

The identification, acquisition and evolution of constructions

or: Data-Oriented Language Learning

or: Probabilistic Construction Grammar

Jelle Zuidema

ILLC

University of Amsterdam

Plan of the talk

1. Linguistic Motivation for Probabilistic Construction Grammar
2. Stochastic Tree Substitution Grammars & DOP
3. Estimation: choosing the right weights
4. Subtrees: twigs, prunes, supertwigs, superprunes
5. Expected Frequency
6. Identifying constructions in a corpus

Probabilistic Linguistics

(Abney 1996, *Statistical Methods and Linguistics*)

- (1) a. #the a are of I word salad?
b. John saw Mary unambiguous?
- (2) a. a hectare is a hundred ares
b. As described in section I paragraph a ...
c. The a paragraph of I is hardly readable.
d. Typhoid Mary
e. the Russia house butler

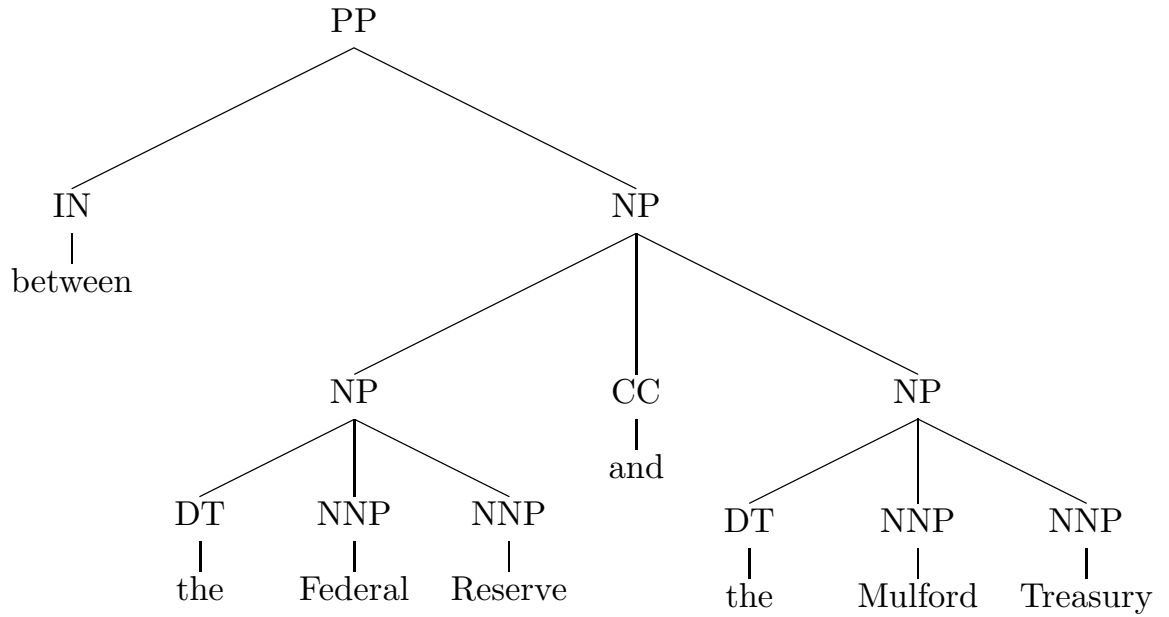
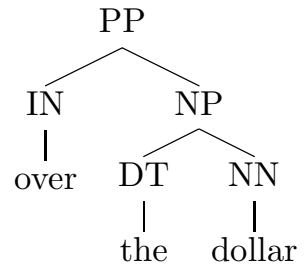
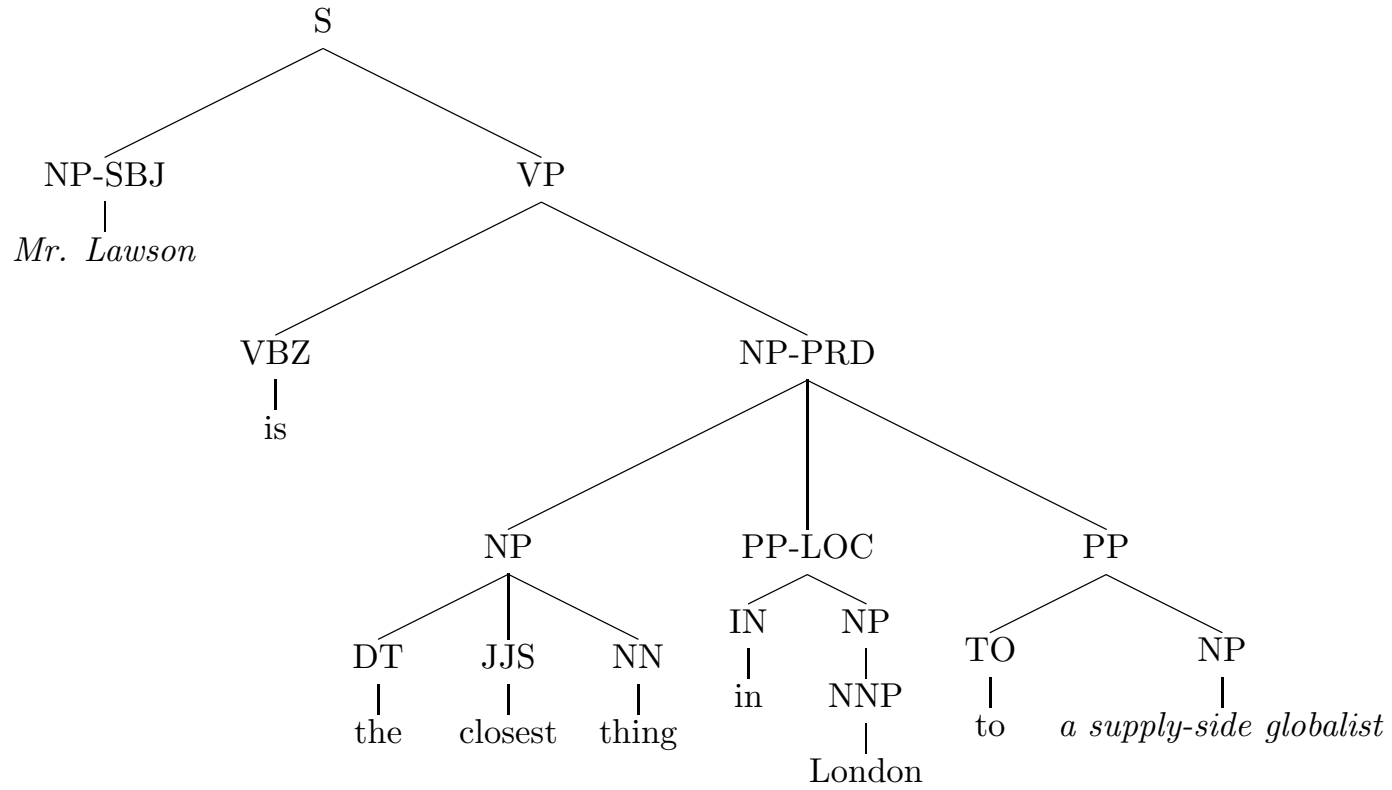
Constructions

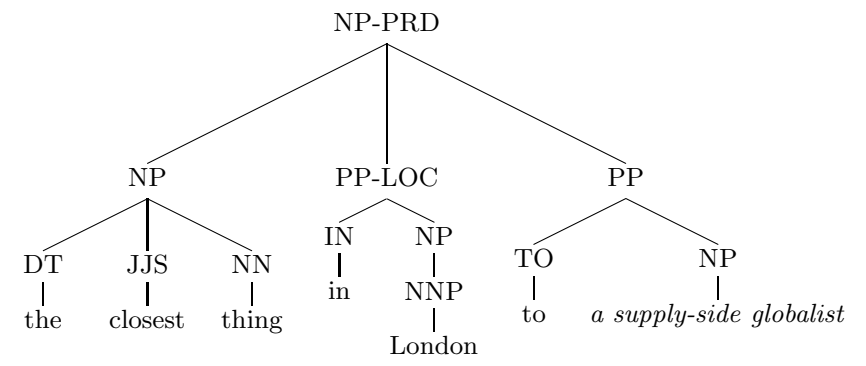
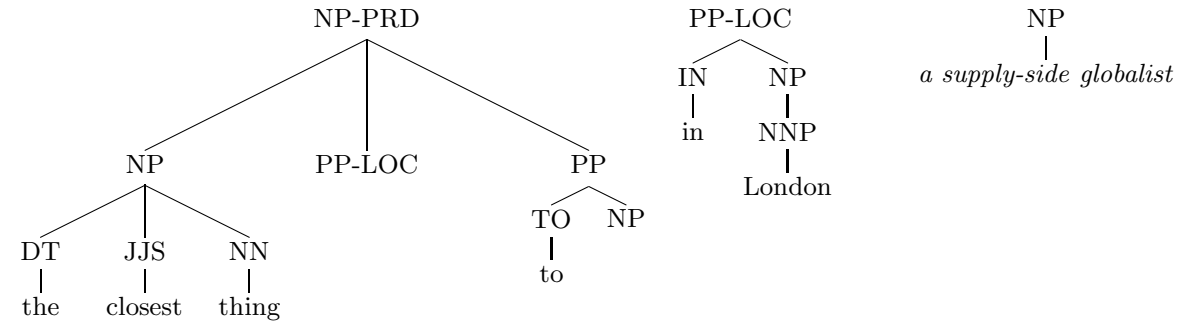
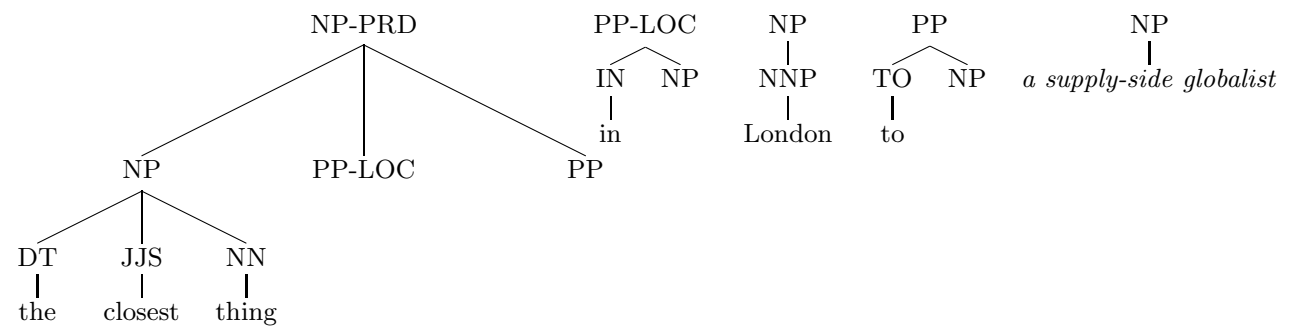
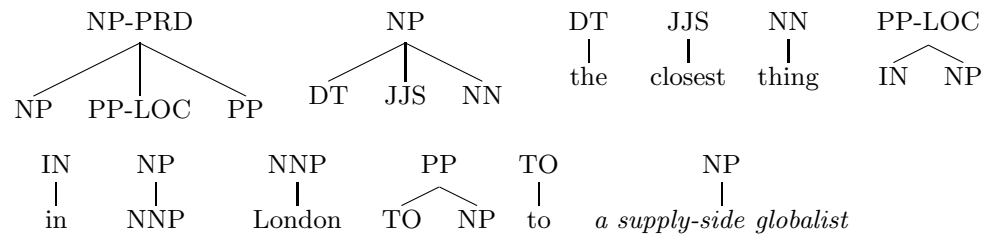
- (3)
 - a. Make one's way through life
 - b. What's X doing to Y
 - c. Construction after construction

- (4)
 - a. What time is it?
 - b. #How late is it?

- (5)
 - a. When is the next train *from Amsterdam to* Paris?
 - b. BA carried *more people than* cargo in 1987.
 - c. Lawson is *the closest thing* in London *to* a supply-side globalist.

(Fillmore, Kay, Goldberg, Jackendoff, Croft, Verhagen; Bod, 1998)





Stochastic Tree Substitution Grammars

An STSG is a 5-tuple $\langle V_n, V_t, S, T, w \rangle$

$$w : T \rightarrow [0, 1], \text{ such that } \forall r \sum_{t:r(t)=r} w(t) = 1$$

The probability of a derivation:

$$P(d = t_1 \circ \dots \circ t_n) = \prod_{i=1}^n (w(t_i))$$

The probability of a parse:

$$P(p) = \sum_{d:\hat{d}=p} (P(d))$$

Stochastic Tree Substitution Grammars

Expected Usage Frequency:

$$u(t) = \sum_{d:t \in d} P(d)$$
$$w(t) = \frac{u(t)}{\sum_{t':r(t')=r(t)} u(t')}$$

Expected Occurrence Frequency:

$$\mathbf{E}[f(t)] = \sum_{p:t \in p^*} (P(p) C(t, p^*)),$$

(where p^* is set of all subtrees of p , and $C(t, p^*) = \#$ occurrences of t in p^*)

Data-Oriented Parsing

(Scha, 1990; Bod, 1993)

Given a corpus of phrase-tree annotated sentences

divided in a train set and a test set

all subtrees of all trees in the train set form the symbolic grammar

with which the test set sentences are parsed.

Data-Oriented Parsing

DOP1 (Bod, 1993, 1998)

$$w(t) = \frac{f(t)}{\sum_{t':r(t')=r(t)} f(t')}$$

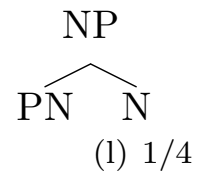
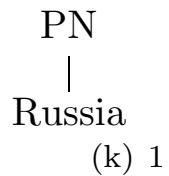
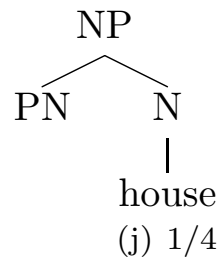
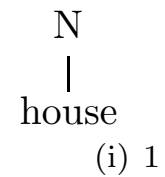
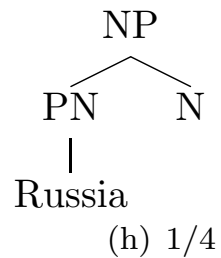
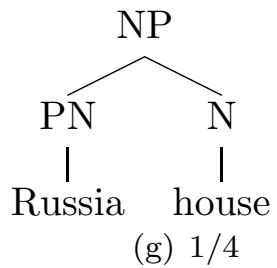
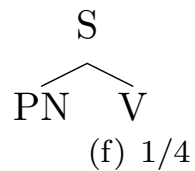
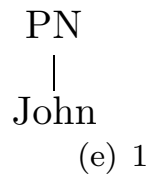
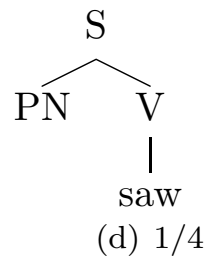
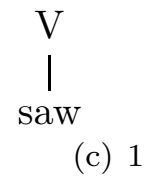
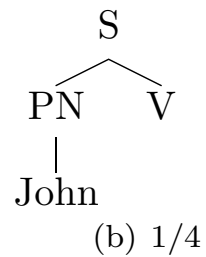
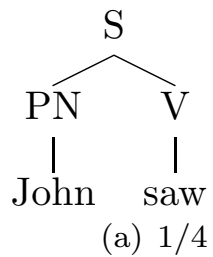
EACL'03 (Bod, 2003)

$$w(t) = \frac{f(t)\alpha}{\sum_{t':r(t')=r(t)} f(t')\alpha}$$

(where α is a scaling factor)

Maximum Probable Parse

Example



Excellent empirical results

E.g. on Wall Street Journal sentences of less than 100 words:

parser	LP	LR	<i>F</i>
Collins '96	.857	.853	.855
Collins '99	.883	.881	.882
Charniak '00	.895	.896	.895
Likelihood-DOP	.897	.895	.896
SL-DOP	.908	.907	.907
Charniak & Johnson '05			.910

LP=labeled precision (# correctly labeled constituents / # labeled constituents)

LR=labeled recall (# correctly labeled constituents / # target constituents)

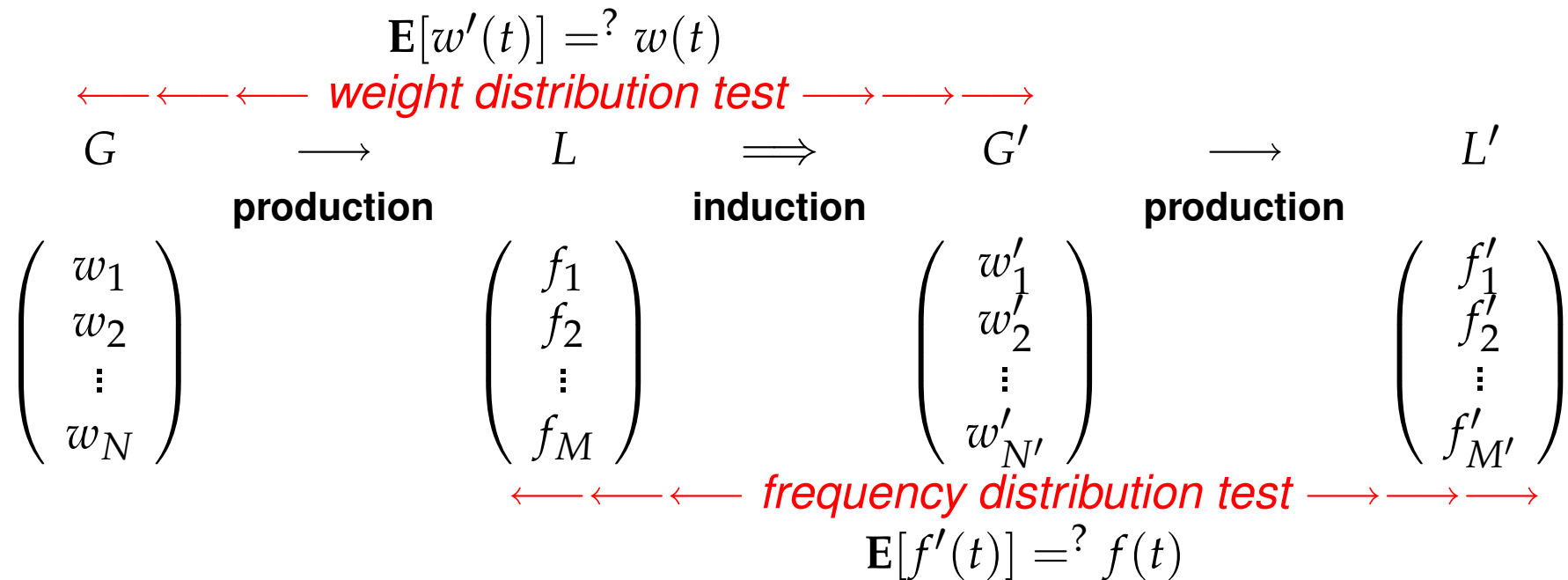
$$F = 2 \cdot LP \cdot LR / (LP + LR) \quad \text{(harmonic mean)}$$

Computational problems

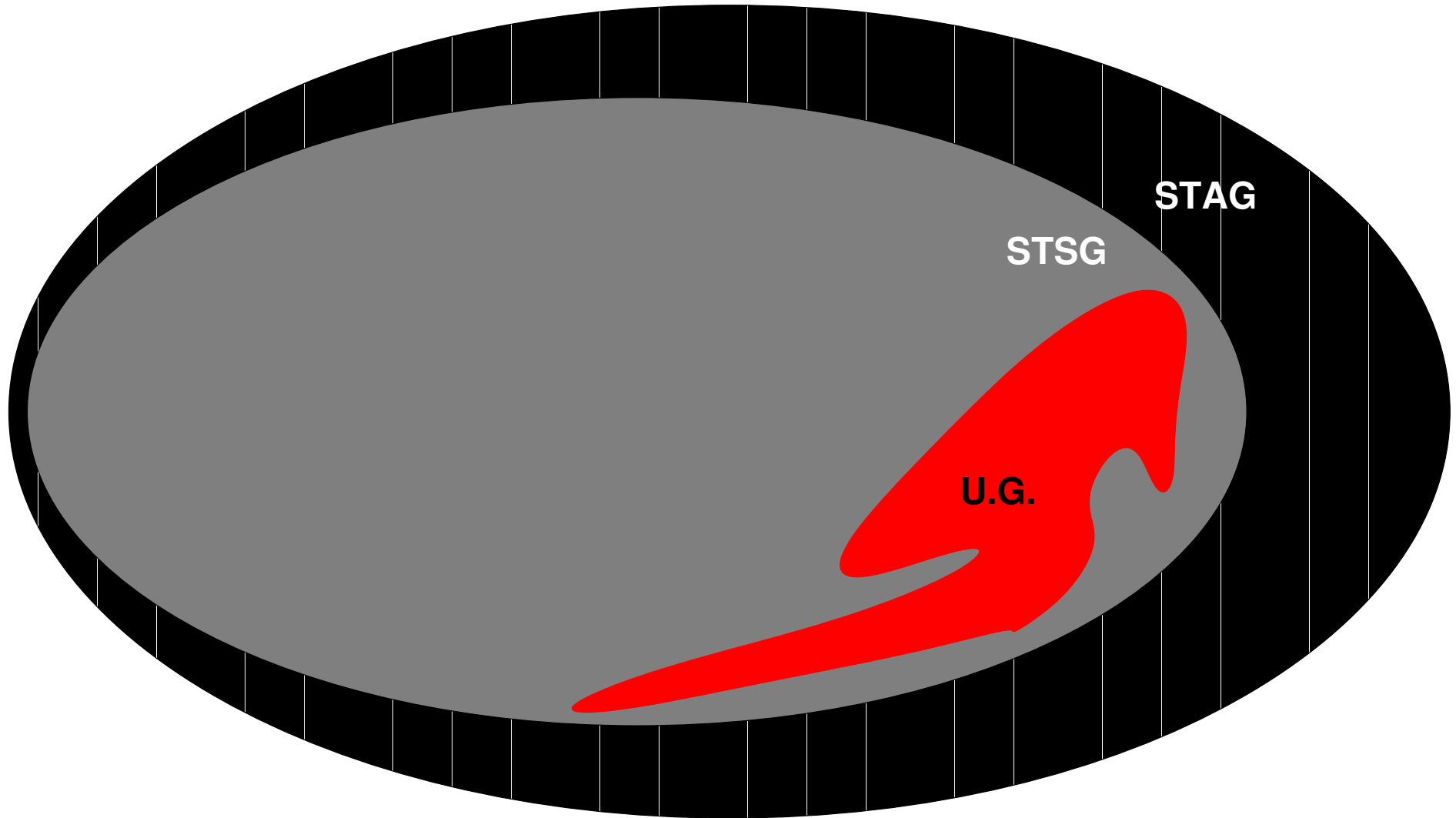
- Time complexity approximations of MPP
- Space complexity PCFG reduction
- Estimation
- Data Annotation

Bias and Inconsistency

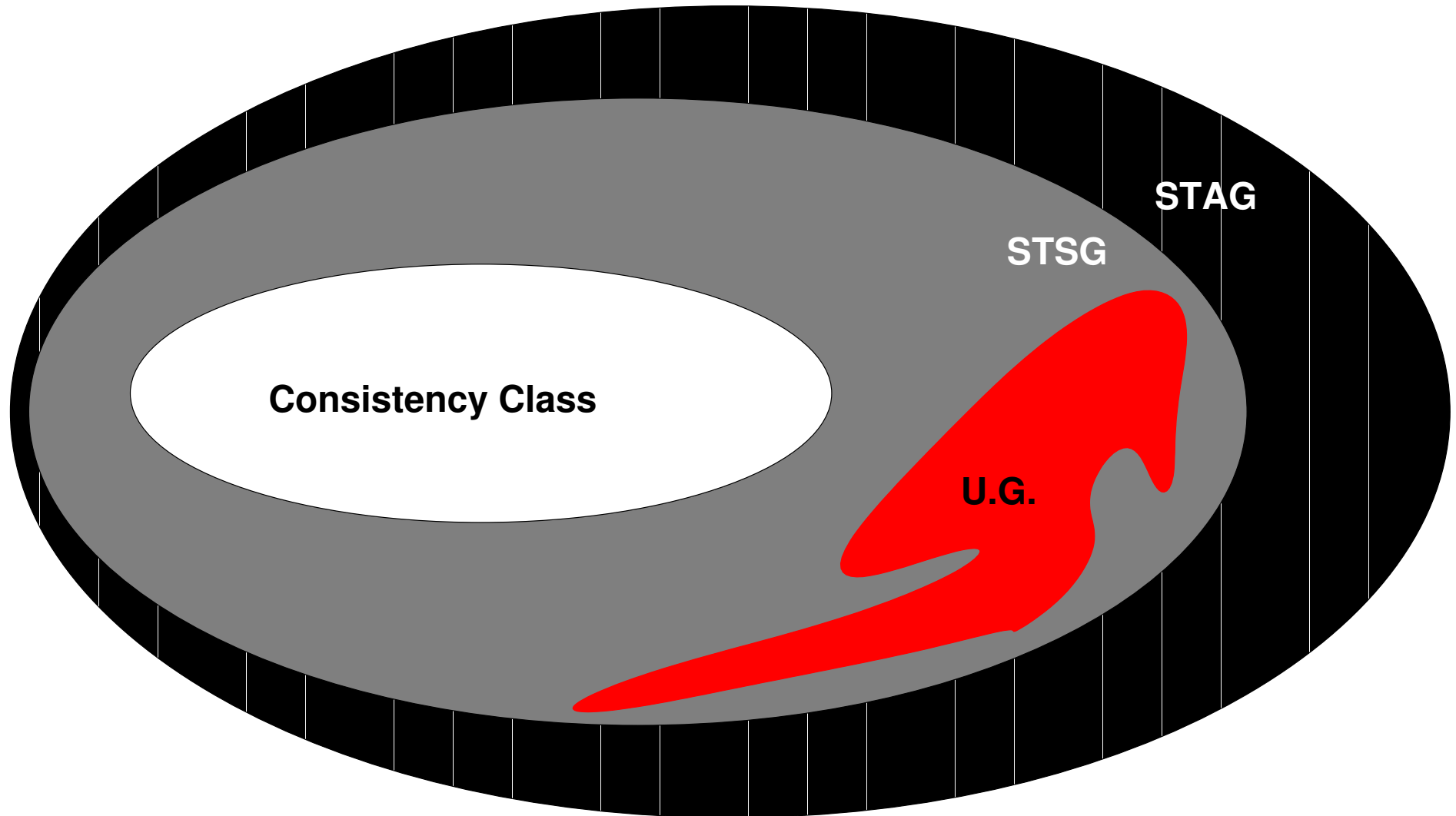
Johnson (2002) presents an example STSG for which the $\mathbf{E}[w'(t)] \neq w(t)$, independent of how much data is seen: DOP1 is *biased* and *inconsistent*.

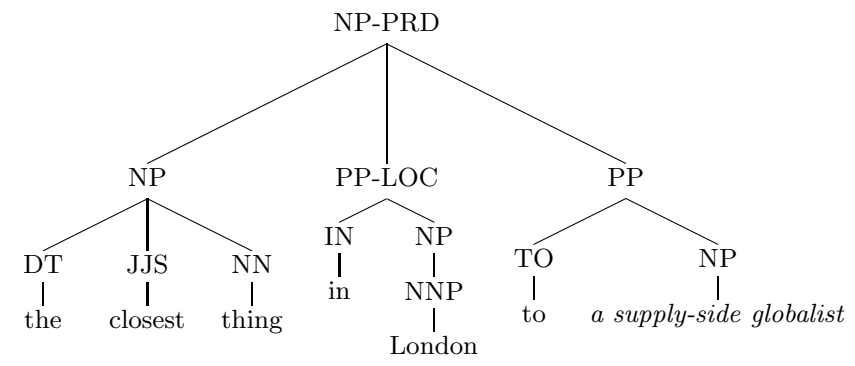
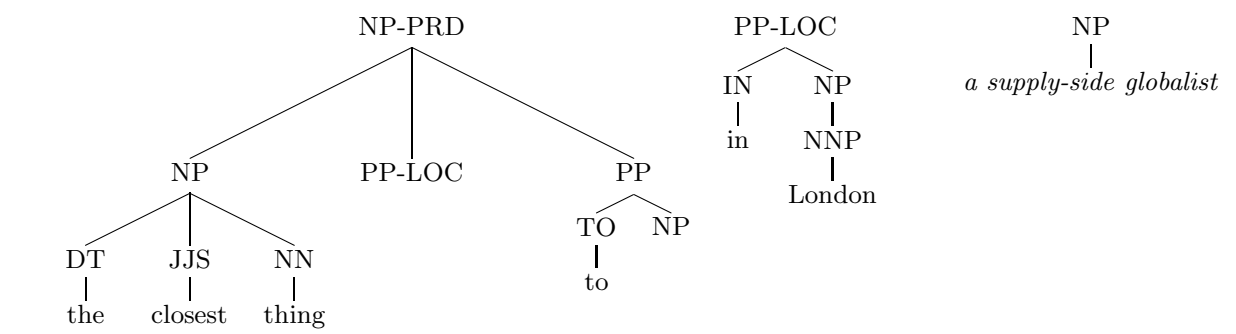
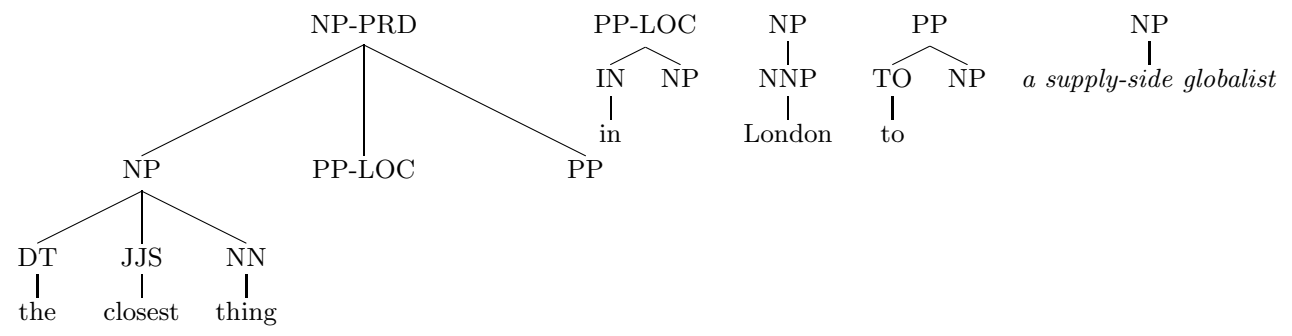
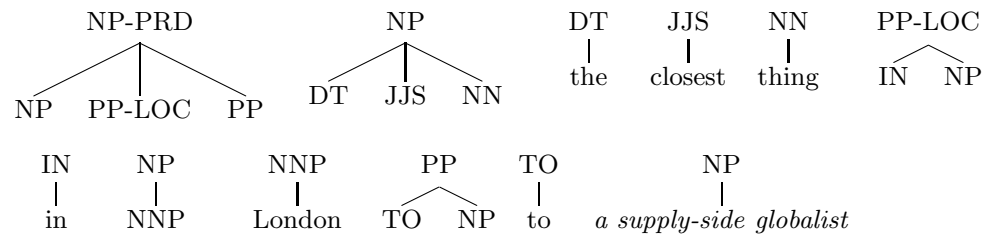


Consistency on a superclass is not necessary



Is the “consistency class” linguistically relevant?



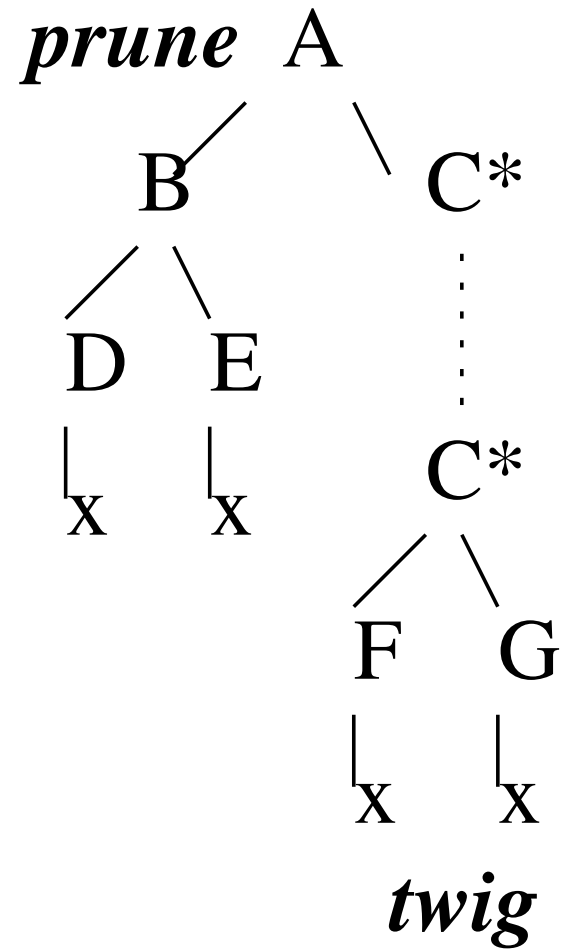
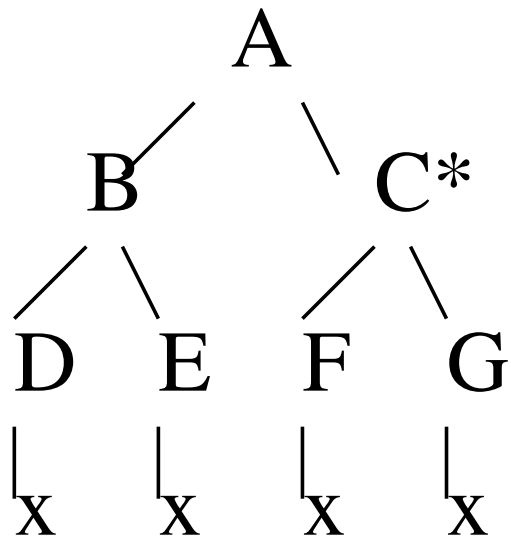


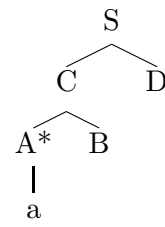
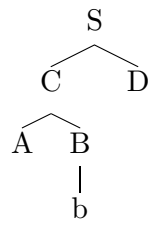
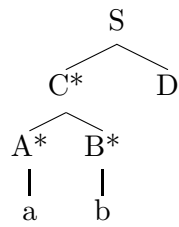
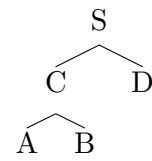
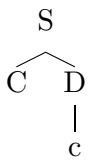
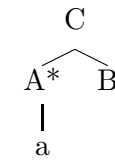
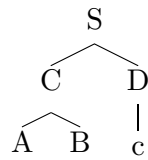
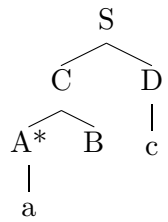
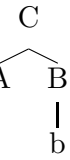
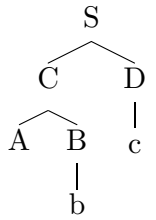
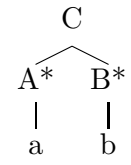
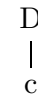
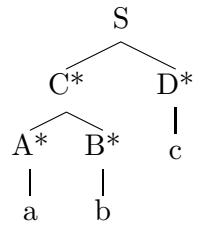
Linguistic desiderata for estimation

- the estimation method converges to the maximally general STSG out of the possibly many correct ones.
- the class of grammars for which the method is consistent in the weight distribution test is linguistically relevant.

Plan of the talk

1. Linguistic Motivation for Probabilistic Construction Grammar
2. Stochastic Tree Substitution Grammars & DOP
3. Estimation: choosing the right weights
4. Subtrees: twigs, prunes, supertwigs, superprunes
5. Expected Frequency
6. Identifying constructions in a corpus





Subtrees, twigs and prunes

the set of twigs:

$$tw(t) = \{t' \mid t' = t \vee \exists t'' (t'' \circ t' = t)\}$$

the set of prunes:

$$pr(t) = \{t' \mid t' = t \vee \exists t'', t''', \dots (t' \circ t'' \circ t''' \circ \dots = t)\}$$

the x-prune:

$pr_x(t)$ = the tree that is created by pruning t at each of the nodes in x

the set of subtrees:

$$st(t) = \{t' \mid \exists t'' (t'' \in tw(t) \wedge t' \in pr(t''))\}.$$

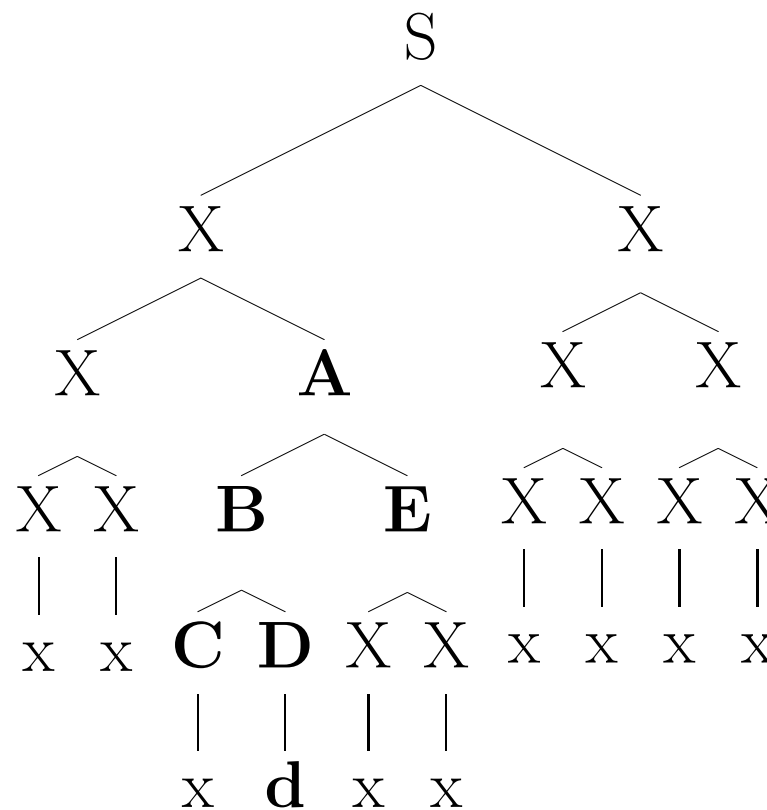
the sets of supertwigs, superprunes and supertrees:

$$\begin{aligned}\widehat{tw}(t) &= \{t' \mid t \in tw(t')\} \\ \widehat{pr}_x(t) &= \{t' \mid t = pr_x(t')\} \\ \widehat{st}(t) &= \{t' \mid t \in st(t')\}\end{aligned}$$

Consider a focal subtree t
 and its derivations $d_1 \circ \dots \circ d_n$

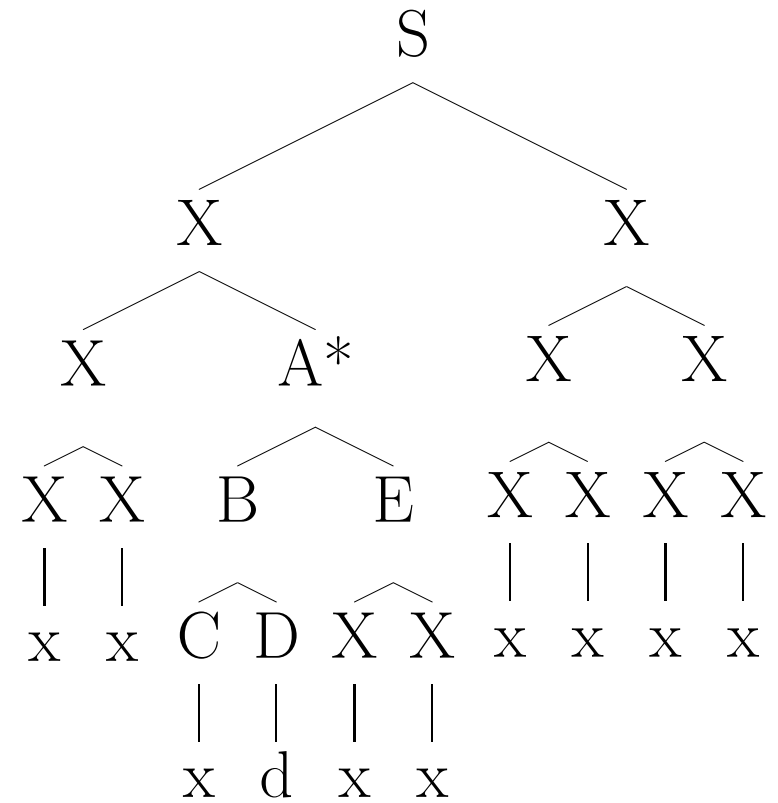
and each of its complete supertrees T
 and its derivations $D_1 \circ \dots \circ D_l$

$$\mathbf{E}[f(t)] = \sum_{p \in T^*} (P(p)C(t, p))$$



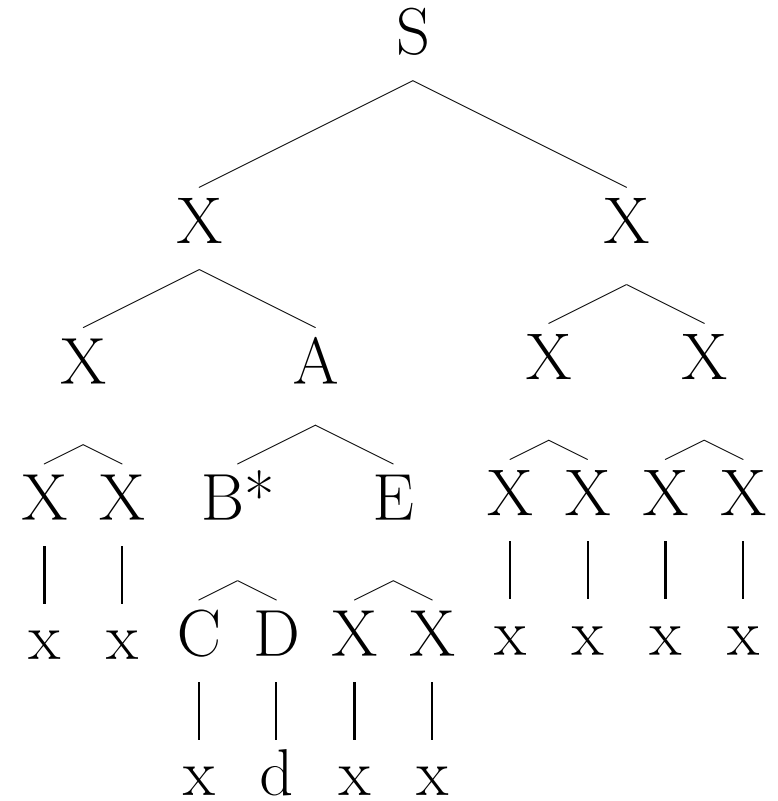
$$D_1 \circ \dots \underbrace{D_k}_{\dots A} \circ \underbrace{D_{k+1}}_{A \dots} \circ \dots \circ D_l$$

$$t = d_1 \circ \dots \circ d_n$$



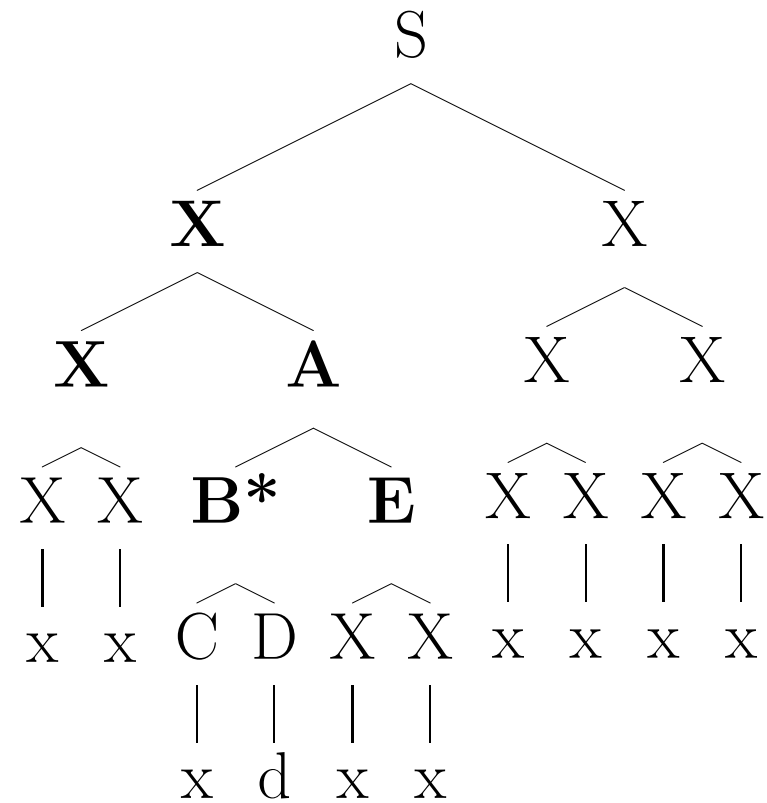
$$D_1 \circ \dots \underbrace{D_k}_{\dots B} \circ \underbrace{D_{k+1}}_{B \dots} \circ \dots \circ D_l$$

$$t = d_1 \circ \dots \circ d_n$$



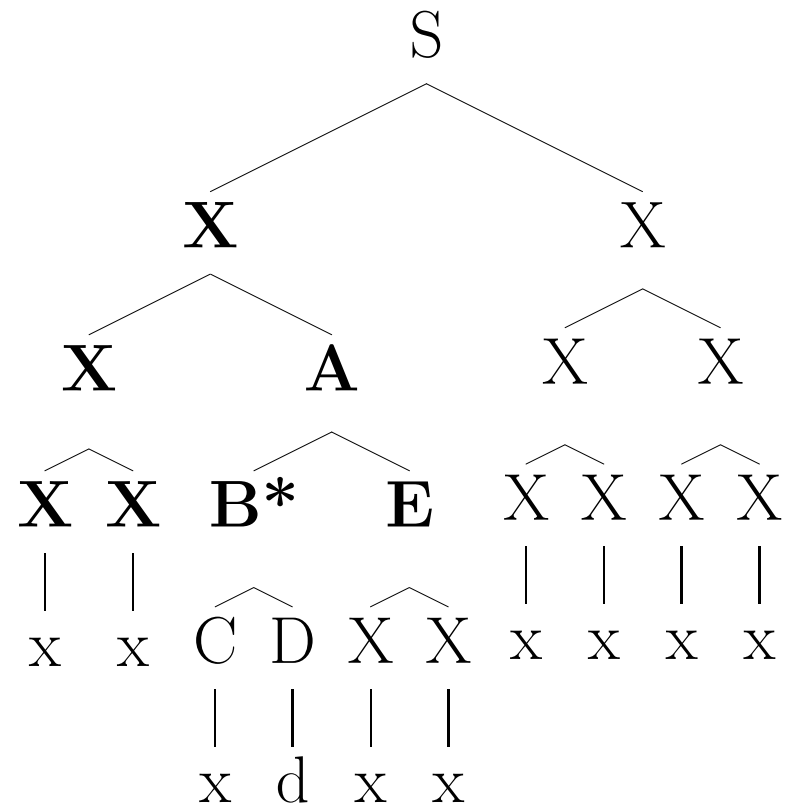
$$D_1 \circ \dots \underbrace{D_k}_{\dots B} \circ \underbrace{D_{k+1}}_{B \dots} \circ \dots \circ D_l$$

D_k can be any supertwig of d_1



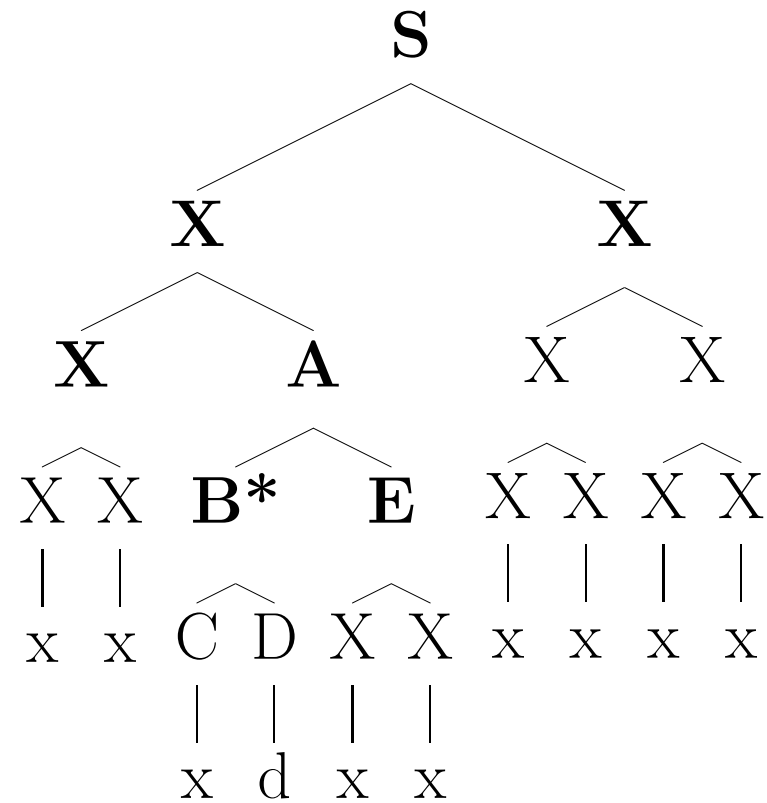
$$D_1 \circ \dots \underbrace{D_k}_{\dots B} \circ \underbrace{D_{k+1}}_{B \dots} \circ \dots \circ D_l$$

D_k can be any supertwig of d_1



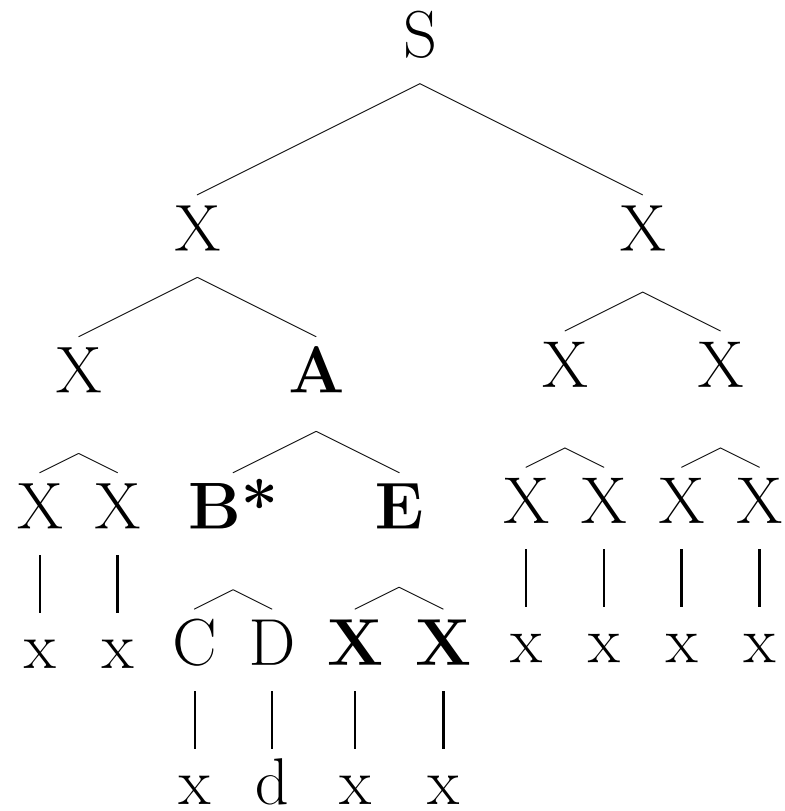
$$D_1 \circ \dots \underbrace{D_k}_{\dots B} \circ \underbrace{D_{k+1}}_{B \dots} \circ \dots \circ D_l$$

D_k can be any supertwig of d_1



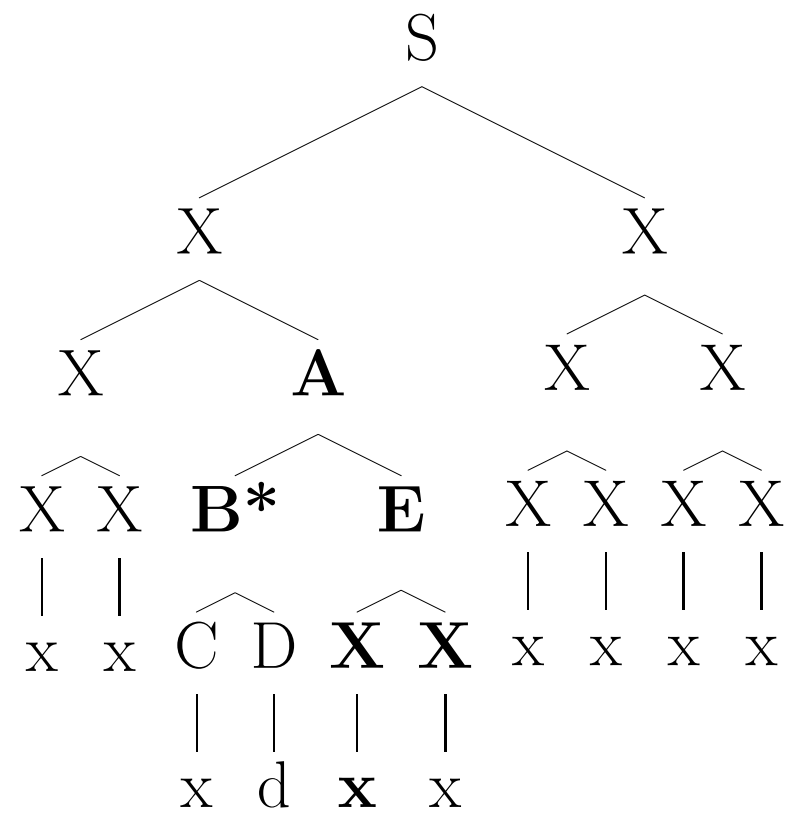
$$D_1 \circ \dots \underbrace{D_k}_{\dots B} \circ \underbrace{D_{k+1}}_{B \dots} \circ \dots \circ D_l$$

D_k can be any E-superprune of d_1



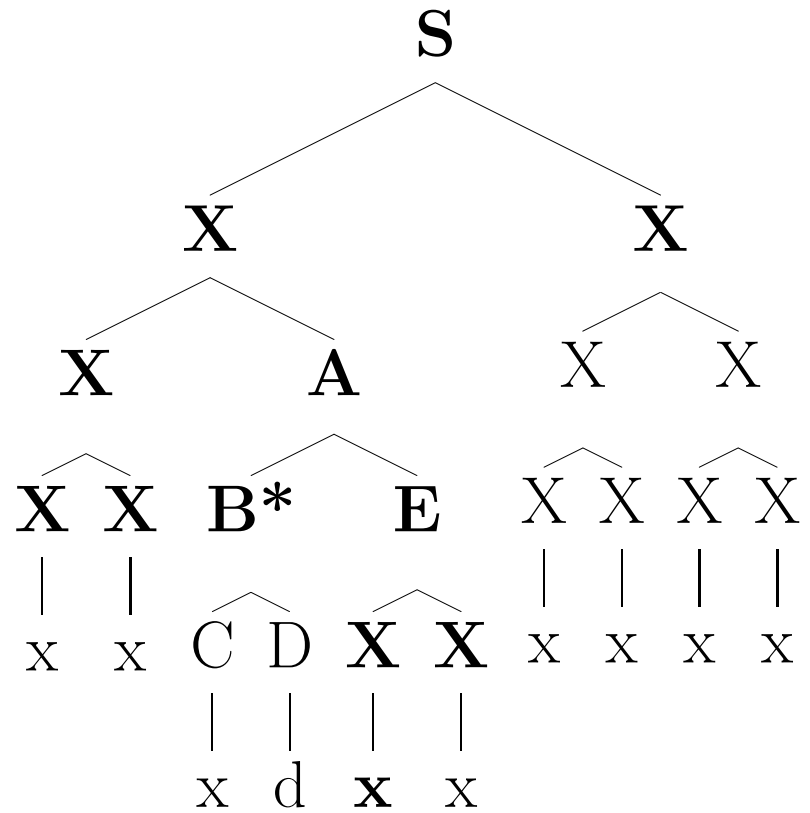
$$D_1 \circ \dots \underbrace{D_k}_{\dots B} \circ \underbrace{D_{k+1}}_{B \dots} \circ \dots \circ D_l$$

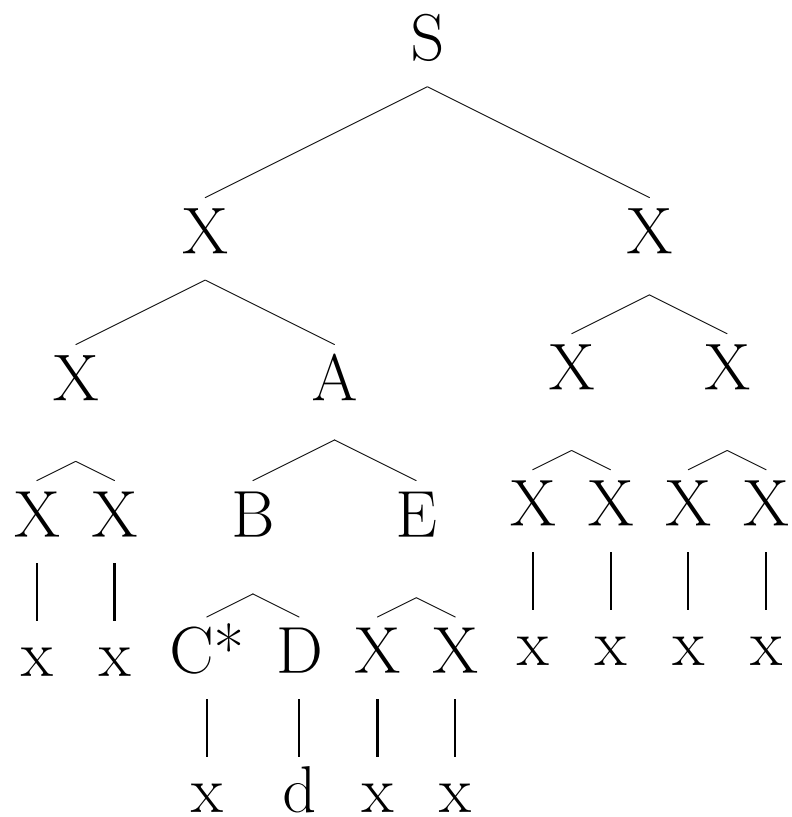
D_k can be any E-superprune of d_1

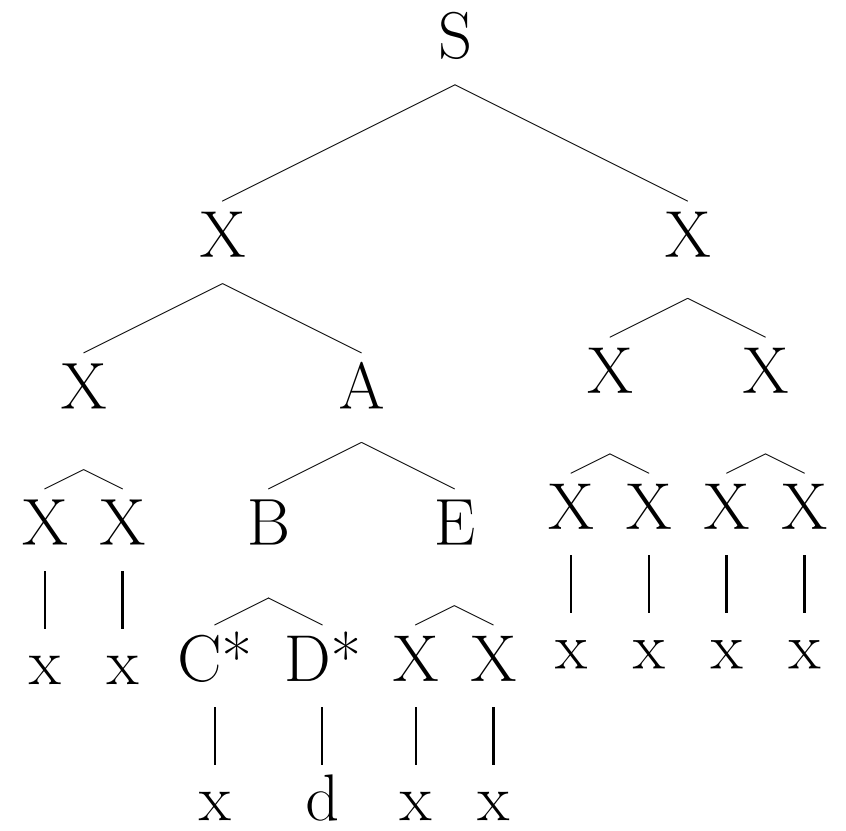


$$D_1 \circ \dots \underbrace{D_k}_{\dots B} \circ \underbrace{D_{k+1}}_{B \dots} \circ \dots \circ D_l$$

D_k can be any E-superprune of any supertwig of d_1







fully lexicalized subtrees

$$\mathbf{E}[f(t)] = \sum_{d \in D(t)} \left[\underbrace{\sum_{\tau \in \widehat{tw}(d_1)} u(\tau)}_{\alpha} \prod_{\substack{\tau' \in \\ \langle d_2, \dots, d_n \rangle}} (w(\tau')) \right]_{\beta}$$

arbitrary subtrees

$$\mathbf{E}[f(t)] = \sum_{d \in D(t)} (\alpha(d)\beta(d))$$
$$\alpha(d) = \sum_{\tau \in \widehat{tw}(d_1)} \left(\sum_{\tau' \in \widehat{pr}_{x(t)}(\tau)} u(\tau') \right)$$
$$\beta(d) = \prod_{\substack{t' \in \\ \langle d_2, \dots, d_n \rangle}} \left(\sum_{\tau' \in \widehat{pr}_{x(t)}(t')} w(\tau') \right)$$

arbitrary subtrees

$$\begin{aligned}
 \mathbf{E}[f(t)] &= \sum_{d \in D(t)} (\alpha(d)\beta(d)) \\
 \alpha(d) &= \sum_{\tau \in \widehat{tw}(d_1)} \left(\sum_{\tau' \in \widehat{pr}_{x(t)}(\tau)} u(\tau') \right) = \alpha_{d_1, x(t)} \\
 \beta(d) &= \prod_{\substack{t' \in \\ \langle d_2, \dots, d_n \rangle}} \left(\sum_{\tau' \in \widehat{pr}_{x(t)}(t')} w(\tau') \right) \\
 &= \prod_{\substack{t' \in \\ \langle d_2, \dots, d_n \rangle}} \frac{1}{\sum_{t'': r(t')=r(t'')} u(t'')} \underbrace{\left(\sum_{\tau' \in \widehat{pr}_{x(t)}(t')} u(\tau') \right)}_{\beta_{t', x(t)}^*}
 \end{aligned}$$

Keeping track of the “supertwig” and “superprune” usage probabilities with every change of these probabilities.

change-usage-probabilities(τ, δ)

$$u(\tau) + = \delta$$

for each prune τ' at sites x

$$\beta_{\tau',x}^* + = \delta$$

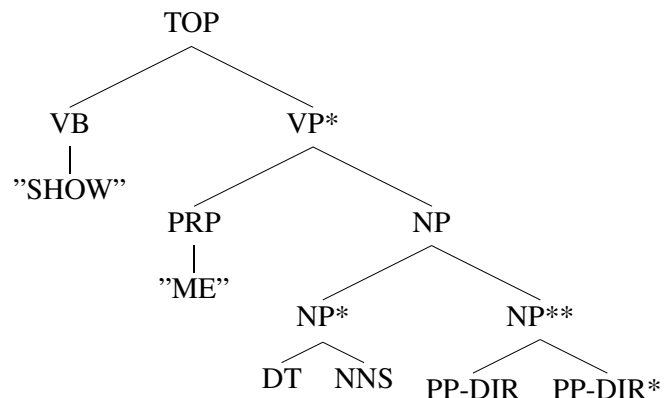
for each twig τ'' of τ

$$\alpha_{\tau'',x} + = \delta$$

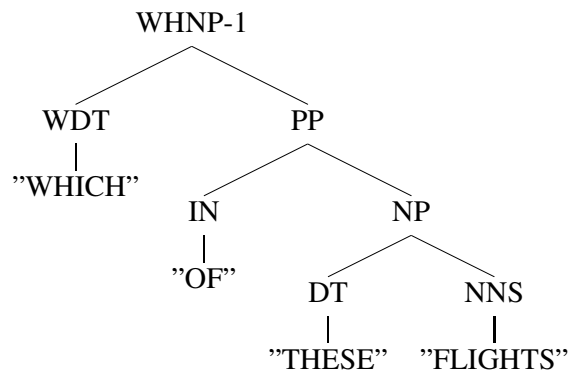
for each observed parse tree p
 for each depth-1 subtree t in p
 update-score($t, 1.0$)
 for each subtree t of p
 $\Delta = \min(sc(t), B + \gamma(\mathbf{E}[f(t)] - f(t)))$
 $\Delta' = 0$
 for each of n derivations d of t
 let $t' \dots t''$ be all elementary trees in d
 $\delta = \min(sc(t'), \dots, sc(t''), -\Delta/n)$
 $\Delta' = \delta$
 for each elementary tree t' in d
 update-score(t', δ)
 update-score (t, Δ')

Promising results on ATIS:

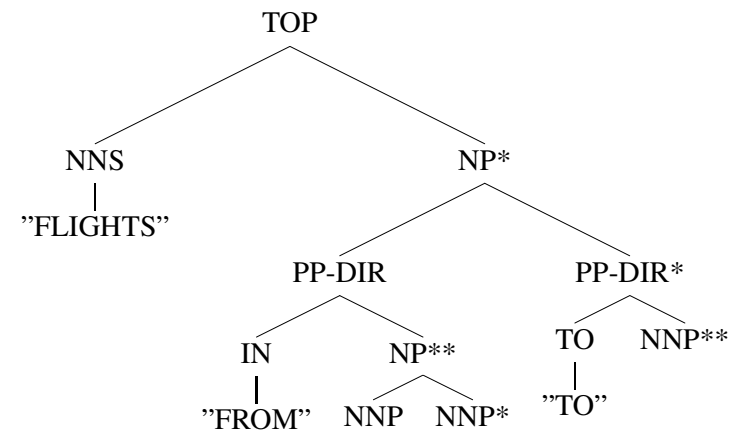
method	# rules	Cov.	LR	LP	EM
DOP1	77852	84%	95.07	95.07	83.5
p-n-p	58799	84%	95.07	95.07	83.5



(a) The “show me NP PP” frame, which occurs very frequently in the training data and is represented in several elementary trees with high weight.



(b) The complete parse tree for the sentence “Which of these flights”, which occurs 16 times in training data.



(c) The frame for “flights from NP to NP”

Unsupervised DOP

(Bod, 2006)

- All binary trees that can be assigned to sentences
- DOP as in (Bod, 2003)
- State of the art results on Unlabeled Precision and Recall

Conclusions

- In the family of CG formalisms, DOP is an early and highly successful approach (in statistical parsing)
- Two crucial steps towards its use as a psycholinguistic model have been taken: (i) a new estimator, (ii) an unsupervised version

Evolution?

Fluid Construction Grammar

Analytic Route

ILM