Discontinuous Data-Oriented Parsing:

A mildly context-sensitive all-fragments grammar

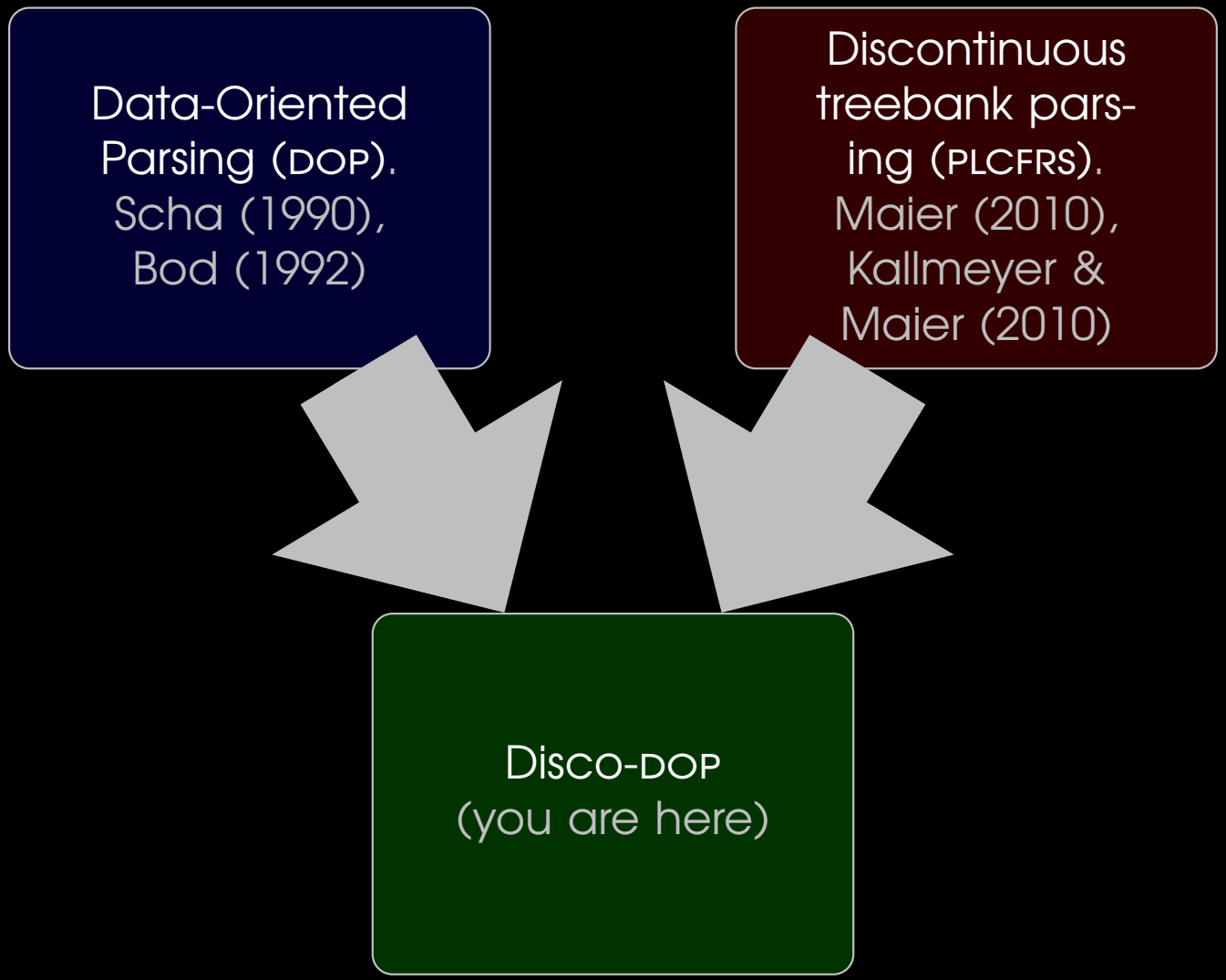
Andreas van Cranenburgh,
Remko Scha, Federico Sangati

University of Amsterdam
Institute for Logic, Language and Computation

October 6, 2011

SPMRL 2011, Dublin

Overview


Data-Oriented Parsing (DOP). Scha (1990), Bod (1992)
Discontinuous treebank parsing (PLCFRS). Maier (2010), Kallmeyer & Maier (2010)
DISCO-DOP (you are here)

Example

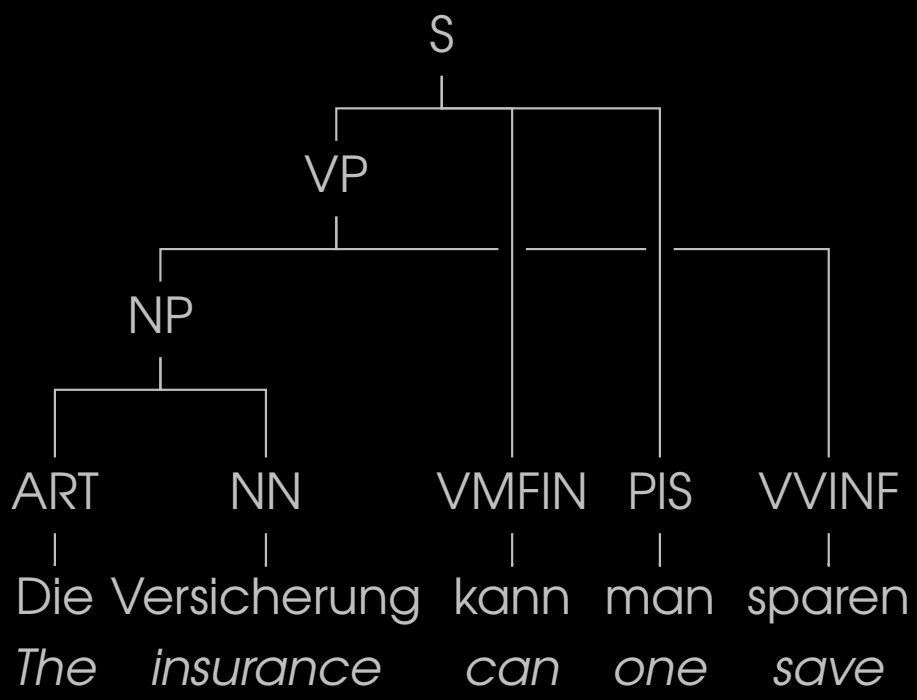


Figure: A discontinuous tree from the Negra corpus.
Translation: *As for the insurance, one can save it.*

Discontinuity

- Cross-serial dependencies
- Extraposition: topicalization, wh-extraction
- Word-order freedom: scrambling

German Negra & Tiger treebanks have discontinuous annotations.
About 30% of sentences contain discontinuity.

Context-Free grammar

CFG: rewrites strings

Productions as deduction schemata:

$$S(ab) \rightarrow NP(a)\ VP(b)$$

where ab is the concatenation of a and b

$\Rightarrow$ cubic time complexity of CKY:

$$\mathcal{O}(|w|^{3 \cdot 1})$$

- where w is the input string,
- 3 is number of non-terminals in a production,
- 1 is number of arguments of each non-terminal

Linear Context-Free Rewriting Systems

LCFRS are a generalization of CFG:

$\Rightarrow$ rewrite tuples, trees or graphs!

LCFRS allow any number of arguments (fan-out):

$$S(abc) \rightarrow NP(b)\ VP(a, c)$$

where abc is the concatenation of a, b, and c

linear: each variable on the left occurs once on the right & vice versa

Vijay-Shanker, Weir, Joshi (1987): Structural descriptions (. . .)

Linear Context-Free Rewriting Systems

A binarized LCFRS has complexity

$$\mathcal{O}(|w|^{3\varphi})$$

where φ is the maximum fan-out in a production of the grammar.

Both CFG & LCFRS $\in$ LOGCFL

Rules can be read off from treebank,
frequencies form MLE $\Rightarrow$ PLCFRS

Treebank grammars

Treebank grammar

trees $\Rightarrow$ productions (+frequencies)

- Why? Straightforward, efficient
- Why not? Arbitrary, coarse grained

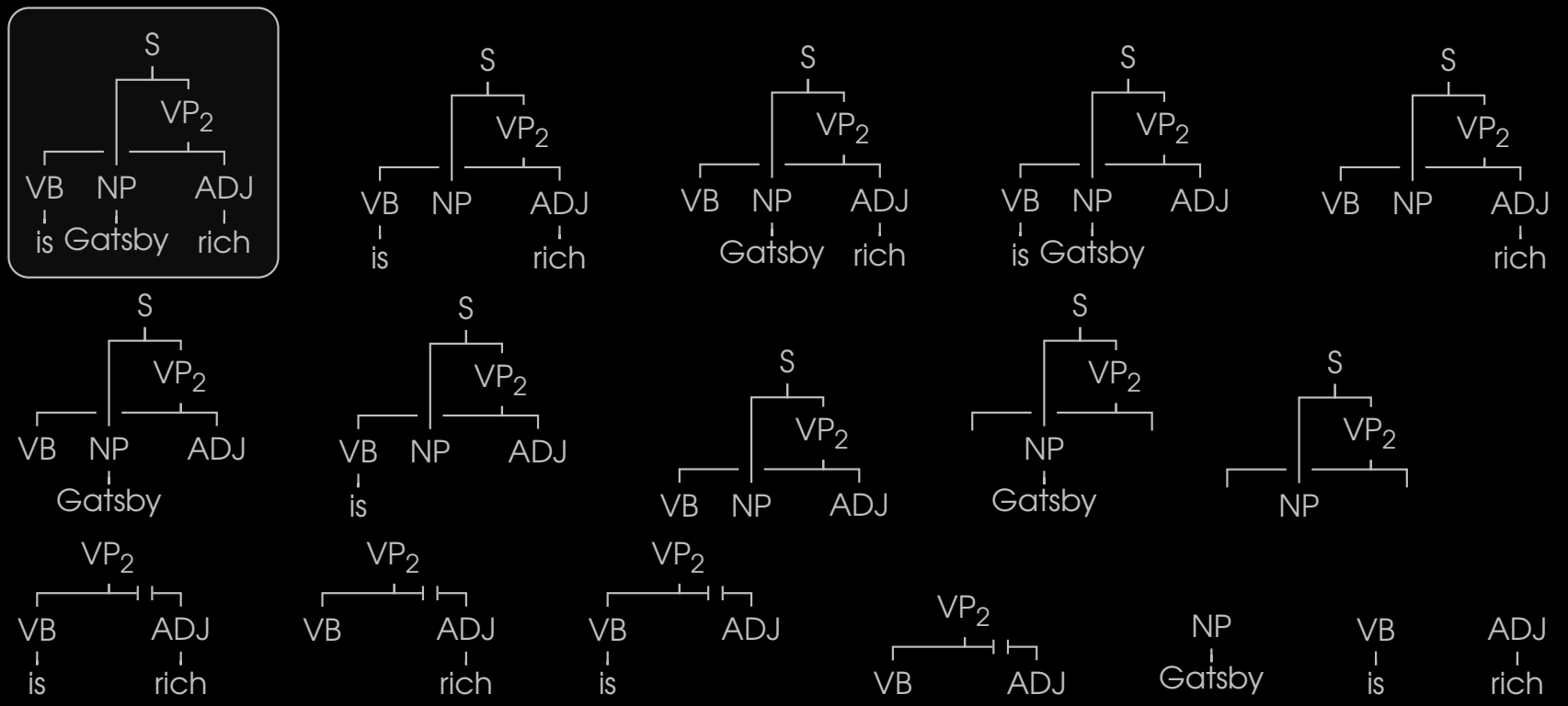
  $\Rightarrow$ strong independence assumptions

Data-Oriented Parsing

Alternative: Data-Oriented Parsing
trees $\Rightarrow$ fragments (+frequencies)

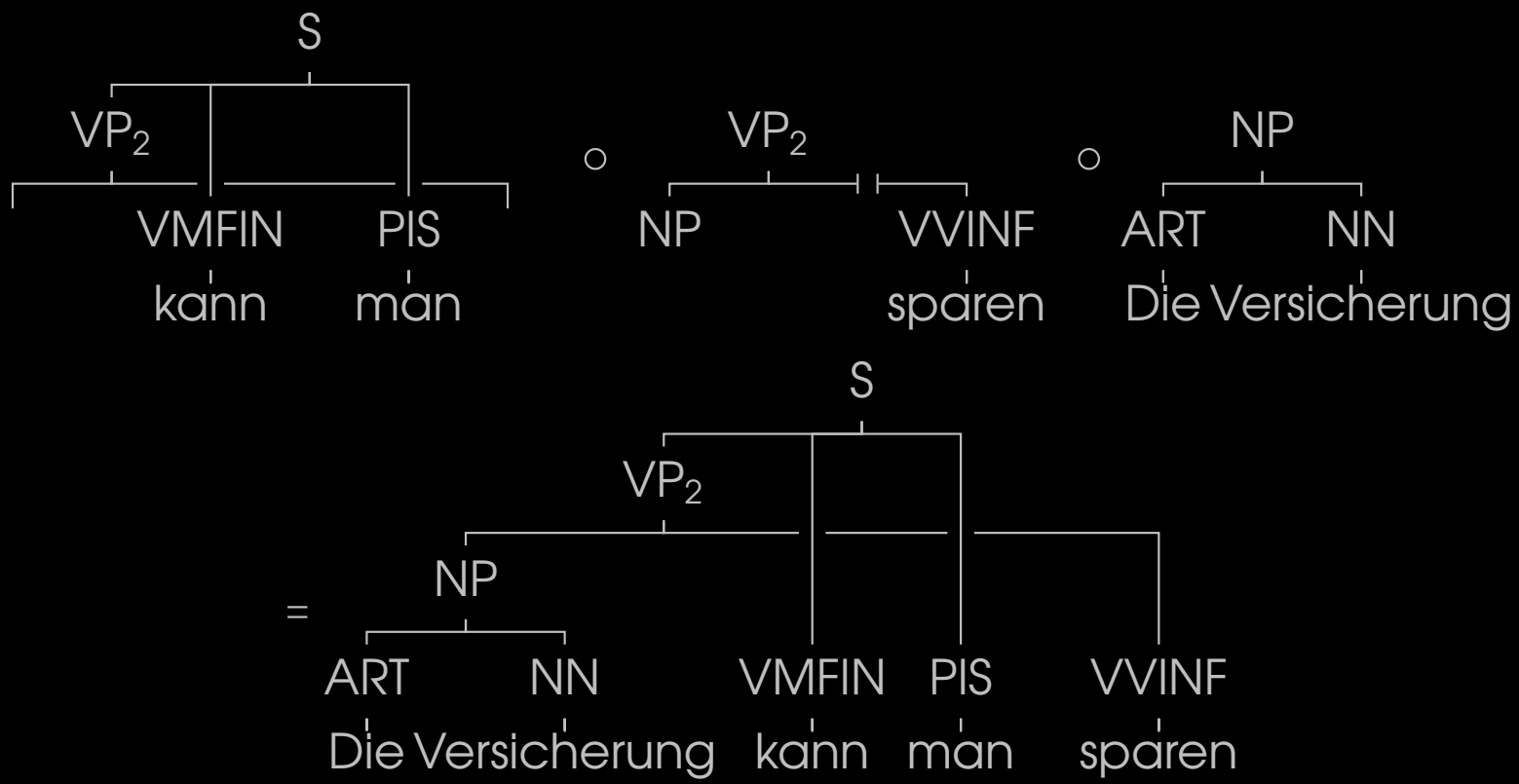
A fragment is a connected subset of a tree in which each node either has all children in common with the original tree, or none (substitution site)

DOP fragments



$$P(f) = \frac{\text{count}(f)}{\sum_{f' \in F} \text{count}(f')} \text{ where } F = \{ f' \mid root(f') = root(f) \}$$

DOP derivation

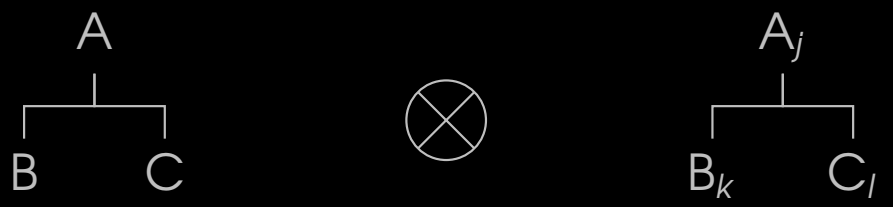


$$P(d) = P(f_1 \circ \cdots \circ f_n) = \prod_{f \in d} p(f)$$

$$P(t) = P(d_1) + \cdots + P(d_n) = \sum_{d \in D(t)} \prod_{f \in d} p(f)$$

DOP Reduction

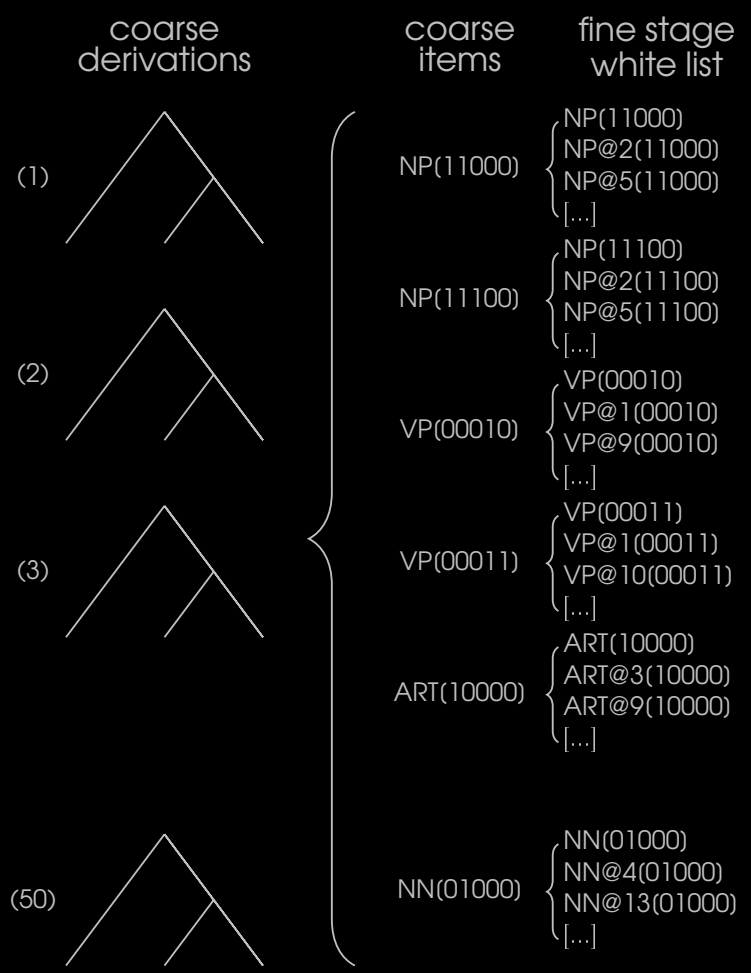
- Treebank refinement: take non-terminal and split according to contexts
- In the limit: each non-terminal becomes a particular occurrence in a tree



$A_j(\vec{\alpha}) \rightarrow B(\vec{\alpha_B})\ C(\vec{\alpha_C})$	$(1/a_j)$	$A(\vec{\alpha}) \rightarrow B(\vec{\alpha_B})\ C(\vec{\alpha_C})$	$(1/(a\bar{a}))$
$A_j(\vec{\alpha}) \rightarrow B_k(\vec{\alpha_B})\ C(\vec{\alpha_C})$	(b_k/a_j)	$A(\vec{\alpha}) \rightarrow B_k(\vec{\alpha_B})\ C(\vec{\alpha_C})$	$(b_k/(a\bar{a}))$
$A_j(\vec{\alpha}) \rightarrow B(\vec{\alpha_B})\ C_l(\vec{\alpha_C})$	(c_l/a_j)	$A(\vec{\alpha}) \rightarrow B(\vec{\alpha_B})\ C_l(\vec{\alpha_C})$	$(c_l/(a\bar{a}))$
$A_j(\vec{\alpha}) \rightarrow B_k(\vec{\alpha_B})\ C_l(\vec{\alpha_C})$	$(b_k c_l/a_j)$	$A(\vec{\alpha}) \rightarrow B_k(\vec{\alpha_B})\ C_l(\vec{\alpha_C})$	$(b_k c_l/(a\bar{a}))$

Goodman (2003): Efficient parsing of DOP with PCFG-reductions

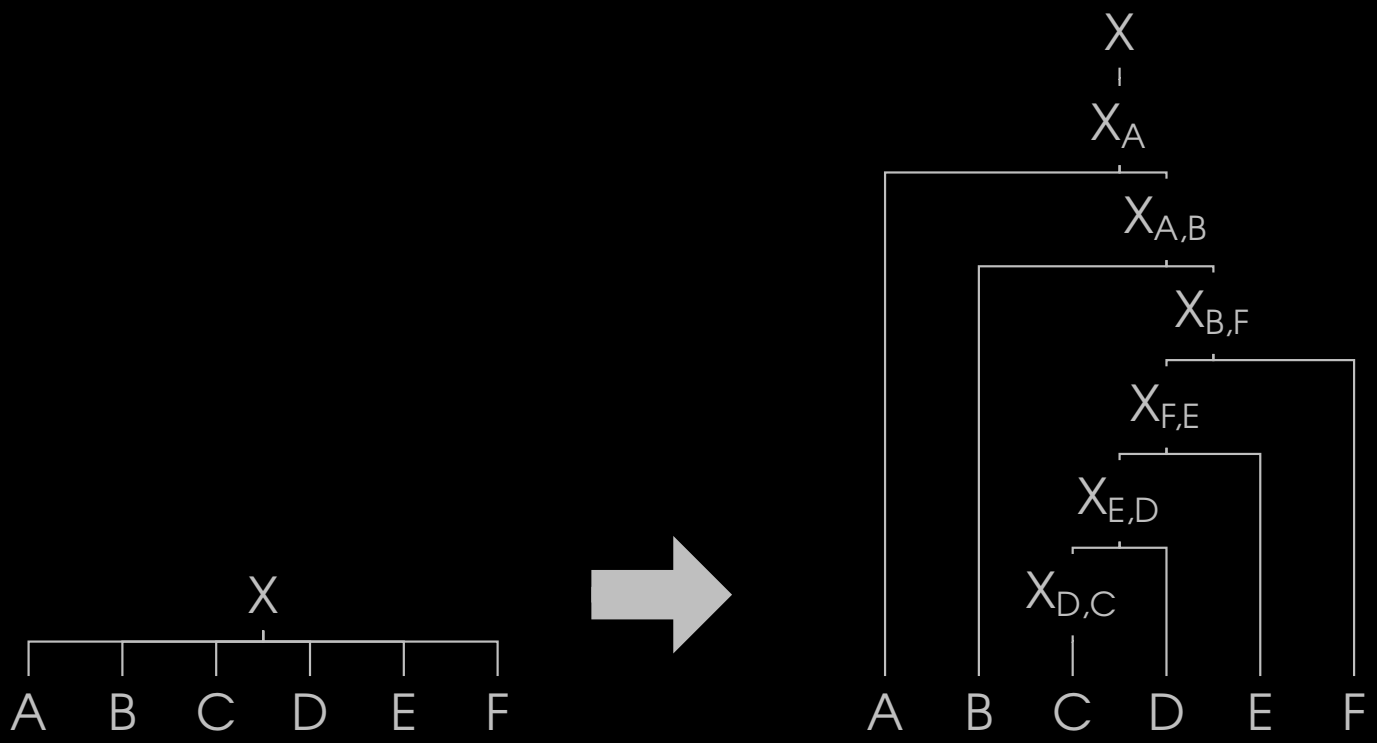
Coarse-to-fine



k-best PLCFRS derivations
help prune DOP derivations.

Binarization

- mark heads
- head-outward binarization
- no parent annotation: $v = 1$
- horizontal Markovization: $h \in \{1, 2, \infty\}$



Evaluation

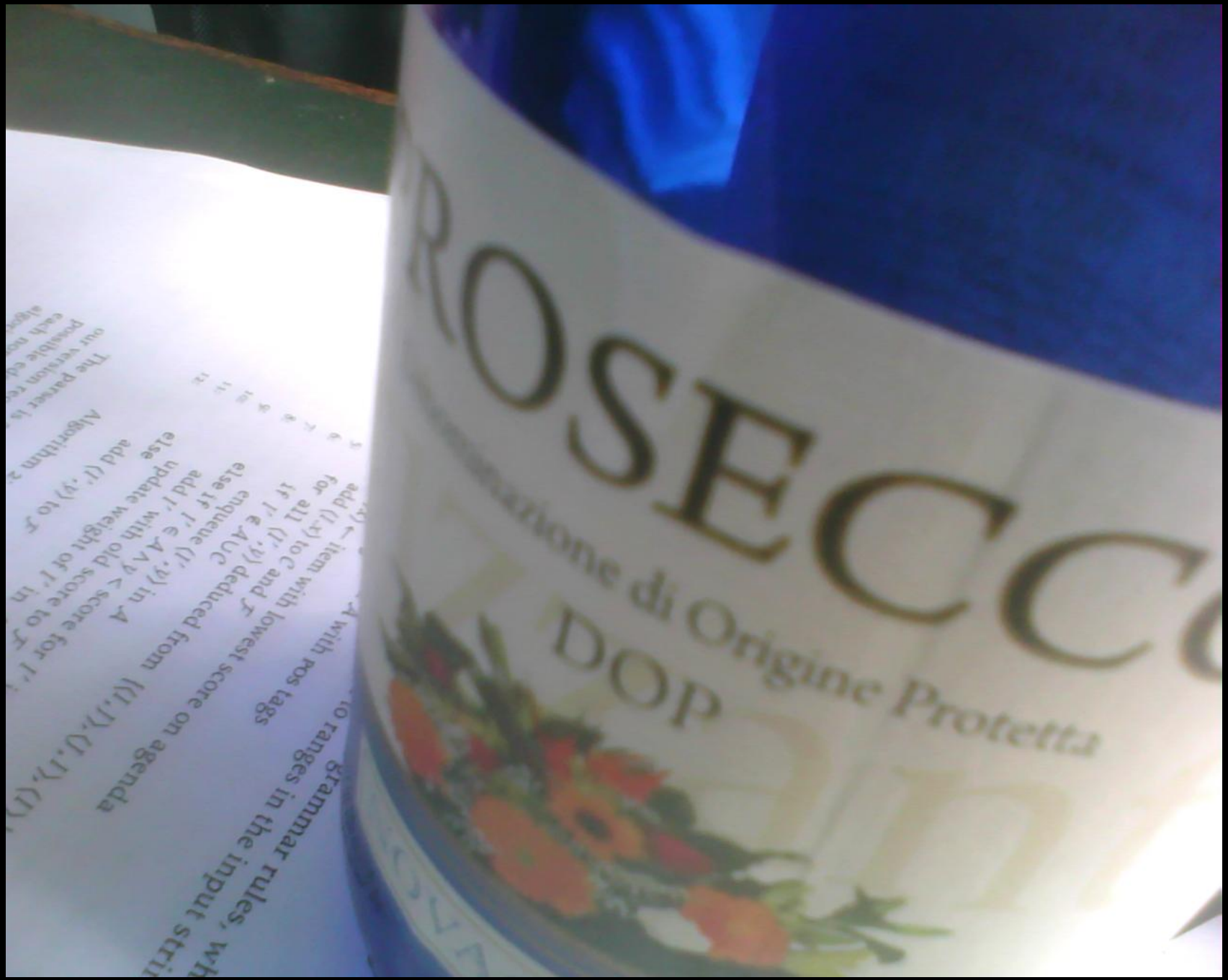
NEGRA	words	F_1	EX	COV.
DPSG Pla2004*	$\leq$ 15	73.16	39.0	96.04
PLCFRS KaMa2010†	$\leq$ 15	81.27	-	-
DISCO-DOP v=1, h=1	$\leq$ 15	**84.56**	54.68	99.90
PLCFRS KaMa2010†	$\leq$ 25	73.25	-	99.45
PLCFRS v=1, h=2	$\leq$ 25	75.98	36.79	98.90
DISCO-DOP v=1, h=2	$\leq$ 25	**78.81**	39.60	98.90
PLCFRS Mai2010‡	$\leq$ 30	71.52	-	97.00
PLCFRS v=1, h=∞	$\leq$ 30	72.34	31.27	96.59
DISCO-DOP v=1, h=∞	$\leq$ 30	**73.98**	34.96	96.59

Table: Discontinuous parsing on the Negra corpus.
Function tags discarded; Gold POS tags given to parser.

All code available from:

`http://github.com/andreasvc/disco-dop`

*Plaehn (2004). †Kallmeyer & Maier (2010). ‡Maier (2010).


ROSECC
Origine Protetta
DOP