Lecture 8: The Berkeley Parser

Jelle Zuidema

Institute for Logic, Language and Computation
University of Amsterdam, The Netherlands

Unsupervised Language Learning 2014
University of Amsterdam

Three ways forward

Maximum Likelihood PCFG (with traditional linguistic categories) does not work well... So:

- Change the generative model
  - CCM, DMV, UDOP
- Change the supervision mode
  - Learn from bracketed sentences - discover constituent labels (categorization) (Lari & Young, 1990; Borensztajn & Zuidema, 2007);
  - Learn from annotated treebank trees - discover *finer* constituent labels (Klein & Manning, 2003; Prescher, 2005; Matzuzaki et al., 2005; Petrov et al., 2006)

Three ways forward (ctd)

- Change the objective function

$$\arg\max_g P(g|d) = \arg\max_g P(g)P(d|g)$$

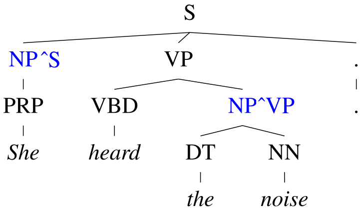
BMM, Bayesian Tree Substitution Grammar (O'Donnell et al. 2009, Cohn & Blunsom 2009, Knight 2009)

For next class, read Knight (2009; p.1-13) instead of Zuidema (2007)!

Learning Accurate, Compact, and Interpretable Tree Annotation

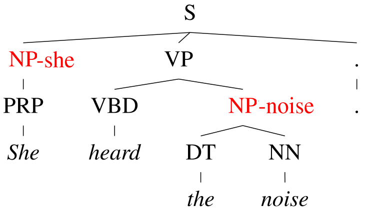
Slav Petrov, Leon Barrett, Romain Thibaux,
Dan Klein

The Game of Designing a Grammar



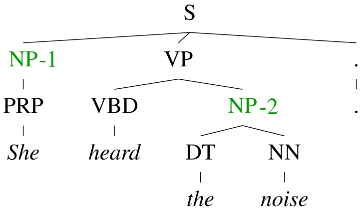
- Annotation refines base treebank symbols to improve statistical fit of the grammar
  - Parent annotation [Johnson '98]

The Game of Designing a Grammar



- Annotation refines base treebank symbols to improve statistical fit of the grammar
  - Parent annotation [Johnson ’98]
  - Head lexicalization [Collins ’99, Charniak ’00]

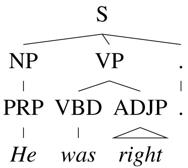
The Game of Designing a Grammar



- Annotation refines base treebank symbols to improve statistical fit of the grammar
  - Parent annotation [Johnson ’98]
  - Head lexicalization [Collins ’99, Charniak ’00]
  - Automatic clustering?

Previous Work: Manual Annotation [Klein & Manning '03]

- Manually split categories
  - NP: subject vs object
  - DT: determiners vs demonstratives
  - IN: sentential vs prepositional



- Advantages:
  - Fairly compact grammar
  - Linguistic motivations
- Disadvantages:
  - Performance leveled out
  - Manually annotated

Model	*F1*
Naïve Treebank Grammar	72.6
Klein & Manning '03	86.3

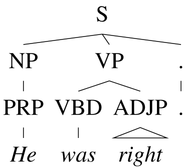
[Matsuzaki et. al ’05, Prescher ’05]

Previous Work: Automatic Annotation Induction

- Advantages:
  - Automatically learned:

    Label *all* nodes with latent variables.

    Same number k of subcategories for all categories.



- Disadvantages:
  - Grammar gets too large
  - Most categories are oversplit while others are undersplit.

Model	*F1*
Klein & Manning ’03	86.3
Matsuzaki et al. ’05	86.7

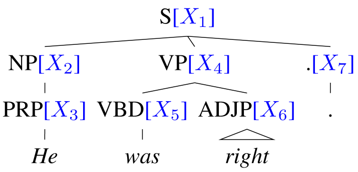
Previous work is complementary

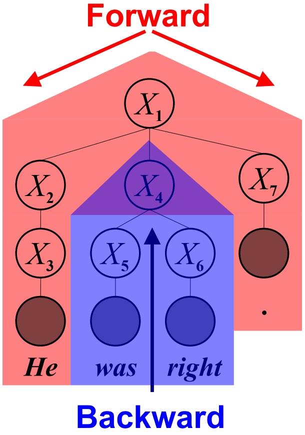
This Work
Allocates splits where needed
Automatically learned
Compact Grammar
Captures many features

Learning Latent Annotations

EM algorithm:

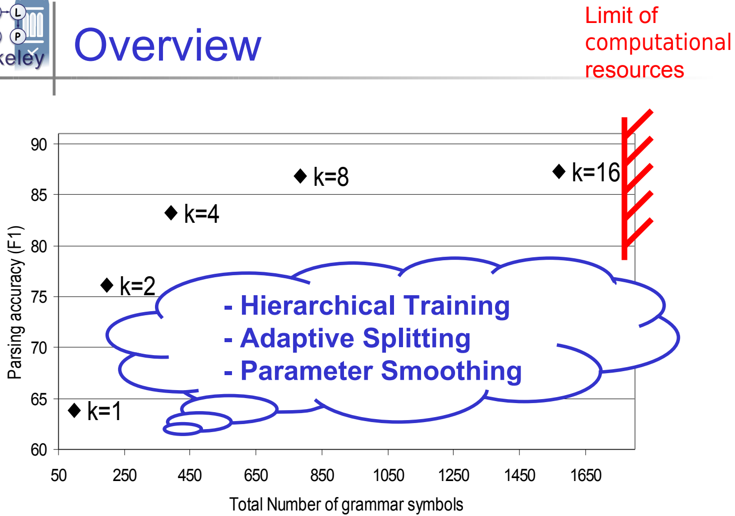
- Brackets are known
- Base categories are known
- Only induce subcategories



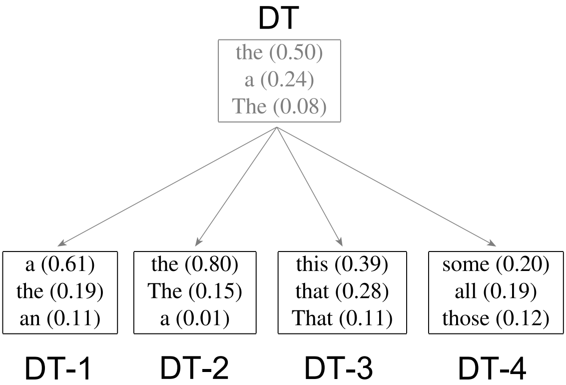


Just like Forward-Backward for HMMs.

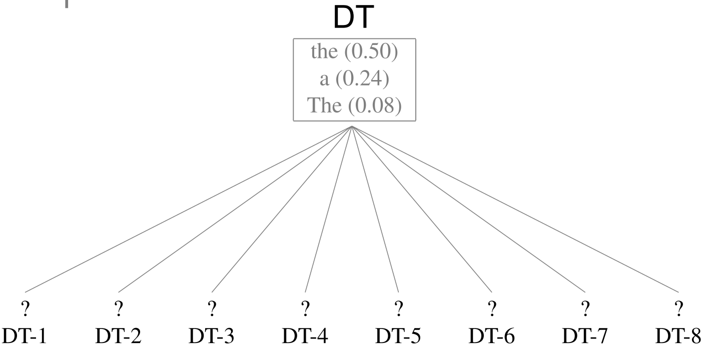
Overview


Limit of computational resources
Parsing accuracy (F1)
90
85
80
75
70
65
60
k=1
k=2
k=4
k=8
k=16
- Hierarchical Training
- Adaptive Splitting
- Parameter Smoothing
50
250
450
650
850
1050
1250
1450
1650
Total Number of grammar symbols

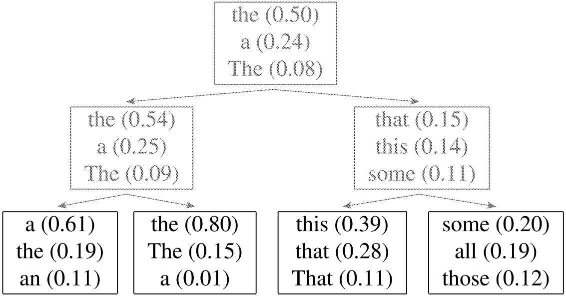
Refinement of the DT tag



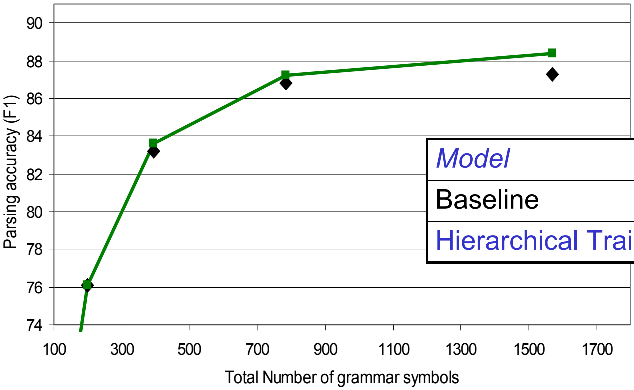
Refinement of the DT tag



Hierarchical refinement of the DT tag



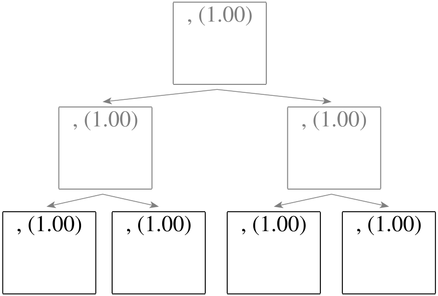
Hierarchical Estimation Results



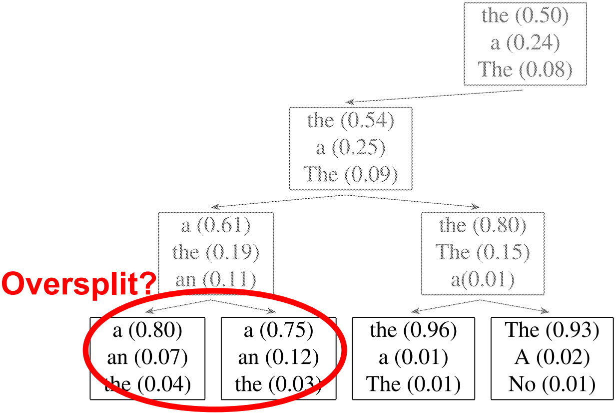
Model	*F1*
Baseline	87.3
Hierarchical Training	88.4

Refinement of the , tag

- Splitting all categories the same amount is wasteful:

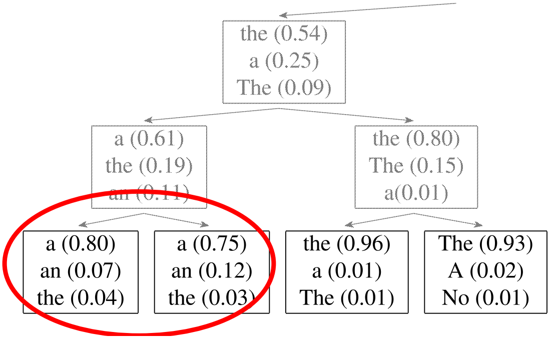


The DT tag revisited



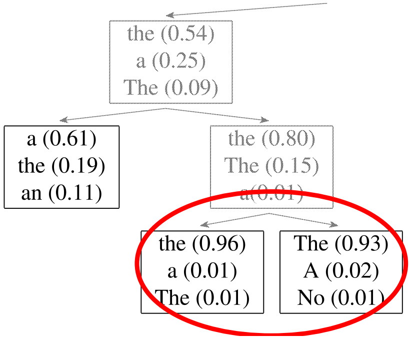
Adaptive Splitting

- Want to split complex categories more
- Idea: split everything, roll back splits which were least useful



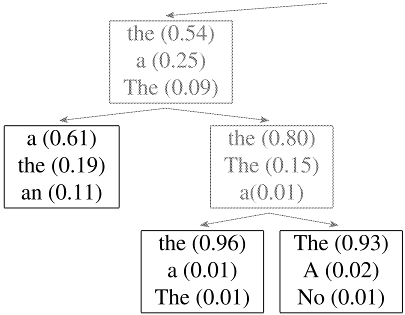
Adaptive Splitting

- Want to split complex categories more
- Idea: split everything, roll back splits which were least useful



Adaptive Splitting

- Want to split complex categories more
- Idea: split everything, roll back splits which were least useful

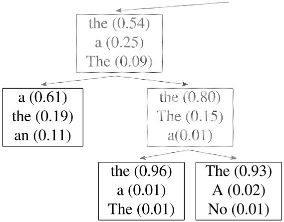


Adaptive Splitting

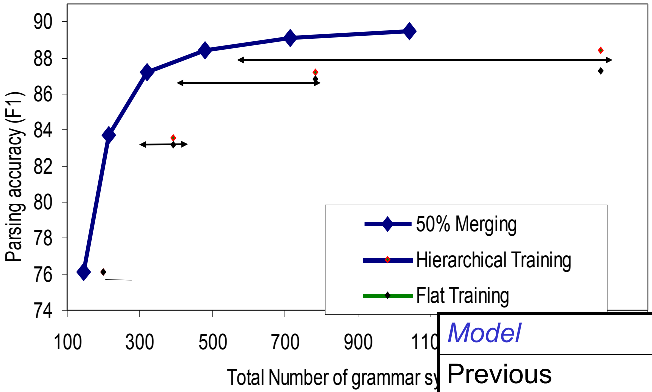
- Evaluate loss in likelihood from removing each split =

$$\frac{\textit{Data likelihood with split reversed}}{\textit{Data likelihood with split}}$$

- No loss in accuracy when 50% of the splits are reversed.

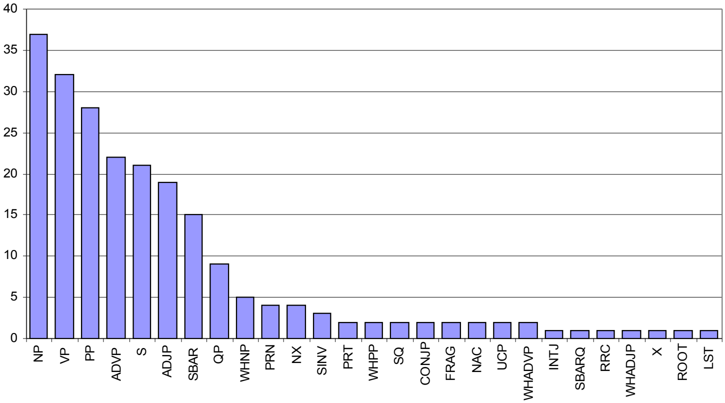


Adaptive Splitting Results

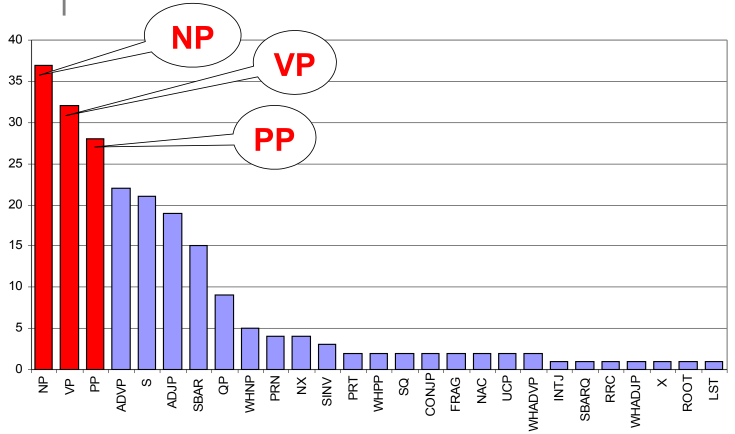


Model	*F1*
Previous	88.4
With 50% Merging	89.5

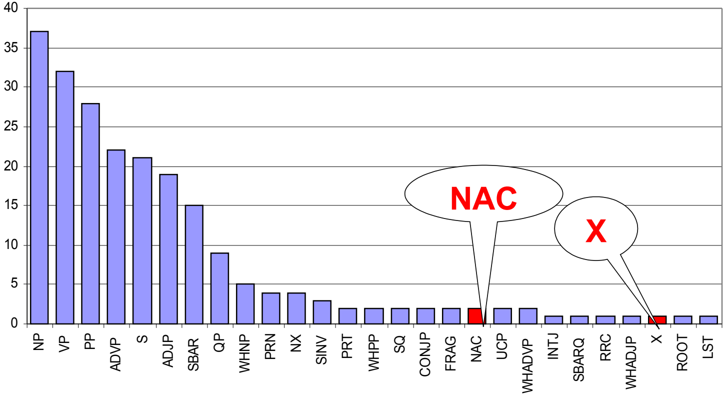
Number of Phrasal Subcategories



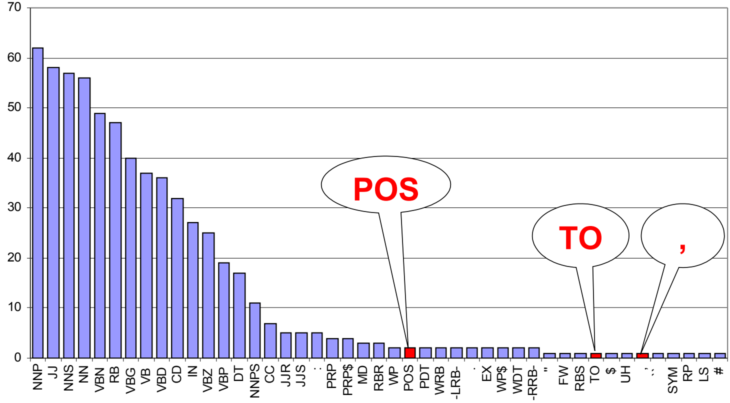
Number of Phrasal Subcategories



Number of Phrasal Subcategories

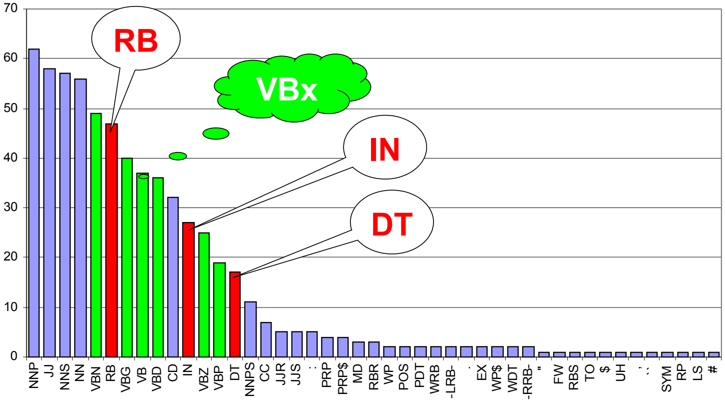


Number of Lexical Subcategories

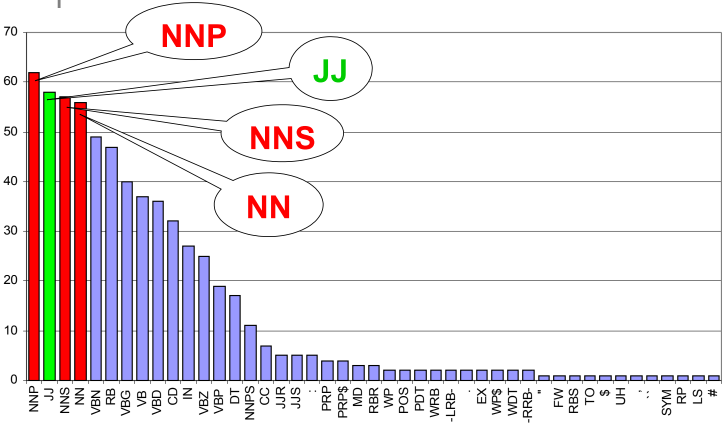


University of California CAL NLP Berkeley

Number of Lexical Subcategories

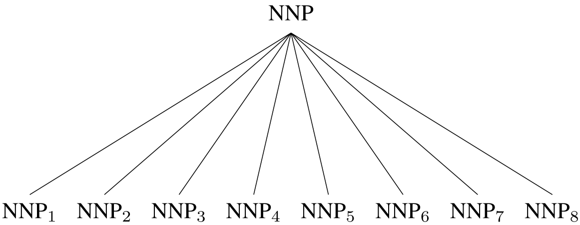


Number of Lexical Subcategories



Smoothing

- Heavy splitting can lead to overfitting
- Idea: Smoothing allows us to pool statistics



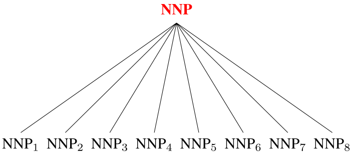
Linear Smoothing

$$p_x = \mathrm{P}(A_x \rightarrow BC)$$

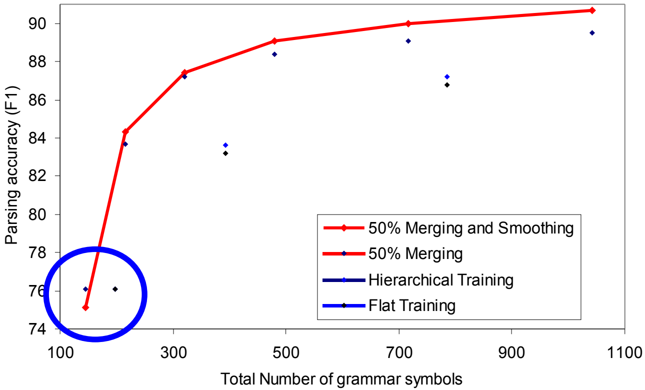
$$\Downarrow$$

$$p'_x = (1-\alpha)p_x + \alpha\bar{p}$$

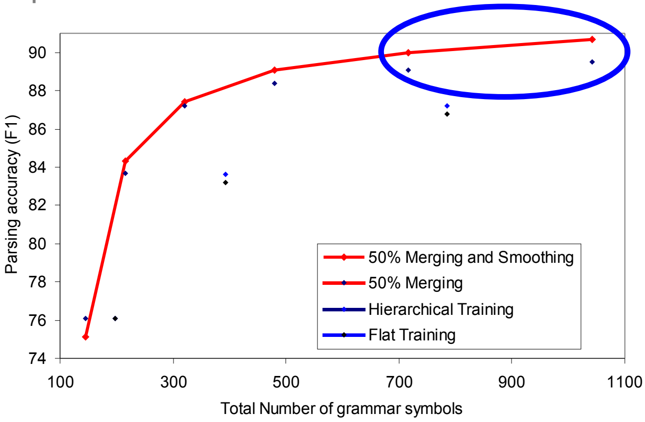
where $\bar{p} = \frac{1}{n}\sum_x p_x$



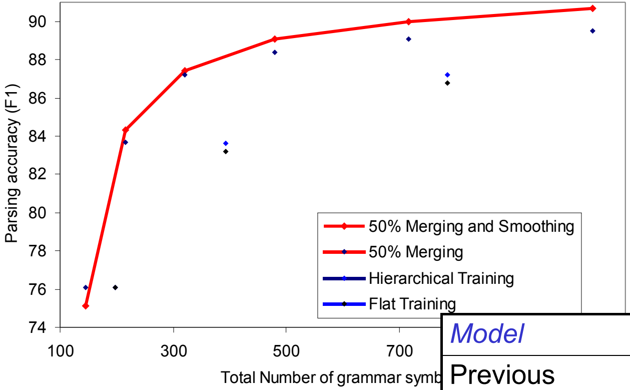
Result Overview



Result Overview



Result Overview



Model	F1
Previous	89.5
With Smoothing	90.7

Final Results

Parser	*F1 ≤ 40 words*	*F1 all words*
Klein & Manning ’03	86.3	85.7
Matsuzaki et al. ’05	86.7	86.1
This Work	**90.2**	**89.7**

Final Results

Parser	*F1* *≤ 40 words*	*F1* *all words*
Klein & Manning '03	86.3	85.7
Matsuzaki et al. '05	86.7	86.1
Collins '99	88.6	88.2
Charniak & Johnson '05	90.1	89.6
This Work	**90.2**	**89.7**

Linguistic Candy

- Proper Nouns (NNP):

NNP-14	Oct.	Nov.	Sept.
NNP-12	John	Robert	James
NNP-2	J.	E.	L.
NNP-1	Bush	Noriega	Peters
NNP-15	New	San	Wall
NNP-3	York	Francisco	Street

- Personal pronouns (PRP):

PRP-0	It	He	I
PRP-1	it	he	they
PRP-2	it	them	him

Linguistic Candy

- Relative adverbs (RBR):

RBR-0	further	lower	higher
RBR-1	more	less	More
RBR-2	earlier	Earlier	later

- Cardinal Numbers (CD):

CD-7	one	two	Three
CD-4	1989	1990	1988
CD-11	million	billion	trillion
CD-0	1	50	100
CD-3	1	30	31
CD-9	78	58	34

Conclusions

- New Ideas:
  - Hierarchical Training
  - Adaptive Splitting
  - Parameter Smoothing
- State of the Art Parsing Performance:
  - Improves from X-Bar initializer 63.4 to 90.2
- Linguistically interesting grammars to sift through.

Multi-Word Expressions

Idiomatic expressions

-
    - by and large
    - lo and behold
    - beat a dead horse
    - make amends
    - cast aspersions
    - a flash in the pan

Multi-Word Expressions

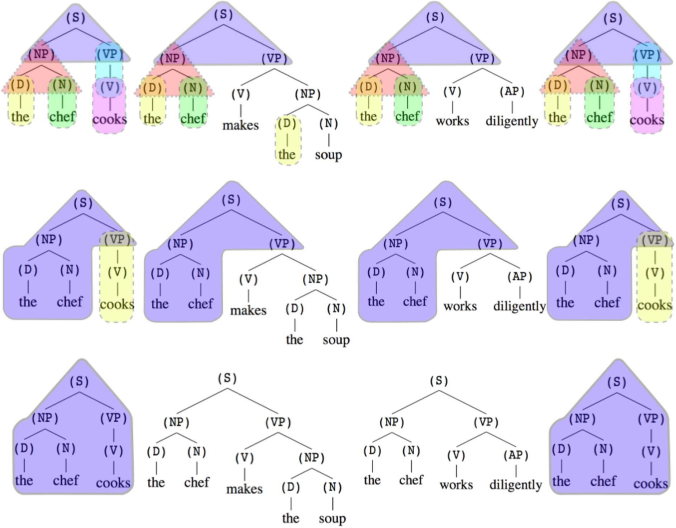
Formulaic expressions

-
    - declined to comment (WSJ)
    - Stocks went up to $ 15.4 from $ 15.3 (WSJ)
    - I want to travel from Baltimore to Oakland (ATIS)
    - Ik wil vandaag van Amsterdam naar Leuven (OVIS)

Minimal

Intermediate

Maximal



PTSGs: Old Generative Story

Probabilistic Tree Substitution Grammars

An PTSG 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: [hidden]

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

The probability of a parse: [observable]

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

PTSGs: New Generative Story

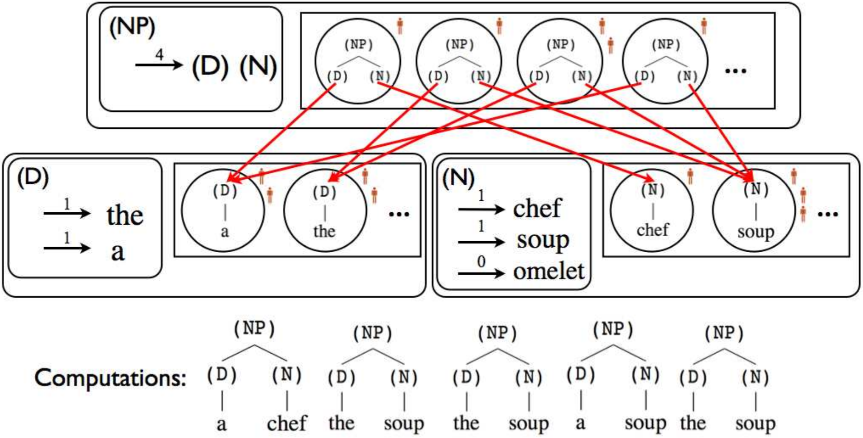
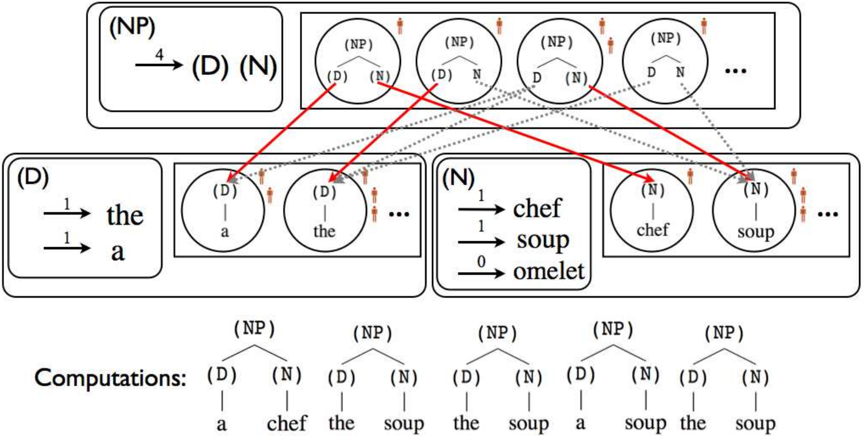


Figure 13: Representation of a possible state of an adaptor grammar after



(NP)
4
(D) (N)
(NP)
(D) (N)
(NP)
(D) N
(NP)
D (N)
(NP)
D N
...
(D)
1
the
1
a
(D)
a
(D)
the
...
(N)
1
chef
1
soup
0
omelet
(N)
chef
(N)
soup
...
Computations:
(NP)
(D) (N)
a chef
(NP)
(D) (N)
the soup
(NP)
(D) (N)
the soup
(NP)
(D) (N)
a soup
(NP)
(D) (N)
the soup