

Was ist die Formel für die WK einer syntaktischen Analyse (ein Parsebaum) T?

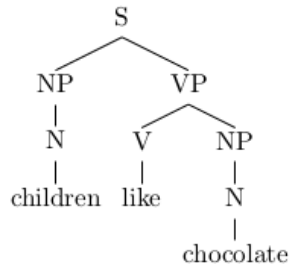
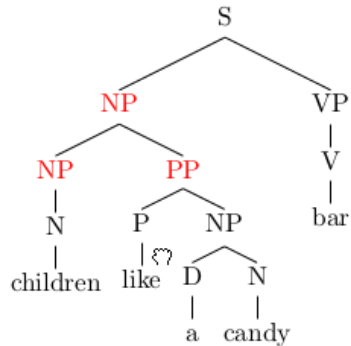
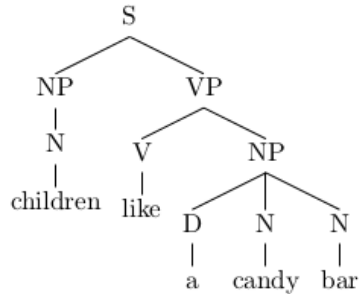
Was ist die Formel für die WK einer syntaktischen Analyse (ein Parsebaum) T?

$$p(T) = p(r_1, \dots, r_n) \stackrel{?}{=} \prod_{i=1}^n p(r_i)$$

das produkt der Regelwahrsrscheinlichkeiten?

## Führe den 3. Schritt durch

### 3 Extraktion der gewichteten Regelhäufigkeiten



Regel	$p_0$	$f_1$
$S \rightarrow NP VP$	1.00	2.00
$NP \rightarrow D N$	0.25	0.50
$NP \rightarrow D N N$	0.25	0.50
$NP \rightarrow N$	0.25	3.00
$NP \rightarrow NP PP$	0.25	0.50
$VP \rightarrow V$	0.50	0.50
$VP \rightarrow V NP$	0.50	1.50
$PP \rightarrow P NP$	1.00	0.50
$D \rightarrow a$	0.50	1.00
$D \rightarrow the$	0.50	0.00
$N \rightarrow bar$	0.25	0.50
$N \rightarrow candy$	0.25	1.00
$N \rightarrow children$	0.25	2.00
$N \rightarrow chocolate$	0.25	1.00
$V \rightarrow bar$	0.50	0.50
$V \rightarrow like$	0.50	1.50
$P \rightarrow like$	1.00	0.50

Wie bekommen wir **f1** für folgenden Regeln?

$$S \rightarrow NP VP = 1 \cdot 0,5 + 1 \cdot 0,5 + 1 \cdot 1 = 2$$

$$NP \rightarrow N = 1 \cdot 0,5 + 1 \cdot 0,5 + 2 \cdot 1 = 3$$

$$NP \rightarrow NP PP = 1 \cdot 0,5 = 0,5$$

$$V \rightarrow like = 1 \cdot 0,5 + 0 \cdot 0,5 + 1 \cdot 1 = 1,5$$

$$f_2 (NP \rightarrow N) = 1 \cdot \text{gewicht}(t_1) + 1 \cdot \text{gewicht}(t_2) + 2 \cdot \text{gewicht}(t_3)$$

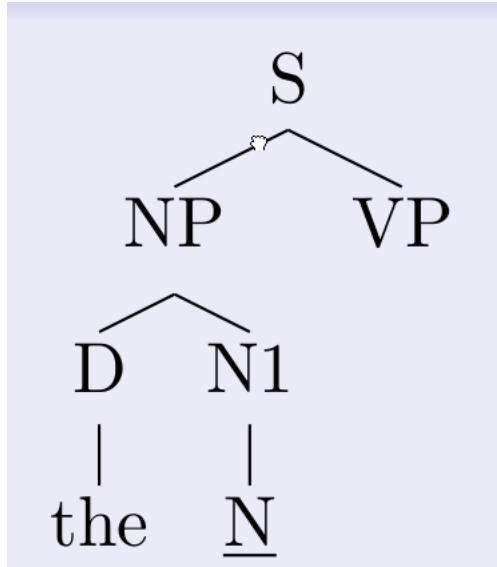
p2

Gegeben folgenden Gewichte:

$$p(t_1 | s) = 0,5$$

$$p(t_2 | s) = 0,5$$

$$p(t_3 | s) = 1$$



1. Schreibe all Regeln von diesem Baum (mit der Reihenfolge der Regeln nach Linksableitung)?

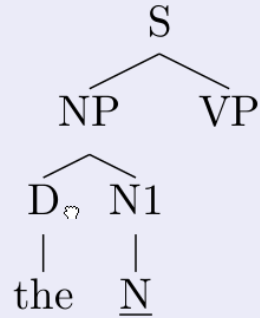
$S \rightarrow NP VP,$   
 $NP \rightarrow D N1,$   
 $D \rightarrow the,$   
 $N1 \rightarrow N$

2.  $P(T)$  für dieses Beispiel aufschreiben

$$p(T) = p(r_1, \dots, r_n) = \prod_{i=1}^n p(r_i)$$

$$p(T) = p(S \rightarrow NP VP) * p(NP \rightarrow D N1) * p(D \rightarrow the) * p(N1 \rightarrow N)$$

$S \rightarrow NP VP$   
 $NP \rightarrow D N1$   
 $D \rightarrow the$   
 $N1 \rightarrow N$



1. Schreibe all Regeln von diesem Baum (mit der Reihenfolge der Regeln nach Linksableitung)?

2.  $P(T)$  für dieses Beispiel aufschreiben

$$p(T) = p(r_1, \dots, r_n) = \prod_{i=1}^n p(r_i)$$

$p(s \rightarrow NP VP,$   
 $NP \rightarrow D N1,$   
 $D \rightarrow the,$   
 $N1 \rightarrow N) = p(\dots$

Wozu brauchen wir EM-Training in syntaktische Desambiguierung (Um was zu berechnen) ?

$p(\text{regel})$  zu berechnen. Parameter zu schätzen

Um die Wortarten zu annotieren? (HMM)  
um die beste **Analyse** zu finden?

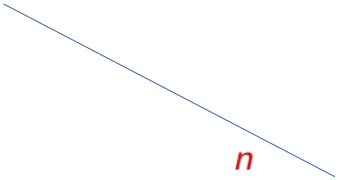
Wen man **die wahrscheinlichste Analyse** eines Satzes haben will ohne Zugriff auf Baumbank

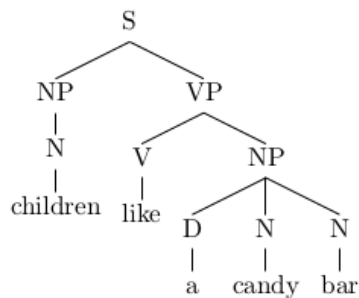
$$p(T) = p(r_1, \dots, r_n) = \prod_{i=1}^n p(r_i)$$

Wozu brauchen wir EM-Training in syntaktische Desambiguierung (Um was zu berechnen) ?

## EM-Training

- 1 Initialisierung der Regelwahrscheinlichkeiten
- 2 Berechnung der Parsebaumgewichte  $p(t|s)$
- 3 Extraktion der gewichteten Regelhäufigkeiten
- 4 Neuschätzung der Regelwahrscheinlichkeiten
- 5 Weiter mit Schritt 2


$$p(T) = p(r_1, \dots, r_n) = \prod_{i=1}^n p(r_i)$$

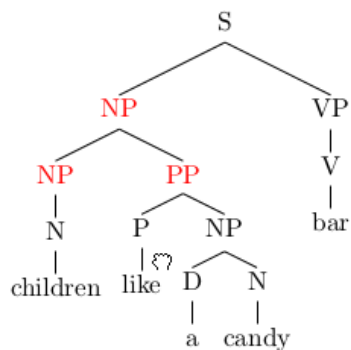


Führe den 1. Schritt durch

$p(S \rightarrow NP VP)$

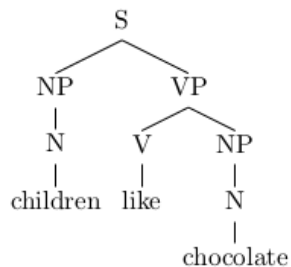
$p(NP \rightarrow N)$

$p(V \rightarrow like)$



EM-Training

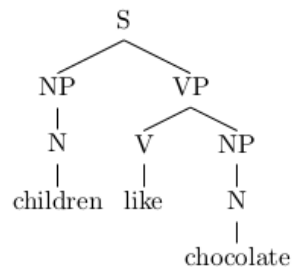
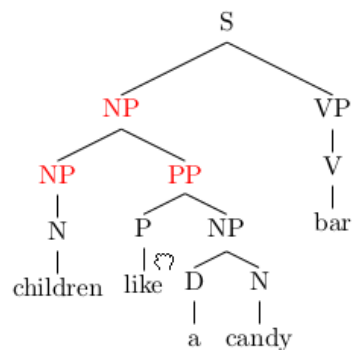
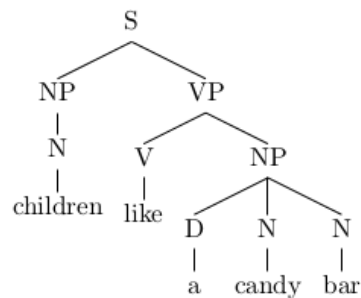
- ① Initialisierung der Regelwahrscheinlichkeiten
- ② Berechnung der Parsebaumgewichte  $p(t|s)$
- ③ Extraktion der gewichteten Regelhäufigkeiten
- ④ Neuschätzung der Regelwahrscheinlichkeiten
- ⑤ Weiter mit Schritt 2





# Führe den 1. Schritt durch

## 1 Initialisierung der Regelwahrscheinlichkeiten



Regel	$p_0$
$S \rightarrow NP VP$	1.00
$NP \rightarrow D N$	0.25
$NP \rightarrow D N N$	0.25
$NP \rightarrow N$	0.25
$NP \rightarrow NP PP$	0.25
$VP \rightarrow V$	0.50
$VP \rightarrow V NP$	0.50
$PP \rightarrow P NP$	1.00
$D \rightarrow a$	0.50
$D \rightarrow the$	0.50
$N \rightarrow bar$	0.25
$N \rightarrow candy$	0.25
$N \rightarrow children$	0.25
$N \rightarrow chocolate$	0.25
$V \rightarrow bar$	0.50
$V \rightarrow like$	0.50
$P \rightarrow like$	1.00

$p_{init}(regel) = 1 / \text{Anzahl der verschiedenen Regel}$

$p(S \rightarrow NP VP)$

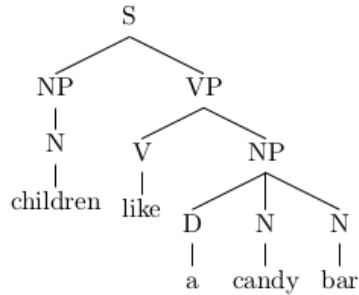
$p(NP \rightarrow N)$

$p(V \rightarrow like)$

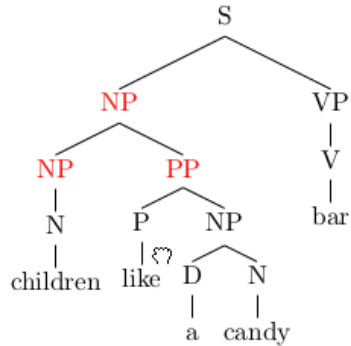
## Führe den 2. Schritt durch

### EM-Training

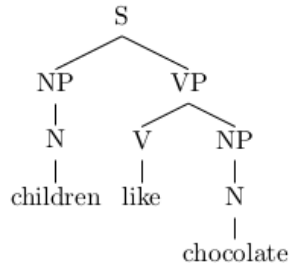
t1



t2



t3



- 1 Initialisierung der Regelwahrscheinlichkeiten
- 2 Berechnung der Parsebaumgewichte  $p(t|s)$
- 3 Extraktion der gewichteten Regelhäufigkeiten
- 4 Neuschätzung der Regelwahrscheinlichkeiten
- 5 Weiter mit Schritt 2

$$p(T) = p(r_1, \dots, r_n) = \prod_{i=1}^n p(r_i)$$

$$p(t_1 | s) =$$

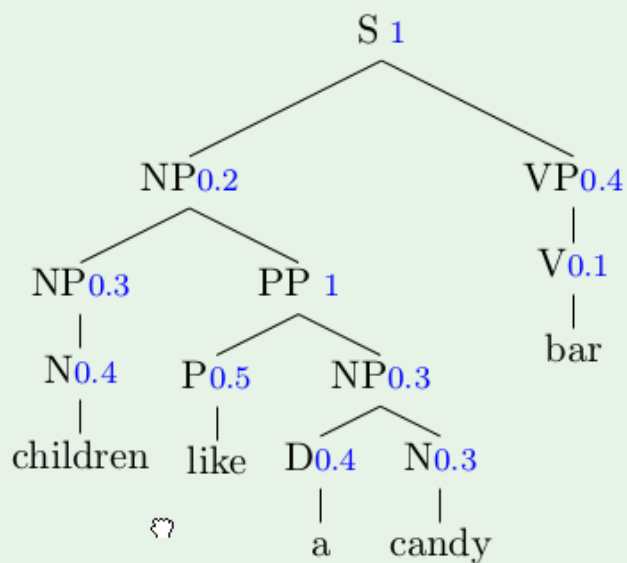
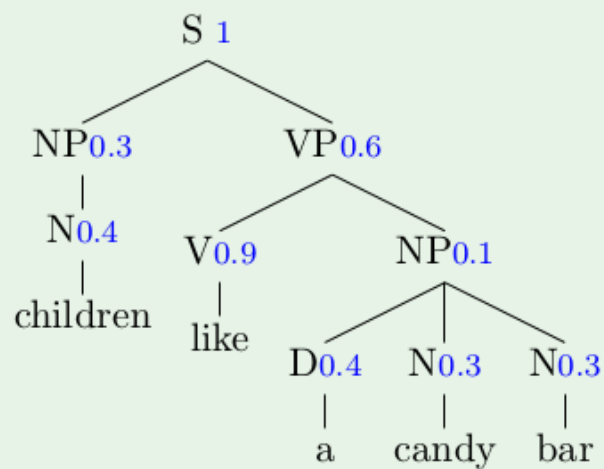
$$p(t_2 | s) =$$

$$p(t_3 | s) =$$

\* der Satz s kann verschieden sein

$$\frac{p(t)}{\sum_{t' \in T(s)} p(t')}$$

Regel	$p_0$
$S \rightarrow NP VP$	1.00
$NP \rightarrow D N$	0.25
$NP \rightarrow D N N$	0.25
$NP \rightarrow N$	0.25
$NP \rightarrow NP PP$	0.25
$VP \rightarrow V$	0.50
$VP \rightarrow V NP$	0.50
$PP \rightarrow P NP$	1.00
$D \rightarrow a$	0.50
$D \rightarrow the$	0.50
$N \rightarrow bar$	0.25
$N \rightarrow candy$	0.25
$N \rightarrow children$	0.25
$N \rightarrow chocolate$	0.25
$V \rightarrow bar$	0.50
$V \rightarrow like$	0.50
$P \rightarrow like$	1.00



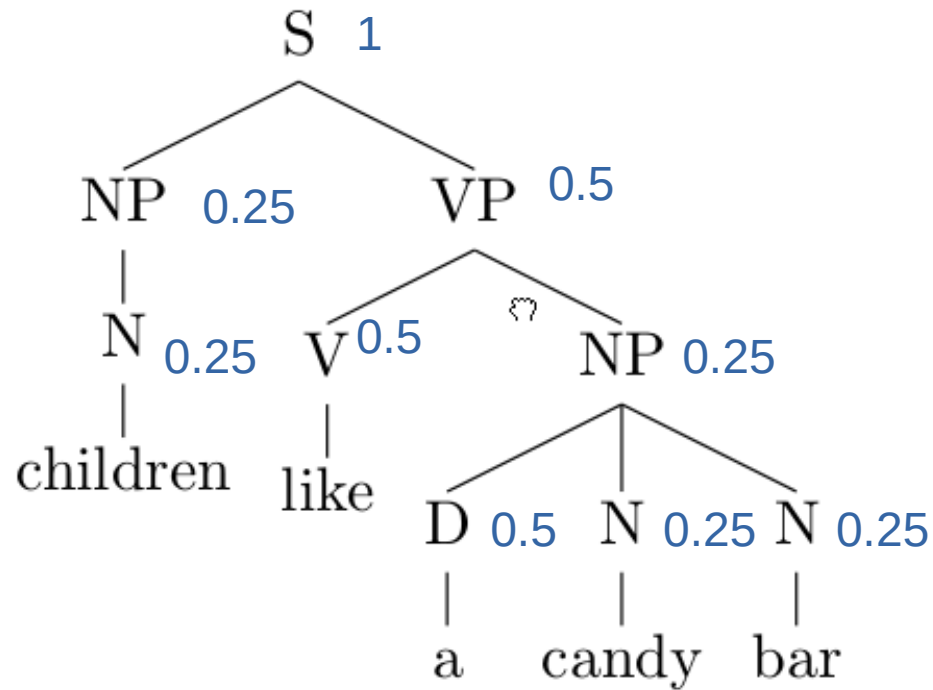
$$p(t_1) = 0.0002333$$

$$p(t_1|s) = \frac{p(t_1)}{p(t_1)+p(t_2)} = 0.93$$

$$p(t_2) = 0.0000173$$

$$p(t_2|s) = \frac{p(t_2)}{p(t_1)+p(t_2)} = 0.07$$

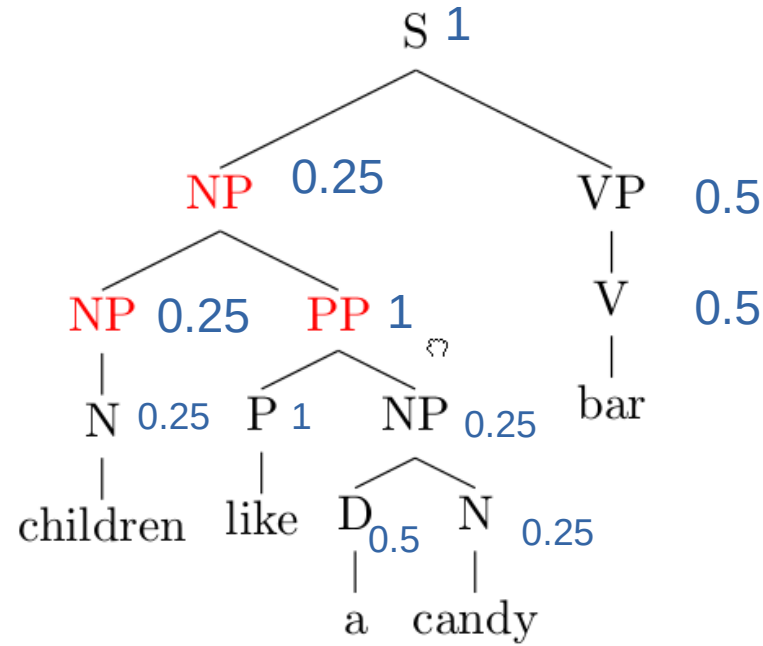
## Führe den 2. Schritt durch



Regel	$p_0$
$S \rightarrow NP VP$	1.00
$NP \rightarrow D N$	0.25
$NP \rightarrow D N N$	0.25
$NP \rightarrow N$	0.25
$NP \rightarrow NP PP$	0.25
$VP \rightarrow V$	0.50
$VP \rightarrow V NP$	0.50
$PP \rightarrow P NP$	1.00
$D \rightarrow a$	0.50
$D \rightarrow the$	0.50
$N \rightarrow bar$	0.25
$N \rightarrow candy$	0.25
$N \rightarrow children$	0.25
$N \rightarrow chocolate$	0.25
$V \rightarrow bar$	0.50
$V \rightarrow like$	0.50
$P \rightarrow like$	1.00

$$p(t_1) = 1 * 0,25 * 0,25 * 0,5 * 0,5 * 0,25 * 0,5 * 0,25 * 0,25$$

$$= 0,00012207$$



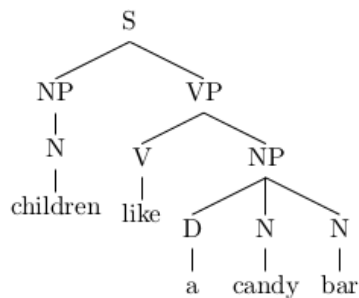
$$p(t_2) = (0,25 \wedge 5) * (0,5 \wedge 3) = 0,00012207$$

Regel	$p_0$
$S \rightarrow NP VP$	1.00
$NP \rightarrow D N$	0.25
$NP \rightarrow D N N$	0.25
$NP \rightarrow N$	0.25
$NP \rightarrow NP PP$	0.25
$VP \rightarrow V$	0.50
$VP \rightarrow V NP$	0.50
$PP \rightarrow P NP$	1.00
$D \rightarrow a$	0.50
$D \rightarrow the$	0.50
$N \rightarrow bar$	0.25
$N \rightarrow candy$	0.25
$N \rightarrow children$	0.25
$N \rightarrow chocolate$	0.25
$V \rightarrow bar$	0.50
$V \rightarrow like$	0.50
$P \rightarrow like$	1.00

$$p(t1 | s1) = 0,00012207 / 0,00012207 + 0,00012207 = 0,5$$

$$p(t2 | s1 ) = 0,00012207 / 0,00012207 + 0,00012207 = 0,5$$

$$p(t1|s2) = \text{nummer} / \text{nummer} = 1$$



Führe den 3. Schritt durch

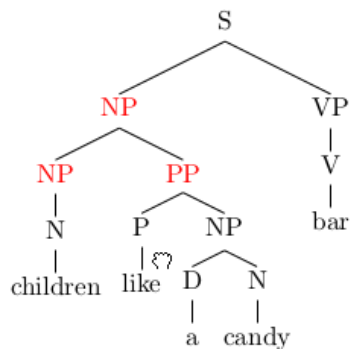
$S \rightarrow NP VP$

$$f(S \rightarrow NP VP) = 1 \cdot 0,5 + 1 \cdot 0,5 + 1 \cdot 1 = 2$$

$$\text{gewicht}(t1) = 0,5$$

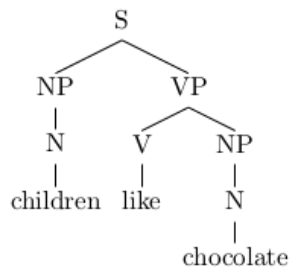
$$\text{gewicht}(t2) = 0,5$$

$$\text{gewicht}(t3) = 1$$



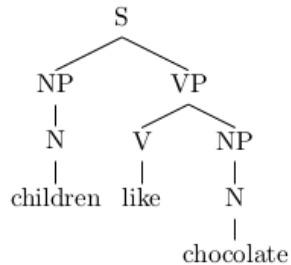
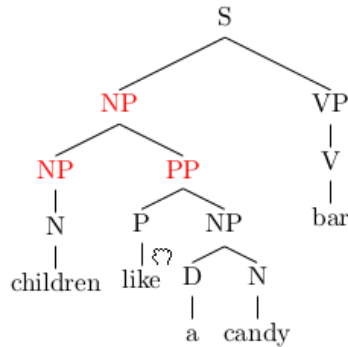
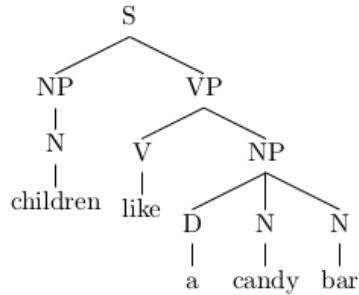
## EM-Training

- 1 Initialisierung der Regelwahrscheinlichkeiten
- 2 Berechnung der Parsebaumgewichte  $p(t|s)$
- 3 Extraktion der gewichteten Regelhäufigkeiten
- 4 Neuschätzung der Regelwahrscheinlichkeiten
- 5 Weiter mit Schritt 2



## Führe den 3. Schritt durch

### 3 Extraktion der gewichteten Regelhäufigkeiten



Regel	$p_0$	$f_1$
$S \rightarrow NP VP$	1.00	2.00
$NP \rightarrow D N$	0.25	0.50
$NP \rightarrow D N N$	0.25	0.50
$NP \rightarrow N$	0.25	3.00
$NP \rightarrow NP PP$	0.25	0.50
$VP \rightarrow V$	0.50	0.50
$VP \rightarrow V NP$	0.50	1.50
$PP \rightarrow P NP$	1.00	0.50
$D \rightarrow a$	0.50	1.00
$D \rightarrow the$	0.50	0.00
$N \rightarrow bar$	0.25	0.50
$N \rightarrow candy$	0.25	1.00
$N \rightarrow children$	0.25	2.00
$N \rightarrow chocolate$	0.25	1.00
$V \rightarrow bar$	0.50	0.50
$V \rightarrow like$	0.50	1.50
$P \rightarrow like$	1.00	0.50

Wie bekommen wir  $f_1$  für folgende Regeln?

$S \rightarrow NP VP$

$NP \rightarrow N = 1 \cdot 0,5 + 1 \cdot 0,5 + 2 \cdot 1$   
 $= 3$

$NP \rightarrow NP PP$

$= 1 \cdot 0,5 = 0,5$

$V \rightarrow like$

$= 1 \cdot 0,5 + 1 \cdot 1 = 1,5$

Gegeben folgenden

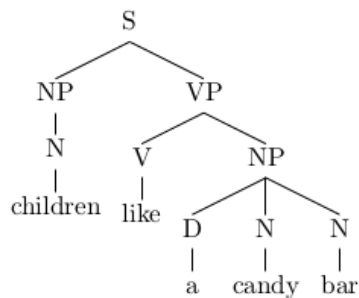
Gewichte:

$p(t_1 | s) = 0,5$

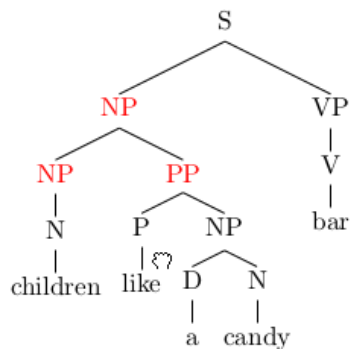
$p(t_2 | s) = 0,5$

$p(t_3 | s) = 1$



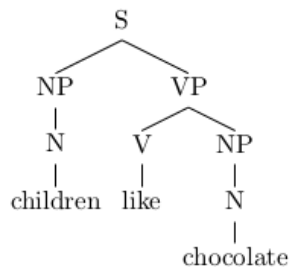


Führe den 4. Schritt durch

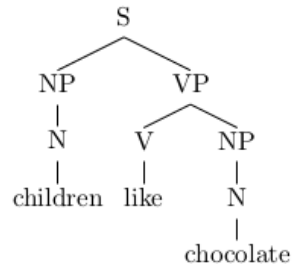
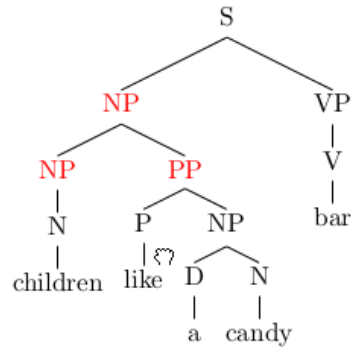
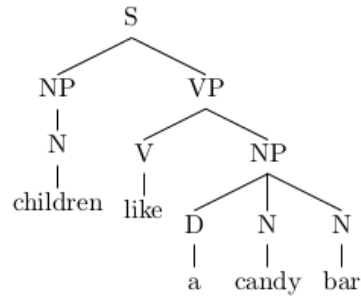


EM-Training

- ① Initialisierung der Regelwahrscheinlichkeiten
- ② Berechnung der Parsebaumgewichte  $p(t|s)$
- ③ Extraktion der gewichteten Regelhäufigkeiten
- ④ Neuschätzung der Regelwahrscheinlichkeiten
- ⑤ Weiter mit Schritt 2



Führe den 4. Schritt durch



#### 4 Neuschätzung der Regelwahrscheinlichkeiten

Regel	$p_0$	$f_1$
$S \rightarrow NP VP$	1.00	2.00
$NP \rightarrow D N$	0.25	0.50
$NP \rightarrow D N N$	0.25	0.50
$NP \rightarrow N$	0.25	3.00
$NP \rightarrow NP PP$	0.25	0.50
$VP \rightarrow V$	0.50	0.50
$VP \rightarrow V NP$	0.50	1.50
$PP \rightarrow P NP$	1.00	0.50
$D \rightarrow a$	0.50	1.00
$D \rightarrow the$	0.50	0.00
$N \rightarrow bar$	0.25	0.50
$N \rightarrow candy$	0.25	1.00
$N \rightarrow children$	0.25	2.00
$N \rightarrow chocolate$	0.25	1.00
$V \rightarrow bar$	0.50	0.50
$V \rightarrow like$	0.50	1.50
$P \rightarrow like$	1.00	0.50

Wie bekommen wir  $p_1$  für folgenden Regeln?

$$S \rightarrow NP VP = f_1(S \rightarrow NP VP) / f_1(S \rightarrow NP VP) = 1$$

$$NP \rightarrow N = f(NP \rightarrow N) / f(NP \rightarrow N) + f(NP \rightarrow D N N) + f(NP \rightarrow D N N N) + f(NP \rightarrow NP PP)$$

$$NP \rightarrow NP PP =$$

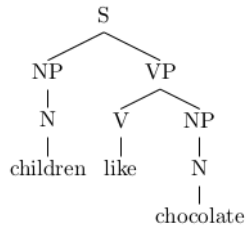
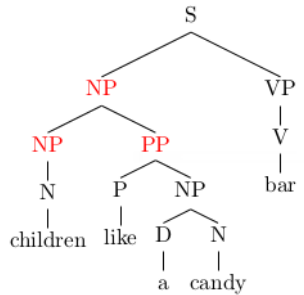
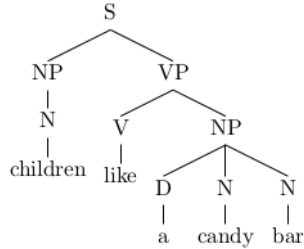
$$V \rightarrow like =$$

$$p(A \rightarrow \alpha) = \frac{f_{A \rightarrow \alpha}}{\sum_{\beta} f_{A \rightarrow \beta}}$$

$$p(N \rightarrow P NP) = p(N \rightarrow P NP) / p(N \rightarrow ..) + p(N \rightarrow ...) + ...$$

## Führe den 4. Schritt durch

### 4 Neuschätzung der Regelwahrscheinlichkeiten



Regel	$p_0$	$f_1$	$p_1$
$S \rightarrow NP VP$	1.00	2.00	1.00
$NP \rightarrow D N$	0.25	0.50	0.11
$NP \rightarrow D N N$	0.25	0.50	0.11
$NP \rightarrow N$	0.25	3.00	0.67
$NP \rightarrow NP PP$	0.25	0.50	0.11
$VP \rightarrow V$	0.50	0.50	0.25
$VP \rightarrow V NP$	0.50	1.50	0.75
$PP \rightarrow P NP$	1.00	0.50	1.00
$D \rightarrow a$	0.50	1.00	1.00
$D \rightarrow the$	0.50	0.00	0.00
$N \rightarrow bar$	0.25	0.50	0.11
$N \rightarrow candy$	0.25	1.00	0.22
$N \rightarrow children$	0.25	2.00	0.44
$N \rightarrow chocolate$	0.25	1.00	0.22
$V \rightarrow bar$	0.50	0.50	0.25
$V \rightarrow like$	0.50	1.50	0.75
$P \rightarrow like$	1.00	0.50	1.00
-logprob		15.2	

Wie bekommen wir  $p_1$  für folgenden Regeln?

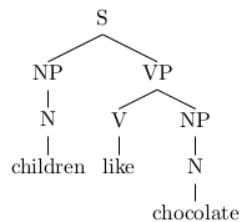
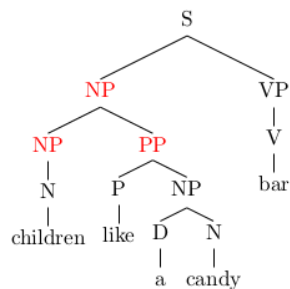
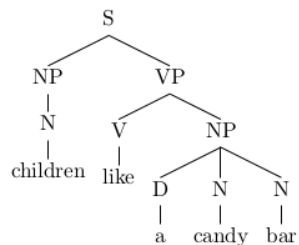
$$S \rightarrow NP VP = 2 / 2 = 1$$

$$NP \rightarrow N = 3 / (0,5 + 0,5 + 3 + 0,5) = 0,67$$

$$NP \rightarrow NP PP = 0,5 / (0,5 + 0,5 + 3 + 0,5) = 0,11$$

$$V \rightarrow like = 1,5 / (1,5 + 0,5) = 0,75$$

$$p(A \rightarrow \alpha) = \frac{f_{A \rightarrow \alpha}}{\sum_{\beta} f_{A \rightarrow \beta}}$$

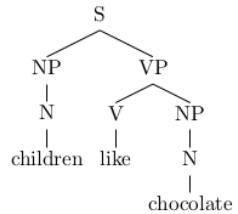
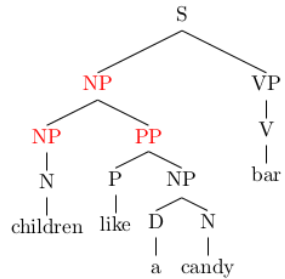
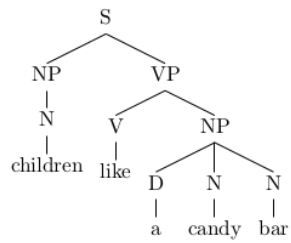


Regel	$p_0$	$f_1$	$p_1$	$f_2$	$p_2$	$f_3$	$p_3$
$S \rightarrow NP VP$	1.00	2.00	1.00	2.00	1.00	2.00	1.00
$NP \rightarrow D N$	0.25	0.50	0.11	0.10	0.02	0.00	0.00
$NP \rightarrow D N N$	0.25	0.50	0.11	0.90	0.22	1.00	0.25
$NP \rightarrow N$	0.25	3.00	0.67	3.00	0.73	3.00	0.75
$NP \rightarrow NP PP$	0.25	0.50	0.11	0.10	0.02	0.00	0.00
$VP \rightarrow V$	0.50	0.50	0.25	0.10	0.05	0.00	0.00
$VP \rightarrow V NP$	0.50	1.50	0.75	1.90	0.95	2.00	1.00
$PP \rightarrow P NP$	1.00	0.50	1.00	0.10	1.00	0.00	1.00
$D \rightarrow a$	0.50	1.00	1.00	1.00	1.00	1.00	1.00
$D \rightarrow the$	0.50	0.00	0.00	0.00	0.00	0.00	0.00
$N \rightarrow bar$	0.25	0.50	0.11	0.90	0.18	1.00	0.20
$N \rightarrow candy$	0.25	1.00	0.22	1.00	0.20	1.00	0.20
$N \rightarrow children$	0.25	2.00	0.44	2.00	0.41	2.00	0.40
$N \rightarrow chocolate$	0.25	1.00	0.22	1.00	0.20	1.00	0.20
$V \rightarrow bar$	0.50	0.50	0.25	0.10	0.05	0.00	0.00
$V \rightarrow like$	0.50	1.50	0.75	1.90	0.95	2.00	1.00
$P \rightarrow like$	1.00	0.50	1.00	0.10	1.00	0.00	1.00
-logprob		15.2		11.3		9.3	

$p(\text{regel}) =$   
 $\frac{\text{erwarte\_f}(\text{regel})}{\text{Sum erwartet\_f}(\text{regel mit gleiche symbol})}$

Wie wird  $f_2(NP \rightarrow NP PP)$  berechnet ?

$= f(NP \rightarrow NP PP) * \text{gewicht}(t_2 \text{ geschätzt aus } p_1) = 1 * \text{gewicht}(t_2) = 0,10$



Regel	$p_0$	$f_1$	$p_1$	$f_2$	$p_2$	$f_3$	$p_3$
$S \rightarrow NP VP$	1.00	2.00	1.00	2.00	1.00	2.00	1.00
$NP \rightarrow D N$	0.25	0.50	0.11	0.10	0.02	0.00	0.00
$NP \rightarrow D N N$	0.25	0.50	0.11	0.90	0.22	1.00	0.25
$NP \rightarrow N$	0.25	3.00	0.67	3.00	0.73	3.00	0.75
$NP \rightarrow NP PP$	0.25	0.50	0.11	0.10	0.02	0.00	0.00
$VP \rightarrow V$	0.50	0.50	0.25	0.10	0.05	0.00	0.00
$VP \rightarrow V NP$	0.50	1.50	0.75	1.90	0.95	2.00	1.00
$PP \rightarrow P NP$	1.00	0.50	1.00	0.10	1.00	0.00	1.00
$D \rightarrow a$	0.50	1.00	1.00	1.00	1.00	1.00	1.00
$D \rightarrow the$	0.50	0.00	0.00	0.00	0.00	0.00	0.00
$N \rightarrow bar$	0.25	0.50	0.11	0.90	0.18	1.00	0.20
$N \rightarrow candy$	0.25	1.00	0.22	1.00	0.20	1.00	0.20
$N \rightarrow children$	0.25	2.00	0.44	2.00	0.41	2.00	0.40
$N \rightarrow chocolate$	0.25	1.00	0.22	1.00	0.20	1.00	0.20
$V \rightarrow bar$	0.50	0.50	0.25	0.10	0.05	0.00	0.00
$V \rightarrow like$	0.50	1.50	0.75	1.90	0.95	2.00	1.00
$P \rightarrow like$	1.00	0.50	1.00	0.10	1.00	0.00	1.00
-logprob		15.2		11.3		9.3	

Wie wird  $f_2$  ( $NP \rightarrow NP PP$ ) berechnet ?

$f_2(NP \rightarrow NP PP) = f(NP \rightarrow NP PP)_{t2} * \text{Gewicht}(t2)$

$f_2(VP \rightarrow V NP) = f(VP \rightarrow V NP)_{t1} * \text{Gewicht}(t1) + f(VP \rightarrow V NP)_{t3} * \text{Gewicht}(t3)$

Welche Methode wird in **überwachtes Training** verwendet?  
Wie funktioniert das (wie wird  $p(\text{regel})$  geschätzt)?

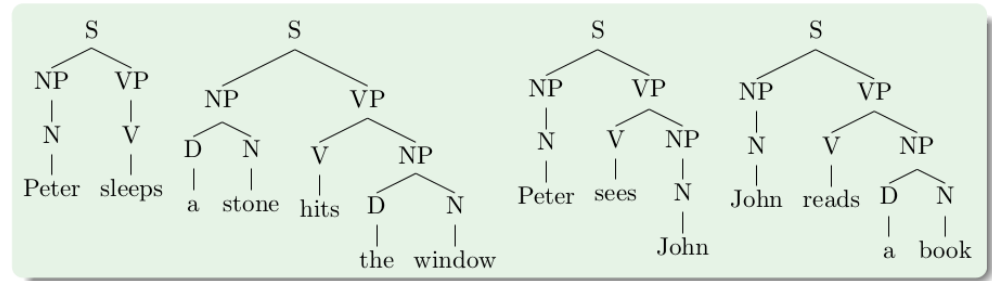
# Welche Methode wird in überwachtes Training verwendet? Wie funktioniert das (wie wird $p(\text{regel})$ geschätzt)?

## • Baumbanktraining

- ▶ benötigt eine manuell erstellte Baumbank.
- ▶ Die Regelhäufigkeiten werden gezählt.
- ▶ Die Regelwahrscheinlichkeiten werden mit relativen Häufigkeiten geschätzt:

$$p(A \rightarrow \alpha) = \frac{f_{A \rightarrow \alpha}}{\sum_{\beta} f_{A \rightarrow \beta}}$$

## Baumbank-Training



Extraktion der Grammatikregeln und Regelhäufigkeiten:

S → NP VP	4	1	D → a	2	0.67	N → Peter	2	0.29
VP → V NP	3	0.75	D → the	1	0.33	N → John	2	0.29
VP → V	1	0.25	V → sleeps	1	0.25	N → stone	1	0.14
NP → D N	3	0.43	V → hits	1	0.25	N → window	1	0.14
NP → N	4	0.57	V → sees	1	0.25	N → book	1	0.14
			V → reads	1	0.25			

