0\}$.

A *generalized n-gram model* is a tuple $N = (\Gamma, \mu_1, \ldots, \mu_n)$ with

- the alphabet $\Gamma$
- a mapping $\mu_i \colon \Gamma^i \to \mathbb{R}_{\geq 0}$ for every $i \in \{1, \ldots, n\}$

# Generalized $n$-Gram Models

Let $n \in \mathbb{N} \setminus \{0\}$.

A *generalized n-gram model* is a tuple $N = (\Gamma, \mu_1, \dots, \mu_n)$ with

- the alphabet $\Gamma$
- a mapping $\mu_i \colon \Gamma^i \to \mathbb{R}_{\geq 0}$ for every $i \in \{1, \dots, n\}$

The semantics of $N$ is $[\![N]\!] \colon \Gamma^* \to \mathbb{R}_{\geq 0} \cup \{\bot\}$ where

$$[\![N]\!](w) = \begin{cases} \prod_{w' \text{ } n\text{-gram of w}} \mu_n(w') & \text{if } |w| \geq n \\ \mu_{|w|}(w) & \text{if } 1 \leq |w| < n \\ \bot & \text{if } |w| = 0 \end{cases}$$

# Generalized $n$-Gram Models (Example)

Let $n = 3$, $\Gamma = \{\text{Garcia}, \text{y}, \text{tres}, \text{asociados}, \textbf{.}\}$, and

$$\mu_3\colon \begin{array}{ll} \text{Garcia y tres} & \mapsto 1/5 \\ \text{y tres asociados} & \mapsto 1/3 \\ \text{tres asociados .} & \mapsto 1/4 \\ \dots & \end{array} \qquad \mu_2\colon \begin{array}{ll} \text{Garcia .} & \mapsto 1/2 \\ \dots & \end{array} \qquad \dots$$

# Generalized $n$-Gram Models (Example)

Let $n = 3$, $\Gamma = \{\text{Garcia}, \text{y}, \text{tres}, \text{asociados}, \textbf{.}\}$, and

$$\mu_3\colon \begin{array}{ll} \text{Garcia y tres} & \mapsto 1/5 \\ \text{y tres asociados} & \mapsto 1/3 \\ \text{tres asociados .} & \mapsto 1/4 \\ \dots & \end{array} \qquad \mu_2\colon \begin{array}{ll} \text{Garcia .} & \mapsto 1/2 \\ \dots & \end{array} \qquad \dots$$

$$[\![N]\!](\text{Garcia y tres asociados .}) =$$

# Generalized $n$-Gram Models (Example)

Let $n = 3$, $\Gamma = \{\text{Garcia}, \text{y}, \text{tres}, \text{asociados}, \textbf{.}\}$, and

$$\mu_3\colon \begin{array}{ll} \text{Garcia y tres} & \mapsto 1/5 \\ \text{y tres asociados} & \mapsto 1/3 \\ \text{tres asociados .} & \mapsto 1/4 \\ \ldots & \end{array} \qquad \mu_2\colon \begin{array}{ll} \text{Garcia .} & \mapsto 1/2 \\ \ldots & \end{array} \qquad \ldots$$

$$[\![N]\!](\text{Garcia y tres asociados .}) = 1/5$$

# Generalized $n$-Gram Models (Example)

Let $n = 3$, $\Gamma = \{\text{Garcia}, \text{y}, \text{tres}, \text{asociados}, \textbf{.}\}$, and

$$\mu_3\colon \begin{array}{lll} \text{Garcia y tres} & \mapsto & 1/5 \\ \text{y tres asociados} & \mapsto & 1/3 \\ \text{tres asociados .} & \mapsto & 1/4 \\ \ldots & & \end{array} \qquad \mu_2\colon \begin{array}{lll} \text{Garcia .} & \mapsto & 1/2 \\ \ldots & & \end{array} \qquad \ldots$$

$$[\![N]\!](\text{Garcia y tres asociados .}) = 1/5 \cdot 1/3$$

# Generalized $n$-Gram Models (Example)

Let $n = 3$, $\Gamma = \{\text{Garcia}, \text{y}, \text{tres}, \text{asociados}, \textbf{.}\}$, and

$$\mu_3\colon \begin{array}{ll} \text{Garcia y tres} & \mapsto 1/5 \\ \text{y tres asociados} & \mapsto 1/3 \\ \text{tres asociados .} & \mapsto 1/4 \\ \ldots & \end{array} \qquad \mu_2\colon \begin{array}{ll} \text{Garcia .} & \mapsto 1/2 \\ \ldots & \end{array} \qquad \ldots$$

$$[\![N]\!](\text{Garcia y tres asociados .}) = 1/5 \cdot 1/3 \cdot 1/4$$

# Generalized $n$-Gram Models (Example)

Let $n = 3$, $\Gamma = \{\text{Garcia}, \text{y}, \text{tres}, \text{asociados}, \textbf{.}\}$, and

$$\mu_3 : \begin{array}{ll} \text{Garcia y tres} & \mapsto 1/5 \\ \text{y tres asociados} & \mapsto 1/3 \\ \text{tres asociados .} & \mapsto 1/4 \\ \ldots & \end{array} \qquad \mu_2 : \begin{array}{ll} \text{Garcia .} & \mapsto 1/2 \\ \ldots & \end{array} \qquad \ldots$$

$$[\![N]\!](\text{Garcia y tres asociados .}) = 1/5 \cdot 1/3 \cdot 1/4$$

$$[\![N]\!](\text{Garcia .}) = 1/2$$

# Generalized $n$-Gram Models (Example)

Let $n = 3$, $\Gamma = \{\text{Garcia}, \text{y}, \text{tres}, \text{asociados}, \textbf{.}\}$, and

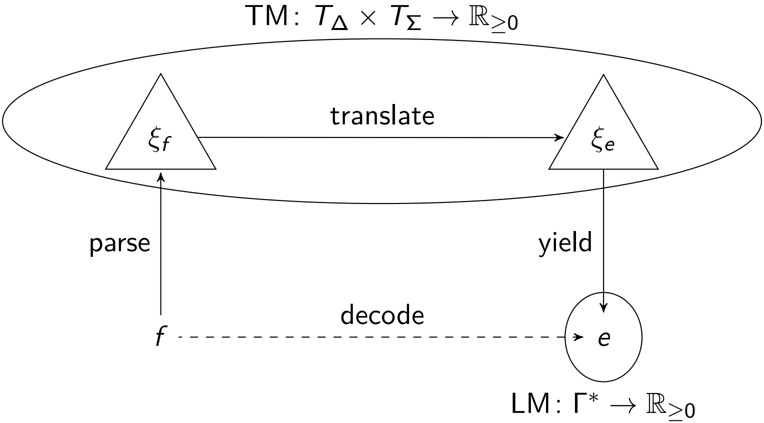
$$\mu_3 \colon \begin{array}{ll} \text{Garcia y tres} & \mapsto 1/5 \\ \text{y tres asociados} & \mapsto 1/3 \\ \text{tres asociados .} & \mapsto 1/4 \\ \ldots & \end{array} \qquad \mu_2 \colon \begin{array}{ll} \text{Garcia .} & \mapsto 1/2 \\ \ldots & \end{array} \qquad \ldots$$

$$[\![N]\!](\text{Garcia y tres asociados .}) = 1/5 \cdot 1/3 \cdot 1/4$$

$$[\![N]\!](\text{Garcia .}) = 1/2$$

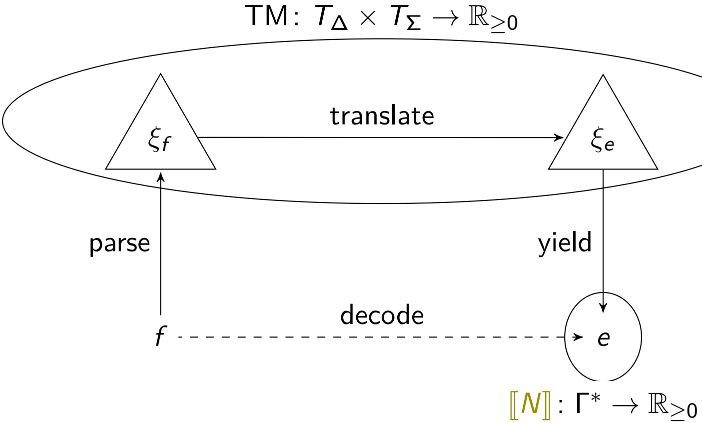
$$[\![N]\!](\varepsilon) = \bot$$

# Recall: Statistical Machine Translation



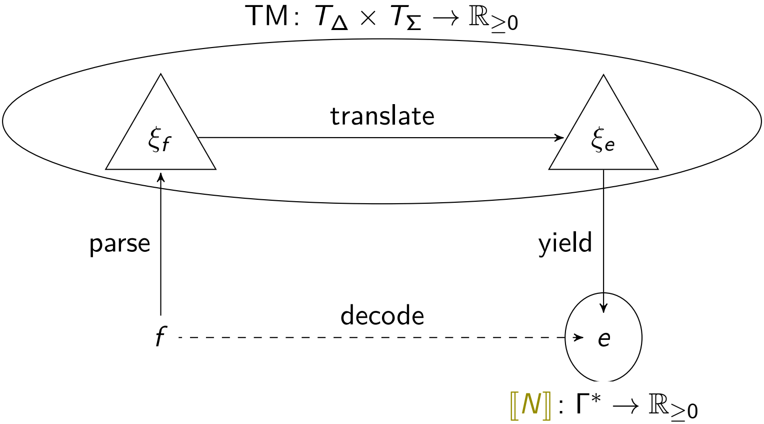

$$\hat{e} = \pi_{\text{out}}\Big( \text{argmax}_d \, \big( \, (f \triangleleft \text{TM}) \, \triangleright \, \text{LM} \, \big)(d) \Big)$$

# Recall: Statistical Machine Translation




$$\hat{e} = \pi_{\text{out}}\Big(\operatorname{argmax}_d \big( (f \triangleleft \text{TM}) \triangleright [\![N]\!] \big)(d)\Big)$$

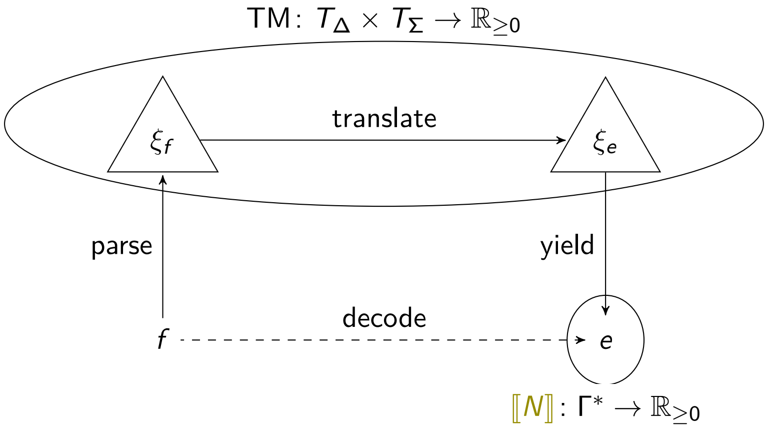
# Recall: Statistical Machine Translation




$$\hat{e} = \pi_{\text{out}}\Big(\operatorname{argmax}_d \big( (f \triangleleft \text{TM}) \triangleright [\![N]\!] \big)(d)\Big)$$

$T_\Delta \times T_\Sigma \to \mathbb{R}_{\geq 0}$ $\quad$ $\Gamma^* \to \mathbb{R}_{\geq 0}$

# Recall: Statistical Machine Translation




$$\hat{e} = \pi_{\text{out}}\Big(\operatorname{argmax}_d \big((f \lhd \text{TM}) \rhd [\![N]\!]\big)(d)\Big)$$

$T_\Delta \times T_\Sigma \to \mathbb{R}_{\geq 0}$ $\quad$ $\Gamma^* \to \mathbb{R}_{\geq 0}$

lift to trees
$\text{yield}^{-1}([\![N]\!])$

# Lifting using Bar-Hillel, Perles, Shamir (BHPS)

- for every $\mathcal{L} \in \mathrm{CF}$ and $R \in \mathrm{REG}$: $\mathcal{L} \cap R \in \mathrm{CF}$

[Bar-Hillel, Perles, Shamir, 1961]

---

**CF** . . . context-free languages
**REG** . . . regular languages

# Lifting using Bar-Hillel, Perles, Shamir (BHPS)

- for every $\mathcal{L} \in \mathrm{CF}$ and $R \in \mathrm{REG}$: $\mathcal{L} \cap R \in \mathrm{CF}$

  [Bar-Hillel, Perles, Shamir, 1961]

- for every $\mathcal{L} \in \mathrm{WRT}$ and $R \in \mathrm{WREG}$: $\mathcal{L} \odot \mathrm{yield}^{-1}(R) \in \mathrm{WRT}$

  [Maletti, Satta, 2009]

**CF** . . . context-free languages
**REG** . . . regular languages

**WRT** . . . weighted regular tree languages
**WREG** . . . weighted regular languages

# Lifting using Bar-Hillel, Perles, Shamir (BHPS)

- for every $\mathcal{L} \in \mathrm{CF}$ and $R \in \mathrm{REG}$: $\mathcal{L} \cap R \in \mathrm{CF}$

  [Bar-Hillel, Perles, Shamir, 1961]

- for every $\mathcal{L} \in \mathrm{WRT}$ and $R \in \mathrm{WREG}$: $\mathcal{L} \odot \mathrm{yield}^{-1}(R) \in \mathrm{WRT}$

  [Maletti, Satta, 2009]

- in our case: $\mathcal{L}$ is the characteristic mapping of $T_\Sigma$ and $R = [\![N]\!]$

  $\mathrm{yield}^{-1}([\![N]\!]) \in \mathrm{WRT}$

---

**CF** . . . context-free languages
**REG** . . . regular languages

**WRT** . . . weighted regular tree languages
**WREG** . . . weighted regular languages

# Lifting using Bar-Hillel, Perles, Shamir (BHPS)

- for every $\mathcal{L} \in \mathrm{CF}$ and $R \in \mathrm{REG}$: $\mathcal{L} \cap R \in \mathrm{CF}$

  [Bar-Hillel, Perles, Shamir, 1961]

- for every $\mathcal{L} \in \mathrm{WRT}$ and $R \in \mathrm{WREG}$: $\mathcal{L} \odot \mathrm{yield}^{-1}(R) \in \mathrm{WRT}$

  [Maletti, Satta, 2009]

- in our case: $\mathcal{L}$ is the characteristic mapping of $T_\Sigma$ and $R = [\![N]\!]$

  $\mathrm{yield}^{-1}([\![N]\!]) \in \mathrm{WRT}$

  - the constructed WTA is *not* bottom-up deterministic
  - *but* for every tree, only one run is relevant

---

**CF** . . . context-free languages
**REG** . . . regular languages
**WTA**. . . weighted tree automaton
**WRT** . . . weighted regular tree languages
**WREG** . . . weighted regular languages

# Intermezzo: Weighted Tree Automata (WTA)

A *weighted tree automaton* is a tuple $\mathcal{A} = (Q, \Sigma, \delta, F)$ with

- $Q$ finite set (states)
- $\Sigma$ ranked alphabet (input symbols)
- $\delta = (\delta_k \mid k \geq 0)$ with $\delta_k \colon Q^k \times \Sigma^{(k)} \times Q \to \mathbb{R}_{\geq 0}$ (transition function)
- $F \subseteq Q$ finite set (final states)

# Intermezzo: Weighted Tree Automata (WTA)

A *weighted tree automaton* is a tuple $\mathcal{A} = (Q, \Sigma, \delta, F)$ with

- $Q$ finite set (states)
- $\Sigma$ ranked alphabet (input symbols)
- $\delta = (\delta_k \mid k \geq 0)$ with $\delta_k \colon Q^k \times \Sigma^{(k)} \times Q \to \mathbb{R}_{\geq 0}$ (transition function)
- $F \subseteq Q$ finite set (final states)

A *run (of $\mathcal{A}$) on $\xi$* is a mapping $\kappa \colon \mathrm{pos}(\xi) \to Q$.

# Intermezzo: Weighted Tree Automata (WTA)

A *weighted tree automaton* is a tuple $\mathcal{A} = (Q, \Sigma, \delta, F)$ with

- $Q$ finite set (states)
- $\Sigma$ ranked alphabet (input symbols)
- $\delta = (\delta_k \mid k \geq 0)$ with $\delta_k \colon Q^k \times \Sigma^{(k)} \times Q \to \mathbb{R}_{\geq 0}$ (transition function)
- $F \subseteq Q$ finite set (final states)

A *run (of $\mathcal{A}$) on $\xi$* is a mapping $\kappa \colon \text{pos}(\xi) \to Q$.

The *weight of $\kappa$* is $\text{wt}(\kappa) = \prod_{\rho \in \text{pos}(\kappa)} \delta_k(\kappa(\rho 1), \ldots, \kappa(\rho k), \sigma, \kappa(\rho))$.

---

$\boldsymbol{\sigma}$. . . label in $\xi$ at $\rho$ $\qquad$ $\boldsymbol{k}$. . . rank of $\sigma$

# Intermezzo: Weighted Tree Automata (WTA)

A *weighted tree automaton* is a tuple $\mathcal{A} = (Q, \Sigma, \delta, F)$ with

- $Q$ finite set (states)
- $\Sigma$ ranked alphabet (input symbols)
- $\delta = (\delta_k \mid k \geq 0)$ with $\delta_k \colon Q^k \times \Sigma^{(k)} \times Q \to \mathbb{R}_{\geq 0}$ (transition function)
- $F \subseteq Q$ finite set (final states)

A *run (of $\mathcal{A}$) on $\xi$* is a mapping $\kappa\colon \operatorname{pos}(\xi) \to Q$.

The *weight of $\kappa$* is $\operatorname{wt}(\kappa) = \prod_{\rho \in \operatorname{pos}(\kappa)} \delta_k(\kappa(\rho 1), \ldots, \kappa(\rho k), \sigma, \kappa(\rho))$.

The *semantics of $\mathcal{A}$* is $[\![\mathcal{A}]\!]\colon T_\Sigma \to \mathbb{R}_{\geq 0}$ where

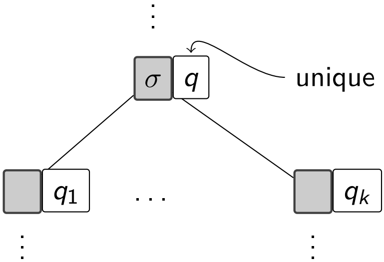
$$[\![\mathcal{A}]\!](\xi) = \sum_{\substack{\kappa \text{ run on } \xi \\ \kappa(\varepsilon) \in F}} \operatorname{wt}(\kappa) \; .$$

---

$\boldsymbol{\sigma}$ . . . label in $\xi$ at $\rho$ $\qquad$ $\boldsymbol{k}$ . . . rank of $\sigma$

# Intermezzo: Weighted Tree Automata (WTA)

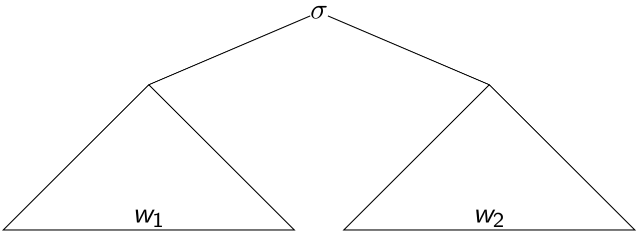
$\mathcal{A}$ is *bottom-up deterministic*:




$$\delta_k(q_1, \ldots, q_k, \sigma, q) > 0$$

# Generalized $n$-Gram WTA (Idea)

$n = 3$

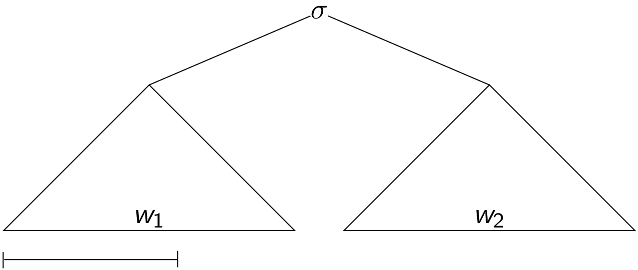



$[\![N]\!](w_1 w_2) = \qquad \cdot \qquad \cdot$

$|w_1| \geq n, |w_2| \geq n$

# Generalized $n$-Gram WTA (Idea)

$n = 3$

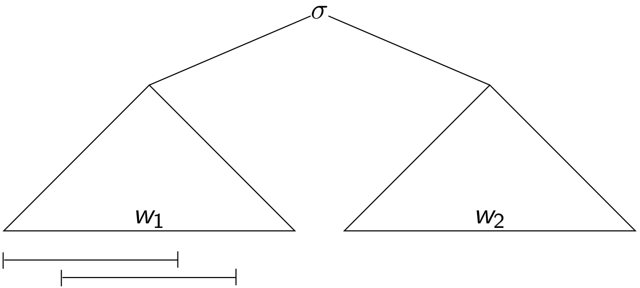



$[\![N]\!](w_1 w_2) = \quad \cdot \quad \cdot$

$|w_1| \geq n, |w_2| \geq n$

# Generalized $n$-Gram WTA (Idea)

$n = 3$

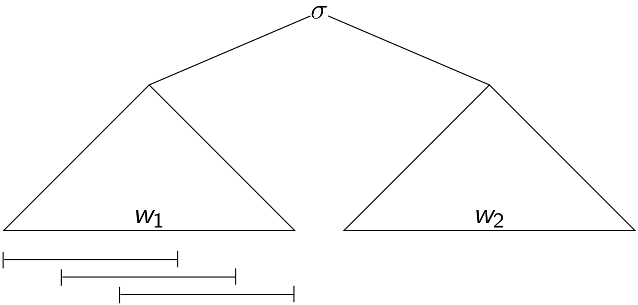



$[\![N]\!](w_1 w_2) = \quad \cdot \quad \cdot$

$|w_1| \geq n, |w_2| \geq n$

# Generalized $n$-Gram WTA (Idea)

$n = 3$

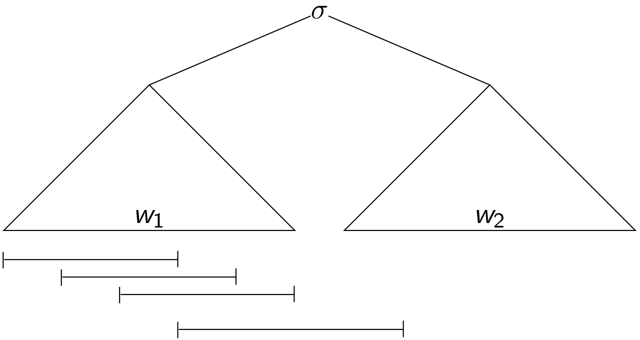



$[\![N]\!](w_1 w_2) = \quad \cdot \quad \cdot$

$|w_1| \geq n, |w_2| \geq n$

# Generalized $n$-Gram WTA (Idea)

$n = 3$

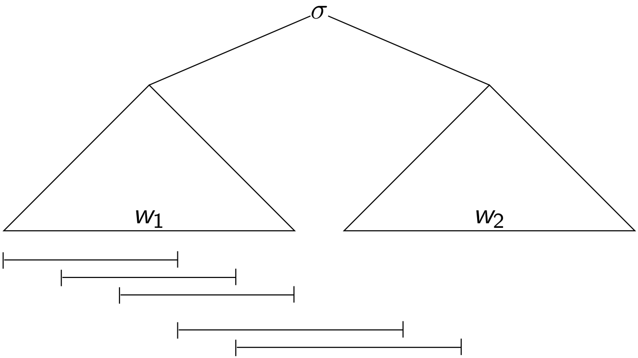



$[\![N]\!](w_1 w_2) = \qquad \cdot \qquad \cdot$

$|w_1| \geq n, |w_2| \geq n$

# Generalized $n$-Gram WTA (Idea)

$n = 3$

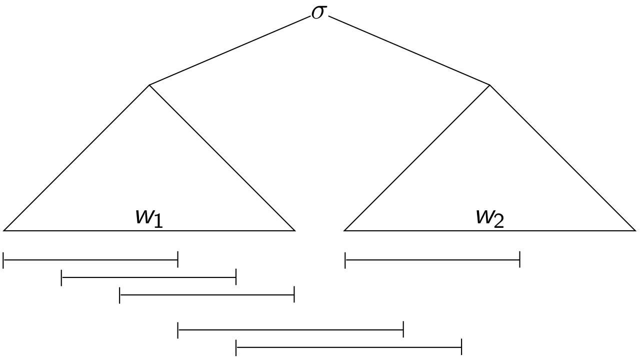



$[\![N]\!](w_1 w_2) = \qquad \cdot \qquad \cdot$

$|w_1| \geq n, |w_2| \geq n$

# Generalized $n$-Gram WTA (Idea)

$n = 3$

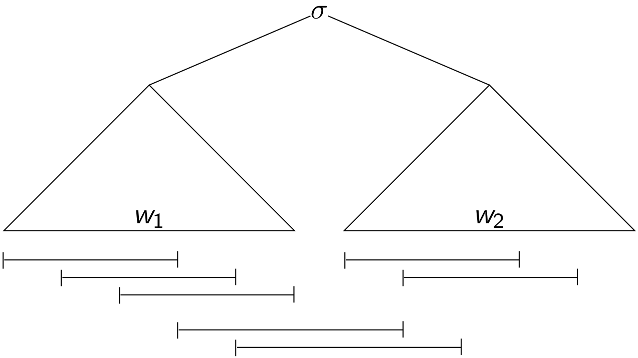



$[\![N]\!](w_1 w_2) = \quad \cdot \quad \cdot$

$|w_1| \geq n, |w_2| \geq n$

# Generalized $n$-Gram WTA (Idea)

$n = 3$

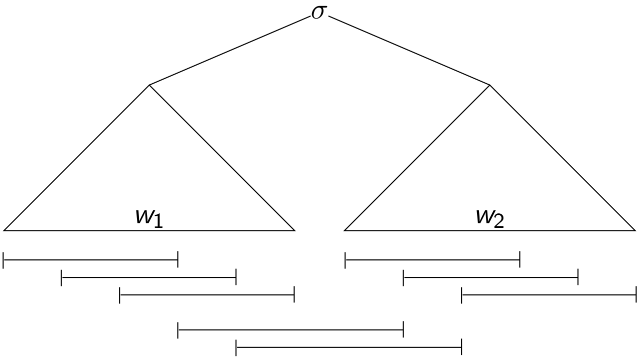



$[\![N]\!](w_1 w_2) = \qquad\qquad \cdot \qquad\qquad\qquad \cdot$

$|w_1| \geq n, |w_2| \geq n$

# Generalized $n$-Gram WTA (Idea)

$n = 3$

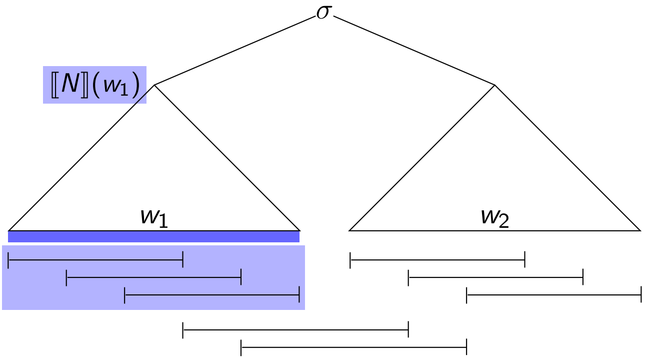



$[\![N]\!](w_1 w_2) = \quad \cdot \quad \cdot$

$|w_1| \geq n, |w_2| \geq n$

# Generalized $n$-Gram WTA (Idea)

$n = 3$

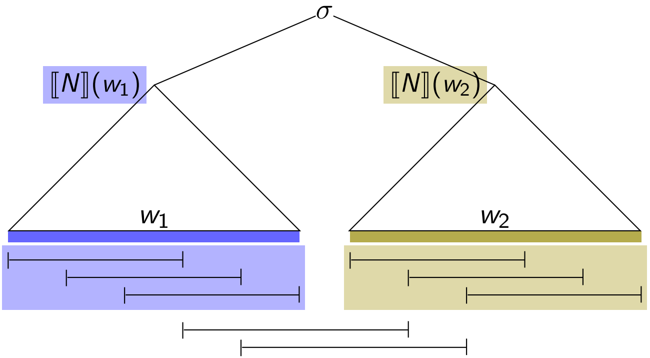



$$[\![N]\!](w_1 w_2) = [\![N]\!](w_1) \cdot \qquad \cdot$$

$|w_1| \geq n, |w_2| \geq n$

# Generalized $n$-Gram WTA (Idea)

$n = 3$

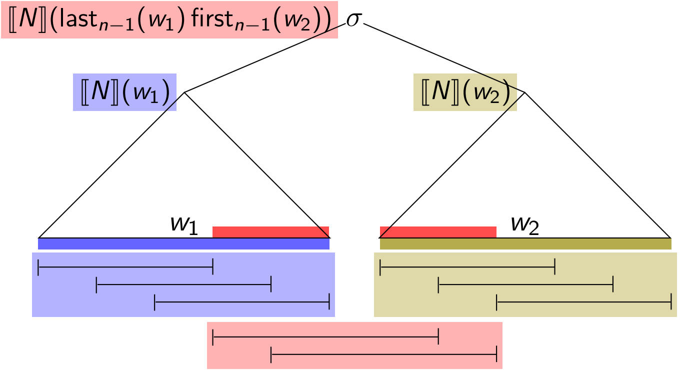



$$[\![N]\!](w_1 w_2) = [\![N]\!](w_1) \cdot \qquad \cdot [\![N]\!](w_2)$$

$|w_1| \geq n, |w_2| \geq n$

# Generalized $n$-Gram WTA (Idea)

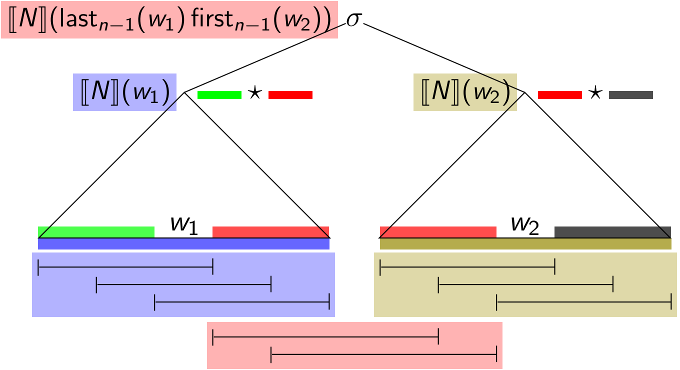
$n = 3$




$$[\![N]\!](w_1 w_2) = [\![N]\!](w_1) \cdot [\![N]\!](\mathrm{last}_{n-1}(w_1)\,\mathrm{first}_{n-1}(w_2)) \cdot [\![N]\!](w_2)$$

$|w_1| \geq n, |w_2| \geq n$

$n = 3$

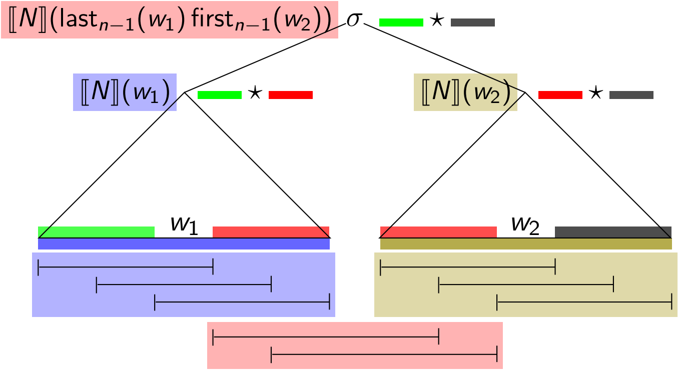



$$[\![N]\!](w_1 w_2) = [\![N]\!](w_1) \cdot [\![N]\!](\mathrm{last}_{n-1}(w_1)\,\mathrm{first}_{n-1}(w_2)) \cdot [\![N]\!](w_2)$$

$|w_1| \geq n, |w_2| \geq n$

# Generalized $n$-Gram WTA (Idea)

$n = 3$

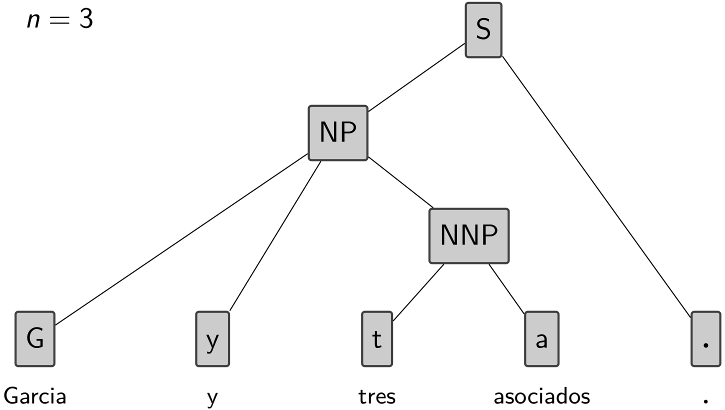



$$[\![N]\!](w_1 w_2) = [\![N]\!](w_1) \cdot [\![N]\!](\text{last}_{n-1}(w_1)\,\text{first}_{n-1}(w_2)) \cdot [\![N]\!](w_2)$$

$|w_1| \geq n, |w_2| \geq n$

# Generalized $n$-Gram WTA (Example)

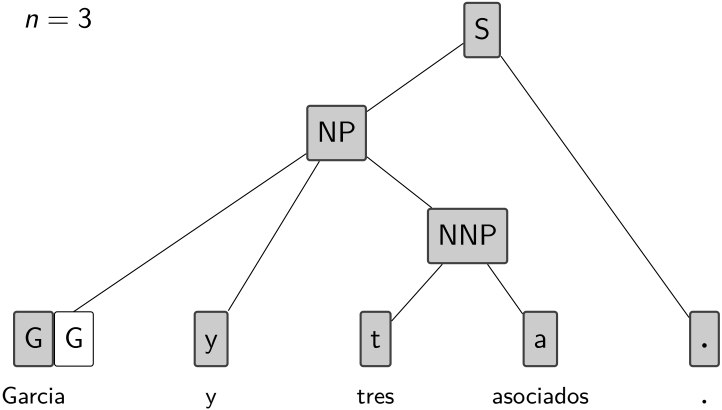
$n = 3$




$\text{wt}(\kappa) =$

# Generalized $n$-Gram WTA (Example)

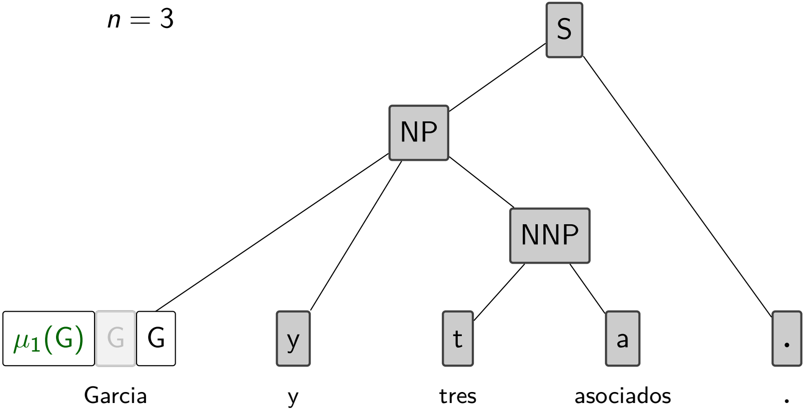
$n = 3$




$\mathrm{wt}(\kappa) =$

# Generalized $n$-Gram WTA (Example)

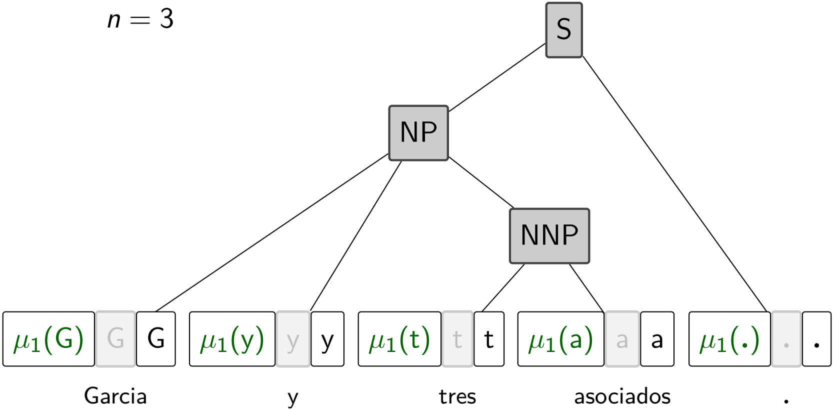
$n = 3$




$\mathrm{wt}(\kappa) = \mu_1(G)$

# Generalized $n$-Gram WTA (Example)

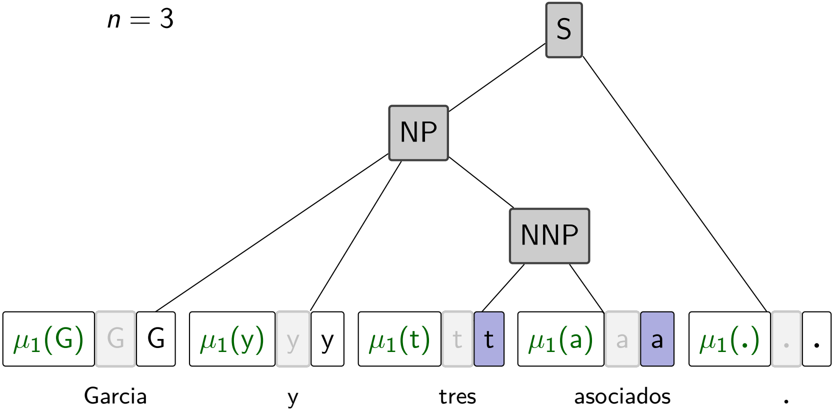
$n = 3$




$$\mathrm{wt}(\kappa) = \mu_1(\mathrm{G}) \cdot \mu_1(\mathrm{y}) \cdot \mu_1(\mathrm{t}) \cdot \mu_1(\mathrm{a}) \cdot \mu_1(.)$$

# Generalized $n$-Gram WTA (Example)

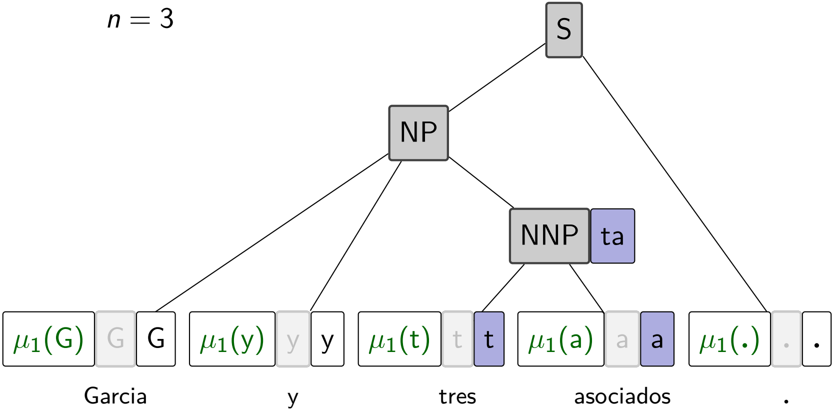
$n = 3$




$$\mathrm{wt}(\kappa) = \mu_1(\mathrm{G}) \cdot \mu_1(\mathrm{y}) \cdot \mu_1(\mathrm{t}) \cdot \mu_1(\mathrm{a}) \cdot \mu_1(.)$$

# Generalized $n$-Gram WTA (Example)

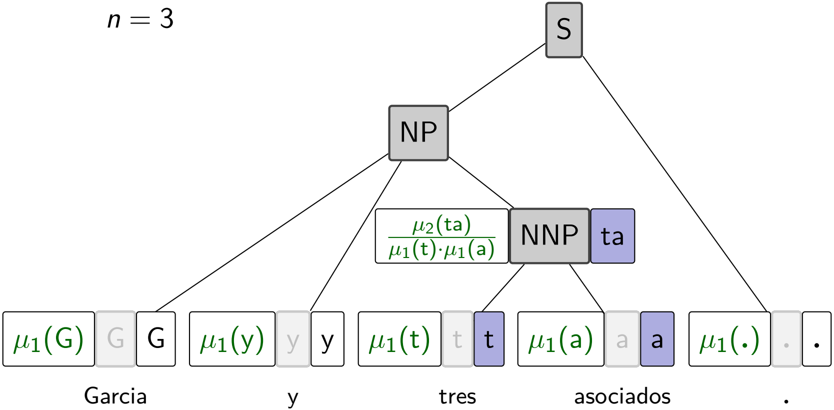
$n = 3$




$$\mathrm{wt}(\kappa) = \mu_1(\mathrm{G}) \cdot \mu_1(\mathrm{y}) \cdot \mu_1(\mathrm{t}) \cdot \mu_1(\mathrm{a}) \cdot \mu_1(.)$$

# Generalized $n$-Gram WTA (Example)

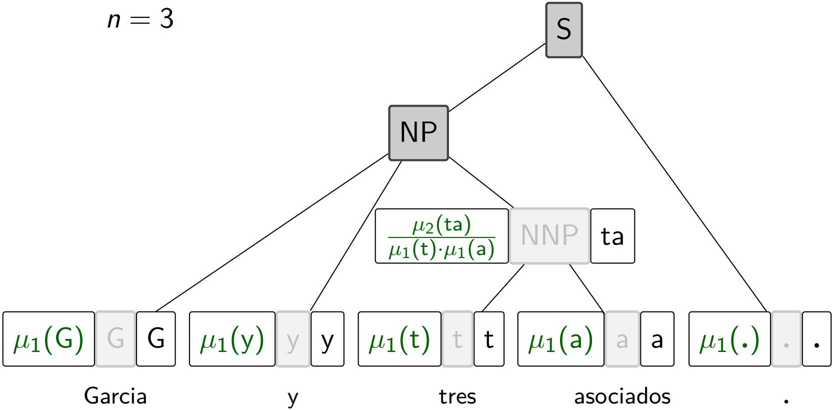
$n = 3$




$$\mathrm{wt}(\kappa) = \mu_1(\mathrm{G}) \cdot \mu_1(\mathrm{y}) \cdot \mu_1(\mathrm{t}) \cdot \mu_1(\mathrm{a}) \cdot \mu_1(.) \cdot \frac{\mu_2(\mathrm{ta})}{\mu_1(\mathrm{t}) \cdot \mu_1(\mathrm{a})}$$

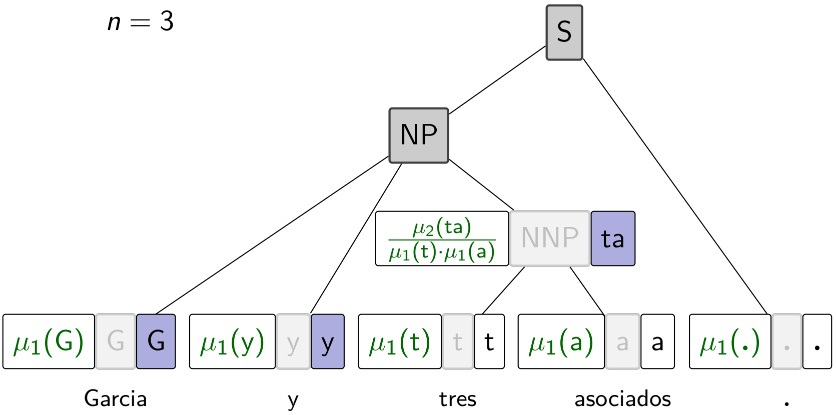
# Generalized $n$-Gram WTA (Example)

$n = 3$




$$\text{wt}(\kappa) = \mu_1(\text{G}) \cdot \mu_1(\text{y}) \cdot \cancel{\mu_1(\text{t})} \cdot \cancel{\mu_1(\text{a})} \cdot \mu_1(.) \cdot \frac{\mu_2(\text{ta})}{\cancel{\mu_1(\text{t})}\cdot\cancel{\mu_1(\text{a})}}$$

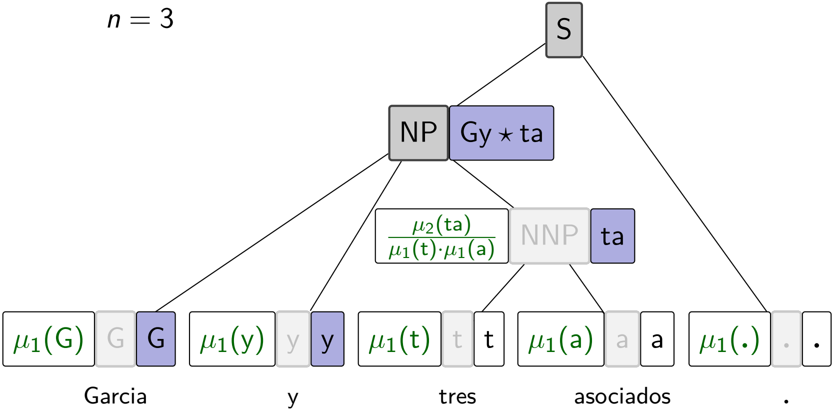
# Generalized $n$-Gram WTA (Example)




$$\text{wt}(\kappa) = \mu_1(\text{G}) \cdot \mu_1(\text{y}) \cdot \cancel{\mu_1(\text{t})} \cdot \cancel{\mu_1(\text{a})} \cdot \mu_1(.) \cdot \frac{\mu_2(\text{ta})}{\cancel{\mu_1(\text{t})}\cdot\cancel{\mu_1(\text{a})}}$$

# Generalized $n$-Gram WTA (Example)

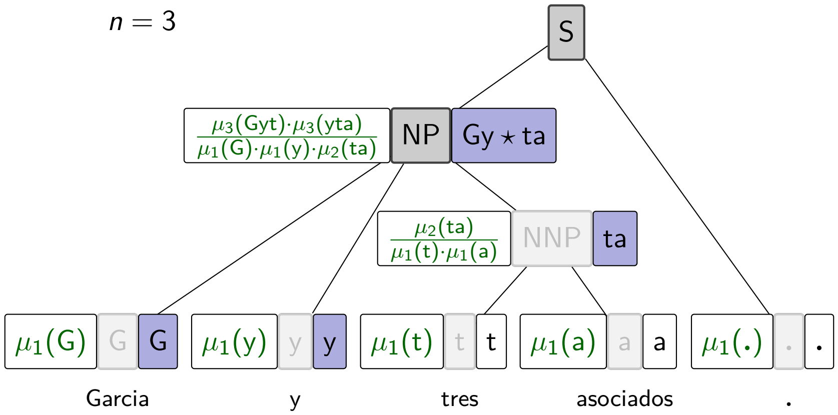
$n = 3$




$$\mathrm{wt}(\kappa) = \mu_1(\mathrm{G}) \cdot \mu_1(\mathrm{y}) \cdot \cancel{\mu_1(\mathrm{t})} \cdot \cancel{\mu_1(\mathrm{a})} \cdot \mu_1(.) \cdot \frac{\mu_2(\mathrm{ta})}{\cancel{\mu_1(\mathrm{t})} \cdot \cancel{\mu_1(\mathrm{a})}}$$

# Generalized $n$-Gram WTA (Example)

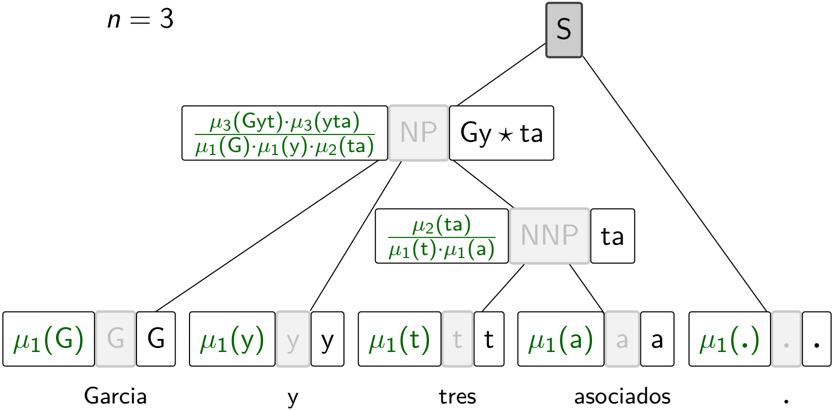
$n = 3$




$$\mathrm{wt}(\kappa) = \mu_1(\mathrm{G}) \cdot \mu_1(\mathrm{y}) \cdot \cancel{\mu_1(\mathrm{t})} \cdot \cancel{\mu_1(\mathrm{a})} \cdot \mu_1(.) \cdot \frac{\mu_2(\mathrm{ta})}{\cancel{\mu_1(\mathrm{t})} \cdot \cancel{\mu_1(\mathrm{a})}} \cdot \frac{\mu_3(\mathrm{Gyt}) \cdot \mu_3(\mathrm{yta})}{\mu_1(\mathrm{G}) \cdot \mu_1(\mathrm{y}) \cdot \mu_2(\mathrm{ta})}$$

# Generalized $n$-Gram WTA (Example)

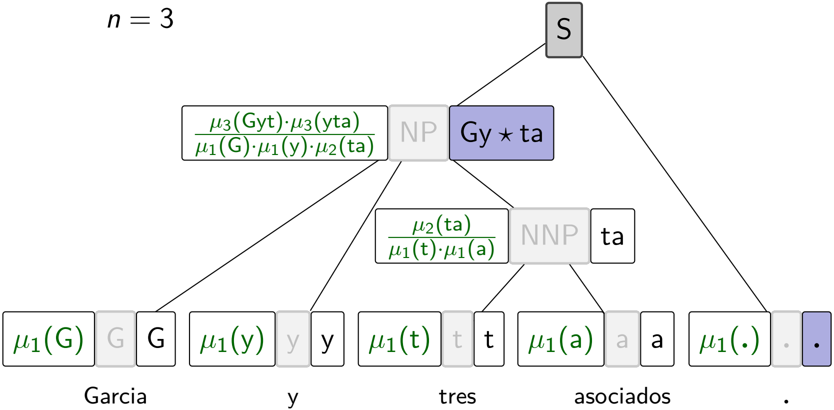
$n = 3$




$$\mathrm{wt}(\kappa) = \cancel{\mu_1(\mathrm{G})} \cdot \cancel{\mu_1(\mathrm{y})} \cdot \cancel{\mu_1(\mathrm{t})} \cdot \cancel{\mu_1(\mathrm{a})} \cdot \mu_1(.) \cdot \frac{\cancel{\mu_2(\mathrm{ta})}}{\cancel{\mu_1(\mathrm{t})} \cdot \cancel{\mu_1(\mathrm{a})}} \cdot \frac{\mu_3(\mathrm{Gyt}) \cdot \mu_3(\mathrm{yta})}{\cancel{\mu_1(\mathrm{G})} \cdot \cancel{\mu_1(\mathrm{y})} \cdot \cancel{\mu_2(\mathrm{ta})}}$$

# Generalized $n$-Gram WTA (Example)

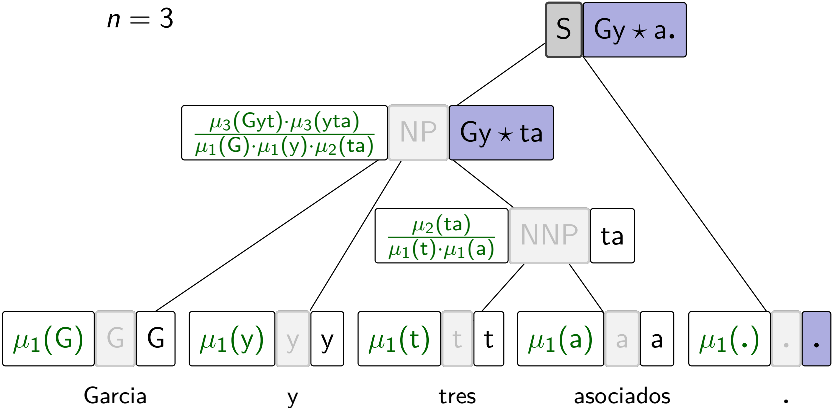
$n = 3$




$$\text{wt}(\kappa) = \cancel{\mu_1(\text{G})} \cdot \cancel{\mu_1(\text{y})} \cdot \cancel{\mu_1(\text{t})} \cdot \cancel{\mu_1(\text{a})} \cdot \mu_1(.) \cdot \frac{\cancel{\mu_2(\text{ta})}}{\cancel{\mu_1(\text{t})}\cdot\cancel{\mu_1(\text{a})}} \cdot \frac{\mu_3(\text{Gyt})\cdot\mu_3(\text{yta})}{\cancel{\mu_1(\text{G})}\cdot\cancel{\mu_1(\text{y})}\cdot\cancel{\mu_2(\text{ta})}}$$

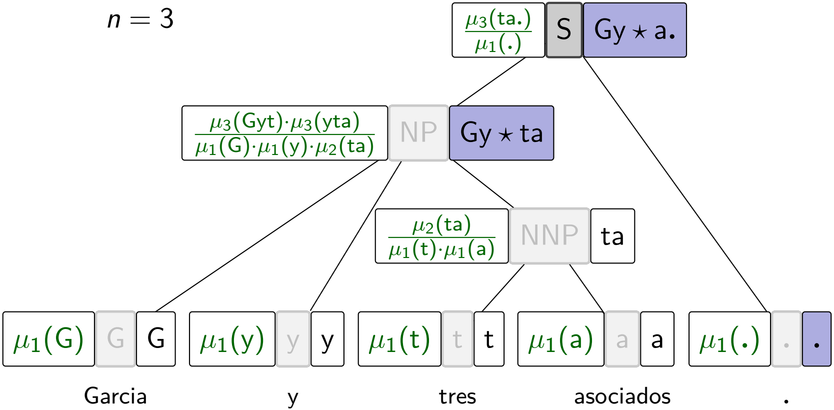
# Generalized $n$-Gram WTA (Example)

$n = 3$




$$\mathrm{wt}(\kappa) = \cancel{\mu_1(\mathrm{G})} \cdot \cancel{\mu_1(\mathrm{y})} \cdot \cancel{\mu_1(\mathrm{t})} \cdot \cancel{\mu_1(\mathrm{a})} \cdot \mu_1(.) \cdot \frac{\cancel{\mu_2(\mathrm{ta})}}{\cancel{\mu_1(\mathrm{t})} \cdot \cancel{\mu_1(\mathrm{a})}} \cdot \frac{\mu_3(\mathrm{Gyt}) \cdot \mu_3(\mathrm{yta})}{\cancel{\mu_1(\mathrm{G})} \cdot \cancel{\mu_1(\mathrm{y})} \cdot \cancel{\mu_2(\mathrm{ta})}}$$

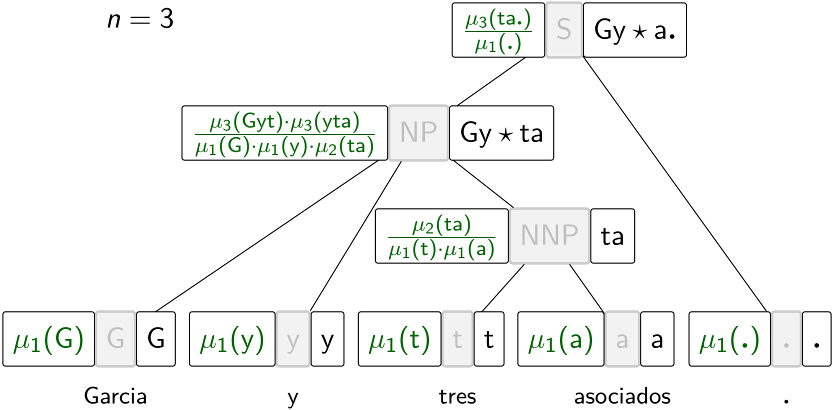
$n = 3$




$$\mathrm{wt}(\kappa) = \cancel{\mu_1(G)} \cdot \cancel{\mu_1(y)} \cdot \cancel{\mu_1(t)} \cdot \cancel{\mu_1(a)} \cdot \mu_1(.) \cdot \frac{\cancel{\mu_2(ta)}}{\cancel{\mu_1(t)} \cdot \cancel{\mu_1(a)}} \cdot \frac{\mu_3(Gyt) \cdot \mu_3(yta)}{\cancel{\mu_1(G)} \cdot \cancel{\mu_1(y)} \cdot \cancel{\mu_2(ta)}} \cdot \frac{\mu_3(ta.)}{\mu_1(.)}$$

# Generalized $n$-Gram WTA (Example)

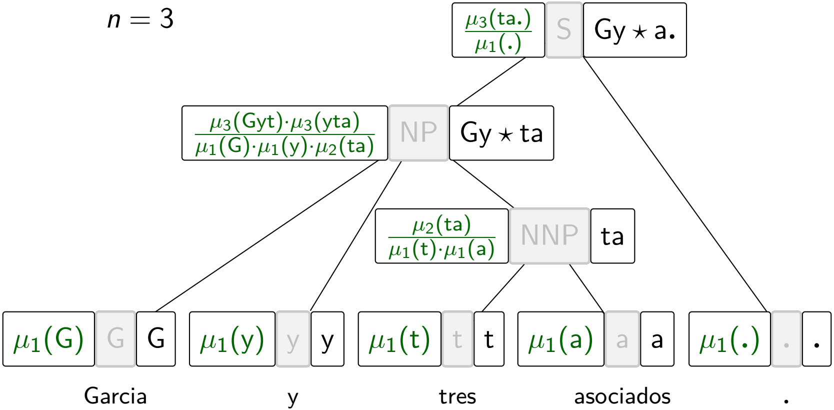
$n = 3$




$$\mathrm{wt}(\kappa) = \cancel{\mu_1(G)} \cdot \cancel{\mu_1(y)} \cdot \cancel{\mu_1(t)} \cdot \cancel{\mu_1(a)} \cdot \cancel{\mu_1(.)} \cdot \frac{\cancel{\mu_2(ta)}}{\cancel{\mu_1(t)} \cdot \cancel{\mu_1(a)}} \cdot \frac{\mu_3(Gyt) \cdot \mu_3(yta)}{\cancel{\mu_1(G)} \cdot \cancel{\mu_1(y)} \cdot \cancel{\mu_2(ta)}} \cdot \frac{\mu_3(ta.)}{\cancel{\mu_1(.)}}$$

# Generalized $n$-Gram WTA (Example)

$n = 3$




$$\begin{aligned}\mathrm{wt}(\kappa) &= \cancel{\mu_1}(G) \cdot \cancel{\mu_1}(y) \cdot \cancel{\mu_1}(t) \cdot \cancel{\mu_1}(a) \cdot \cancel{\mu_1}(.) \cdot \frac{\cancel{\mu_2(\mathrm{ta})}}{\cancel{\mu_1(t)} \cdot \cancel{\mu_1(a)}} \cdot \frac{\mu_3(\mathrm{Gyt}) \cdot \mu_3(\mathrm{yta})}{\cancel{\mu_1(G)} \cdot \cancel{\mu_1(y)} \cdot \cancel{\mu_2(\mathrm{ta})}} \cdot \frac{\mu_3(\mathrm{ta.})}{\cancel{\mu_1(.)}} \\ &= \mu_3(\mathrm{Gyt}) \cdot \mu_3(\mathrm{yta}) \cdot \mu_3(\mathrm{ta.}) = [\![N]\!](\mathrm{Gyta.})\end{aligned}$$

# Generalized $n$-Gram WTA

The WTA $\mathcal{A}_{N,\Sigma}$ is the tuple $(Q, \Sigma, \delta, F)$ with

- $Q = \Gamma^{\leq n-1} \cup \left(\Gamma^{n-1} \times \{\star\} \times \Gamma^{n-1}\right)$
- $F = Q$

# Generalized $n$-Gram WTA

The WTA $\mathcal{A}_{N,\Sigma}$ is the tuple $(Q, \Sigma, \delta, F)$ with

- $Q = \Gamma^{\leq n-1} \cup \left(\Gamma^{n-1} \times \{\star\} \times \Gamma^{n-1}\right)$
- $F = Q$
- $\delta(q_1, \ldots, q_k, \sigma, q) = \begin{cases} g(\sigma) & \text{if } k = 0 \text{ and } q = f(\sigma) \\ \frac{g(q_1 \cdots q_k)}{g'(q_1) \cdot \ldots \cdot g'(q_k)} & \text{if } k \geq 1 \text{ and } q = f(q_1 \cdots q_k) \\ 0 & \text{otherwise} \end{cases}$

# Generalized $n$-Gram WTA

The WTA $\mathcal{A}_{N,\Sigma}$ is the tuple $(Q, \Sigma, \delta, F)$ with

- $Q = \Gamma^{\leq n-1} \cup \left(\Gamma^{n-1} \times \{\star\} \times \Gamma^{n-1}\right)$
- $F = Q$
- $\delta(q_1, \dots, q_k, \sigma, q) = \begin{cases} g(\sigma) & \text{if } k = 0 \text{ and } q = f(\sigma) \\ \frac{g(q_1 \cdots q_k)}{g'(q_1) \cdot \ldots \cdot g'(q_k)} & \text{if } k \geq 1 \text{ and } q = f(q_1 \cdots q_k) \\ 0 & \text{otherwise} \end{cases}$

  for every $v \in (\Gamma \cup \{\star\})^*$

$$f(v) = \begin{cases} v & \text{if } v \in \Gamma^* \text{ and } |v| < n \\ v_1^{n-1} \star v_{|v|-n+2}^{|v|} & \text{if } (v \in \Gamma^* \text{ and } |v| \geq n) \text{ or} \\ & \quad (v \notin \Gamma^* \text{ and } |v| \geq n) \\ \dots & \text{otherwise} \end{cases}$$

# Generalized $n$-Gram WTA

The WTA $\mathcal{A}_{N,\Sigma}$ is the tuple $(Q, \Sigma, \delta, F)$ with

- $Q = \Gamma^{\leq n-1} \cup \left(\Gamma^{n-1} \times \{\star\} \times \Gamma^{n-1}\right)$
- $F = Q$
- $\delta(q_1, \ldots, q_k, \sigma, q) = \begin{cases} g(\sigma) & \text{if } k = 0 \text{ and } q = f(\sigma) \\ \frac{g(q_1 \cdots q_k)}{g'(q_1) \cdot \ldots \cdot g'(q_k)} & \text{if } k \geq 1 \text{ and } q = f(q_1 \cdots q_k) \\ 0 & \text{otherwise} \end{cases}$

  for every $v \in (\Gamma \cup \{\star\})^*$, $v = u_0 \star u_1 \cdots \star u_l$, $u_i \in \Gamma^*$ for all $i \in \{0, \ldots, l\}$, $w \in \Sigma^*$, and $q \in Q$

$$g(v) = \begin{cases} [\![N]\!](u_0) & \text{if } l = 0 \\ N'(u_0) \cdot \left(\prod_{i=1}^{l-1} [\![N]\!](u_i)\right) \cdot N'(u_l) & \text{otherwise} \end{cases}$$

$$N'(w) = \begin{cases} [\![N]\!](w) & \text{if } |w| \geq n \\ 1 & \text{otherwise} \end{cases} \qquad g'(q) = \begin{cases} [\![N]\!](q) & \text{if } q \in \Gamma^* \\ 1 & \text{otherwise} \end{cases}$$

# Theorem

## Theorem

Let $\Sigma$ be a ranked alphabet and $N$ a generalized $n$-gram model over $\Gamma$. Then

- $[\![\mathcal{A}_{N,\Sigma}]\!] = \text{yield}^{-1}([\![N]\!])$
- $\mathcal{A}_{N,\Sigma}$ is bottom-up deterministic

# Theorem

## Theorem

Let $\Sigma$ be a ranked alphabet and $N$ a generalized $n$-gram model over $\Gamma$. Then

- $[\![\mathcal{A}_{N,\Sigma}]\!] = \text{yield}^{-1}([\![N]\!])$
- $\mathcal{A}_{N,\Sigma}$ is bottom-up deterministic

## Conjecture

$\mathcal{A}_{N,\Sigma}$ is well suited for the calculation of $\text{argmax}_d\left((f \triangleleft \text{TM}) \triangleright [\![N]\!]\right)(d)$.

# Further Research

- Implement a decoder for SMT based on
  - BHPS, and
  - the generalized *n*-gram WTA.

# Further Research

- Implement a decoder for SMT based on
  - BHPS, and
  - the generalized *n*-gram WTA.
- Compare the runtime behaviour in practice.

# References


Bar–Hillel, Yehoshua, Micha Perles, and Eliahu Shamir (1961). "On Formal Properties of Simple Phrase Structure Grammars". In: *Z. Phonetik. Sprach. Komm.* 14, pp. 143–172.

Chiang, David (2007). "Hierarchical phrase-based translation". In: *Computational Linguistics* 33(2), pp. 201–228.

Maletti, Andreas and Giorgio Satta (2009). "Parsing Algorithms Based on Tree Automata". In: *Proc. of IWPT '09*. Paris, France: ACL, pp. 1–12.

# BHSP for Trees and $n$-Gram Models

$\mathcal{L} = T_\Sigma$ **as WTA**

$Q_1 = \{\star\}$

**for every** $k \in \mathbb{N}, \sigma \in \Sigma^{(k)}$:

$$\star \xrightarrow{1} \sigma(\underbrace{\star, \ldots, \star}_{k \text{ times}})$$

**WFSA**. . . weighted finite state automaton $\quad \Gamma = \Sigma^{(0)}$

$\mathcal{L} = T_\Sigma$ **as WTA**

$Q_1 = \{\star\}$

**for every** $k \in \mathbb{N}, \sigma \in \Sigma^{(k)}$:

$$\star \xrightarrow{1} \sigma(\underbrace{\star, \ldots, \star}_{k \textbf{ times}})$$

$R = [\![(\Gamma, \mu)]\!]$ **as WFSA**

$Q_2 = \Gamma^0 \cup \cdots \cup \Gamma^{n-1}$

**for every** $q \in Q \setminus \Gamma^{n-1}, \sigma \in \Gamma$:

$$q \xrightarrow{\sigma/1} q\sigma$$

**for every** $q \in \Gamma^{n-2}, \sigma, \sigma' \in \Gamma$:

$$\sigma' q \xrightarrow{\sigma/\mu(\sigma' q \sigma)} q\sigma$$

---

**WFSA**. . . weighted finite state automaton $\quad \boldsymbol{\Gamma} = \boldsymbol{\Sigma}^{(0)}$

# BHSP for Trees and $n$-Gram Models

$\mathcal{L} = T_\Sigma$ **as WTA**

$Q_1 = \{\star\}$

**for every** $k \in \mathbb{N}, \sigma \in \Sigma^{(k)}$:

$$\star \xrightarrow{1} \sigma(\underbrace{\star, \dots, \star}_{k \text{ times}})$$

$R = [\![(\Gamma, \mu)]\!]$ **as WFSA**

$Q_2 = \Gamma^0 \cup \cdots \cup \Gamma^{n-1}$

**for every** $q \in Q \setminus \Gamma^{n-1}, \sigma \in \Gamma$:

$$q \xrightarrow{\sigma/1} q\sigma$$

**for every** $q \in \Gamma^{n-2}, \sigma, \sigma' \in \Gamma$:

$$\sigma' q \xrightarrow{\sigma/\mu(\sigma' q \sigma)} q\sigma$$

$\mathcal{L} \odot R$ **as WTA**

$Q = Q_2 \times Q_1 \times Q_2$

for every $q \in Q_2 \setminus \Gamma^{n-1}, \sigma, \sigma' \in \Gamma$:

$$q \star q\sigma \xrightarrow{1} \sigma, \qquad \sigma' q \star q\sigma \xrightarrow{\mu(\sigma' q \sigma)} \sigma$$

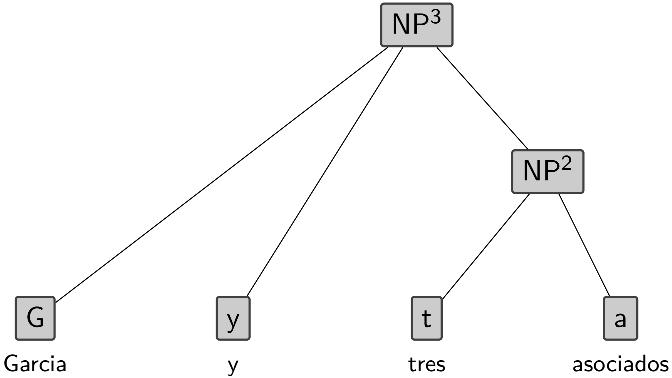
for every $k \in \mathbb{N}, \sigma \in \Sigma^{(k)} \setminus \Gamma, q_0, \dots, q_k \in Q_2$:

$$q_0 \star q_k \xrightarrow{1} \sigma(q_0 \star q_1, q_1 \star q_2, \dots, q_{k-1} \star q_k)$$

---

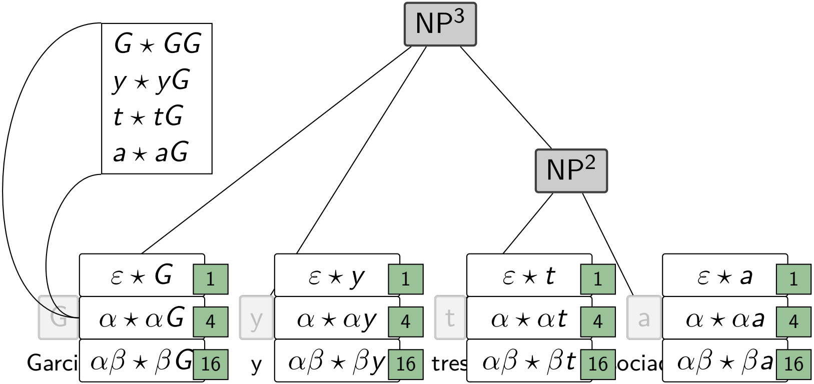
**WFSA**. . . weighted finite state automaton $\qquad \mathbf{\Gamma} = \Sigma^{(0)}$

$n = 3$
$\Gamma = \{G, y, t, a\}$



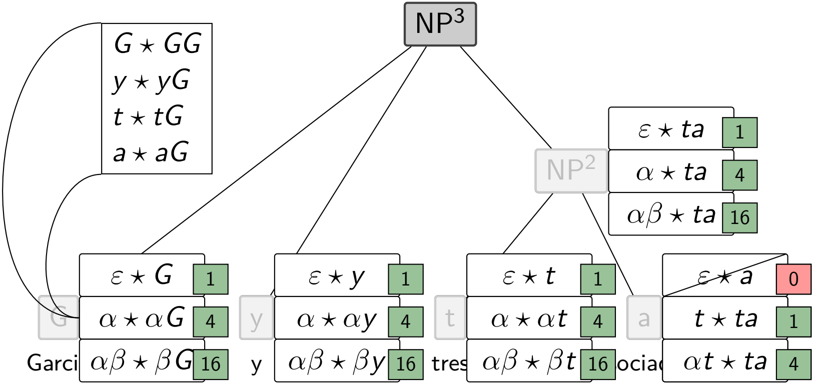
# BHSP for Trees and $n$-Gram Models (Example)

$$n = 3$$
$$\Gamma = \{G, y, t, a\}$$



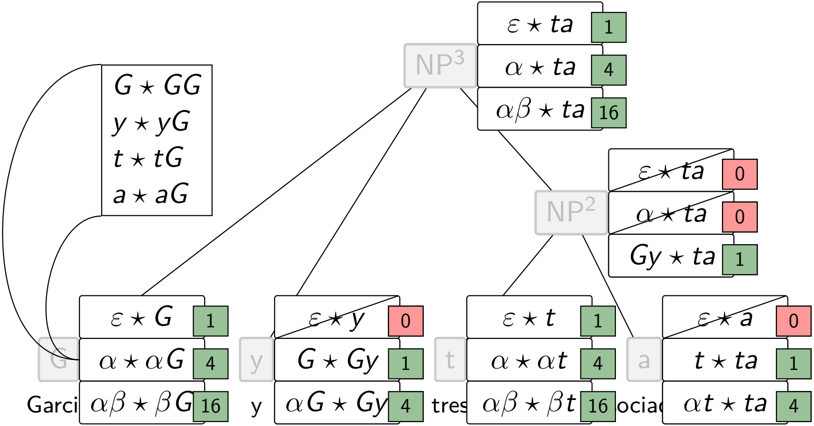
# BHSP for Trees and $n$-Gram Models (Example)

$n = 3$
$\Gamma = \{\mathrm{G}, \mathrm{y}, \mathrm{t}, \mathrm{a}\}$



# BHSP for Trees and $n$-Gram Models (Example)

$n = 3$
$\Gamma = \{G, y, t, a\}$



# BHSP for Trees and $n$-Gram Models (Example)

$n = 3$
$\Gamma = \{G, y, t, a\}$