# Induction of Treebank-Aligned Lexical Resources

## *LREC 2008*

Tejaswini Deoskar, Mats Rooth

Department of Linguistics

Cornell University

# Overview

- Goal: Induction of probabilistic treebank-aligned lexical resources.
- *Treebank-Aligned Lexicon*: a systematic correspondence between features of a probabilistic lexicon and structural annotation in a treebank.
- Features:
  - complex subcategorization frames for verbs or nouns.
  - attachment preference of adverbs

# Overview

- Treebank PCFG and lexicon.
  - ◇ Unlexicalised Treebank PCFG : Clear division between grammar and lexicon.
  - ◇ Good performance (Klein and Manning, 2003)
- Large-scale lexicon: Unsupervised acquisition from unlabeled data.

# Why another Treebank PCFG?

- PCFGs built from Treebanks are reduced representations.

# Why another Treebank PCFG?

- PCFGs built from Treebanks are reduced representations.
  - ```
    Exports which played a key role in
    fueling growth over the last two years
    seem to have stalled.
    ```
    ```

# Why another Treebank PCFG?

- PCFGs built from Treebanks are reduced representations.
  - ```
    Exports which played a key role in
    fueling growth over the last two years
    seem to have stalled.
    ```
- More expressive formalisms can represent these (LFG, HPSG, TAG, CCG, Minimalist grammars)
    ```

# Why another Treebank PCFG?

- PCFGs built from Treebanks are reduced representations.
  - ◇ `Exports which played a key role in fueling growth over the last two years seem to have stalled.`
- More expressive formalisms can represent these (LFG, HPSG, TAG, CCG, Minimalist grammars)
- A sophisticated PCFG that captures the same phenomena as more expressive formalisms.

# Why another Treebank PCFG?

- PCFGs built from Treebanks are reduced representations.
  - `Exports which played a key role in fueling growth over the last two years seem to have stalled.`
- More expressive formalisms can represent these (LFG, HPSG, TAG, CCG, Minimalist grammars)
- A sophisticated PCFG that captures the same phenomena as more expressive formalisms.
  - Linguistic theory neutral.

# Why another Treebank PCFG?

- PCFGs built from Treebanks are reduced representations.
  - ```
    Exports which played a key role in
    fueling growth over the last two years
    seem to have stalled.
    ```
- More expressive formalisms can represent these (LFG, HPSG, TAG, CCG, Minimalist grammars)
- A sophisticated PCFG that captures the same phenomena as more expressive formalisms.
  - Linguistic theory neutral.
  - Focus on commonly observed phenomenon.
    ```

# Treebank Transformation Framework

- Treebank Transformation : Johnson (1999), Klein and Manning (2003), etc.
- Training of PCFG on transformed treebank.

# Treebank Transformation Framework

- Treebank Transformation : Johnson (1999), Klein and Manning (2003), etc.
- Training of PCFG on transformed treebank.

- Methodology for transformation based on addition of linguistically motivated features, and feature-constraint solving.
- Database of Penn Treebank trees annotated with linguistic features as a resource.
- Components usable for transforming existing PTB-style treebanks, and building accurate PCFGs from them.

# Feature Constraint Framework

- Bare-bones CFG extracted from Penn Treebank.
- A feature-constraint grammar is built by adding constraints on CF rules (YAP, Schmid (2000)).
- Each treebank tree converted into a trivial context-free shared forest.
- Constraints in the shared forest solved by YAP constraint solver.

# Adding Constraints

Features on auxiliary verbs:

# Adding Constraints

Features on auxiliary verbs:

$$\text{VP} \quad \rightarrow \quad \text{VB} \qquad \text{ADVP} \qquad \text{VP}$$

# Adding Constraints

Features on auxiliary verbs:

$$\mathrm{VP} \quad \rightarrow \quad \mathrm{VB} \qquad \mathrm{ADVP} \qquad \mathrm{VP}$$

$$\mathrm{VP}\ \{\ \mathrm{Vform} = \mathrm{base};\ \} \quad \rightarrow \quad \mathrm{VB}\ \{\mathrm{Val} = \mathrm{aux};\} \quad \mathrm{ADVP}\ \{\ \} \quad \mathrm{VP}\ \{\ \}$$

# Adding Constraints

Features on auxiliary verbs:

$$\text{VP} \rightarrow \text{VB} \quad \text{ADVP} \quad \text{VP}$$

$$\text{VP}\ \{\ \text{Vform} = \text{base};\ \} \rightarrow \text{VB}\ \{\text{Val} = \text{aux};\}\ \ \text{ADVP}\ \{\ \}\ \ \text{VP}\ \{\ \}$$

$$\text{VP}\ \{\text{Vform} = \text{base};\ \ \text{Slash} = \mathit{sl};\ \} \rightarrow \text{VB}\ \{\text{Val} = \text{aux};\ \ \text{Vsel} = \mathit{vf};\ \}$$

$$\text{ADVP}\ \{\ \}$$

$$\text{VP}\ \{\ \text{Slash} = \mathit{sl};\ \text{Vform} = \mathit{vf}\ \ \}$$

# Adding Constraints

Features on auxiliary verbs:

$$\mathrm{VP} \rightarrow \mathrm{VB} \quad \mathrm{ADVP} \quad \mathrm{VP}$$

$$\mathrm{VP}\ \{\ \mathrm{Vform} = \mathrm{base};\ \} \rightarrow \mathrm{VB}\ \{\mathrm{Val} = \mathrm{aux};\} \quad \mathrm{ADVP}\ \{\ \} \quad \mathrm{VP}\ \{\ \}$$

$$\mathrm{VP}\ \{\mathrm{Vform} = \mathrm{base};\ \ \mathrm{Slash} = sl;\ \} \rightarrow \begin{array}{l} \mathrm{VB}\ \{\mathrm{Val} = \mathrm{aux};\ \ \mathrm{Vsel} = vf;\ \} \\ \mathrm{ADVP}\ \{\ \} \\ \mathrm{VP}\ \{\ \mathrm{Slash} = sl;\ \mathrm{Vform} = vf\ \ \} \end{array}$$

$$\mathrm{VP}\ \{\mathrm{Vform} = \mathrm{base};\ \ \mathrm{Slash} = sl;\ \} \rightarrow \begin{array}{l} \mathrm{VB}\ \{\mathrm{Val} = \mathrm{aux};\ \ \mathrm{Vsel} = vf; \\ \mathrm{Prep} = -\ ;\ \mathrm{Prtcl} = -;\ \mathrm{Sbj} = -;\ \} \\ \mathrm{ADVP}\ \{\ \} \\ \mathrm{VP}\ \{\mathrm{Slash} = sl;\ \mathrm{Vform} = vf\} \end{array}$$

# Adding Constraints

Features on auxiliary verbs:

$$\mathrm{VP} \rightarrow \mathrm{VB} \quad \mathrm{ADVP} \quad \mathrm{VP}$$

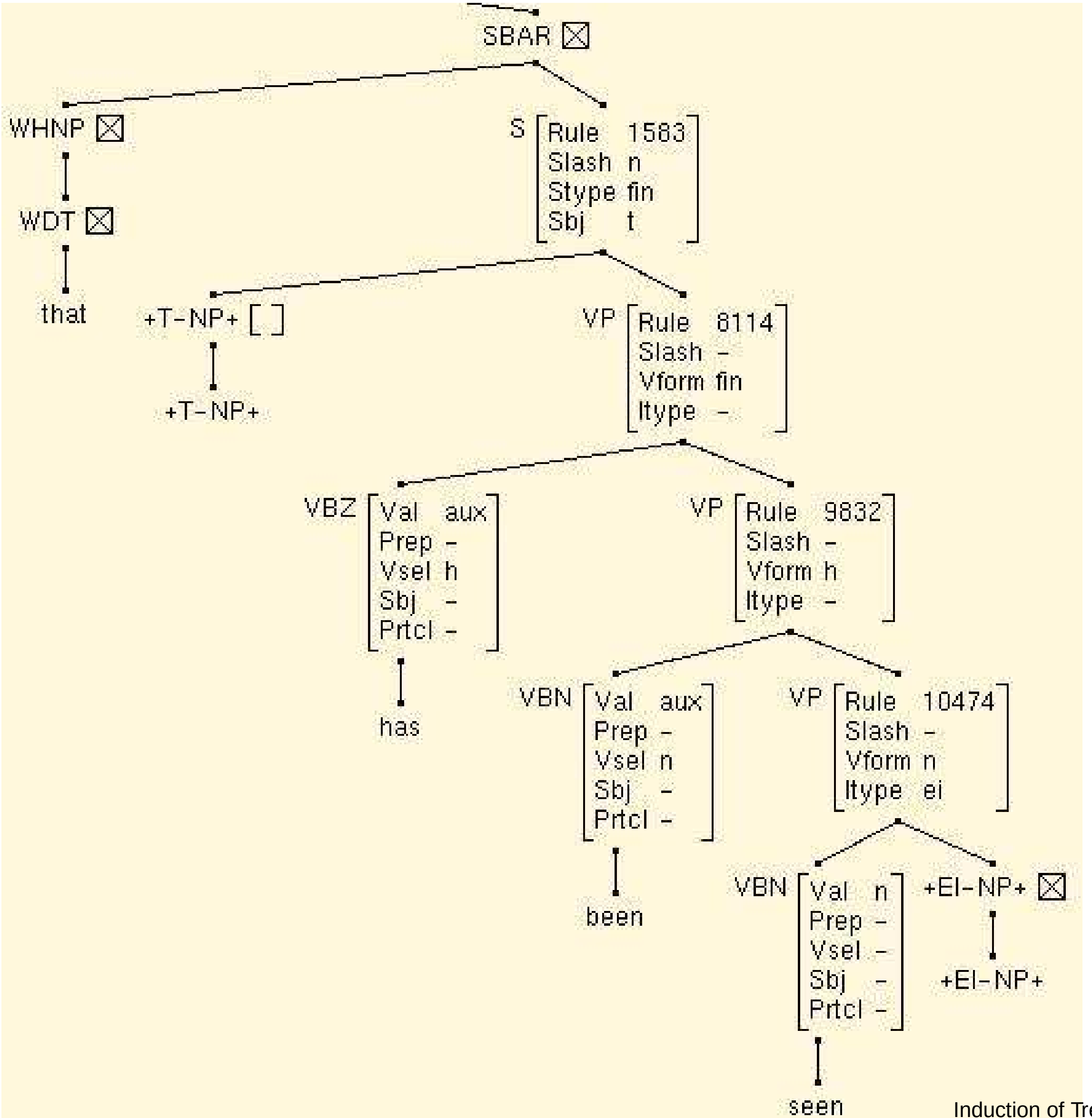
$$\mathrm{VP}\ \{\ \mathrm{Vform} = \mathrm{base};\ \} \rightarrow \mathrm{VB}\ \{\mathrm{Val} = \mathrm{aux};\} \quad \mathrm{ADVP}\ \{\ \} \quad \mathrm{VP}\ \{\ \}$$

$$\mathrm{VP}\ \{\mathrm{Vform} = \mathrm{base};\ \mathrm{Slash} = \mathit{sl};\ \} \rightarrow \mathrm{VB}\ \{\mathrm{Val} = \mathrm{aux};\ \mathrm{Vsel} = \mathit{vf};\ \}\ \mathrm{ADVP}\ \{\ \}\ \mathrm{VP}\ \{\ \mathrm{Slash} = \mathit{sl};\ \mathrm{Vform} = \mathit{vf}\ \}$$

$$\mathrm{VP}\ \{\mathrm{Vform} = \mathrm{base};\ \mathrm{Slash} = \mathit{sl};\ \} \rightarrow \mathrm{VB}\ \{\mathrm{Val} = \mathrm{aux};\ \mathrm{Vsel} = \mathit{vf};\ \mathrm{Prep} = -;\ \mathrm{Prtcl} = -;\ \mathrm{Sbj} = -;\ \}\ \mathrm{ADVP}\ \{\ \}\ \mathrm{VP}\ \{\mathrm{Slash} = \mathit{sl};\ \mathrm{Vform} = \mathit{vf}\}$$

# Relative Clause

..that has been seen.



# Verbal Subcategorization Features

VP $\longrightarrow$ VBD +EI-NP+ S

# Verbal Subcategorization Features

$$\text{VP} \quad \longrightarrow \quad \text{VBD} \quad\quad \text{+EI-NP+} \quad\quad \text{S}$$

$$\text{VP\{ Vform = ns; \}} \quad \longrightarrow \quad \text{VBD \{ Val = ns; \}} \quad \text{+EI-NP+} \quad \text{S \{ \}}$$

# Verbal Subcategorization Features

VP $\longrightarrow$ VBD +EI-NP+ S

VP{ Vform = ns; } $\longrightarrow$ VBD { Val = ns; } +EI-NP+ S { }

VP{ Vform = ns; } $\longrightarrow$ VBD { Val=ns; Sbj = *x*; Vsel = *vf*; }
+EI-NP+
S { Sbj= *x*; Vform = *vf*; }

# Verbal Subcategorization Features

$$\mathrm{VP} \rightarrow \mathrm{VBD} \quad \text{+EI-NP+} \quad \mathrm{S}$$

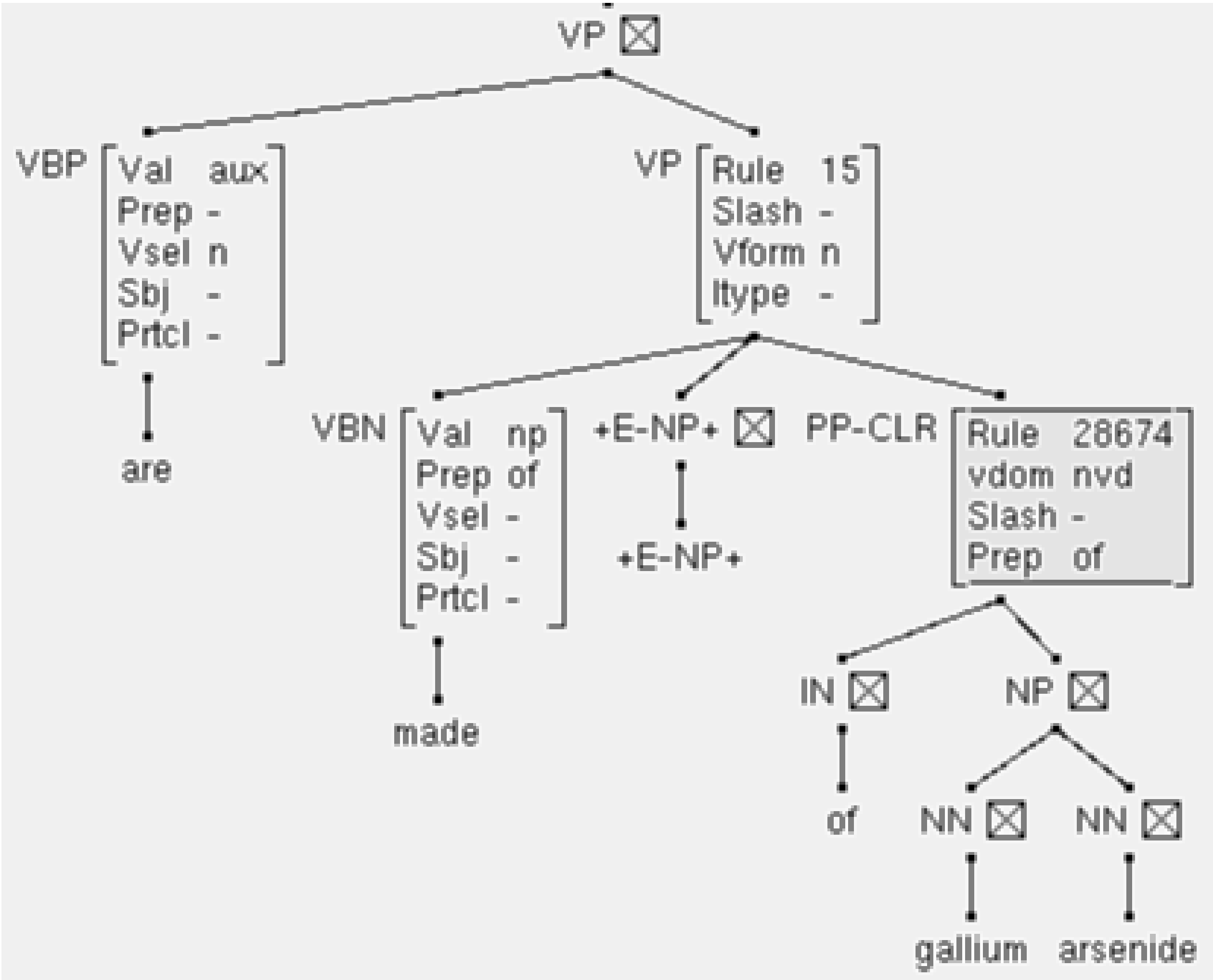
$$\mathrm{VP}\{\ \mathrm{Vform} = \mathrm{ns};\ \} \rightarrow \mathrm{VBD}\ \{\ \mathrm{Val} = \mathrm{ns};\ \} \quad \text{+EI-NP+} \quad \mathrm{S}\ \{\ \}$$

$$\mathrm{VP}\{\ \mathrm{Vform} = \mathrm{ns};\ \} \rightarrow \mathrm{VBD}\ \{\ \mathrm{Val{=}ns};\ \mathrm{Sbj} = x;\ \mathrm{Vsel} = vf;\ \}$$
$$\text{+EI-NP+}$$
$$\mathrm{S}\ \{\ \mathrm{Sbj}{=}\ x;\ \mathrm{Vform} = vf;\ \}$$

$$\mathrm{VP}\{\mathrm{Vform} = \mathrm{ns};\ \mathrm{Slash} = sl;\} \rightarrow \mathrm{VBD}\ \{\mathrm{Val{=}ns};\ \mathrm{Sbj}{=}x;\ \mathrm{Vsel}{=}vf;$$
$$\mathrm{Prep}{=}\text{-};\ \mathrm{Prtcl}{=}\text{-};\ \}$$
$$\text{+EI-NP+}$$
$$\mathrm{S}\ \{\mathrm{Sbj}{=}x;\ \mathrm{Vform}{=}vf;\ \mathrm{Slash}{=}sl;\}$$

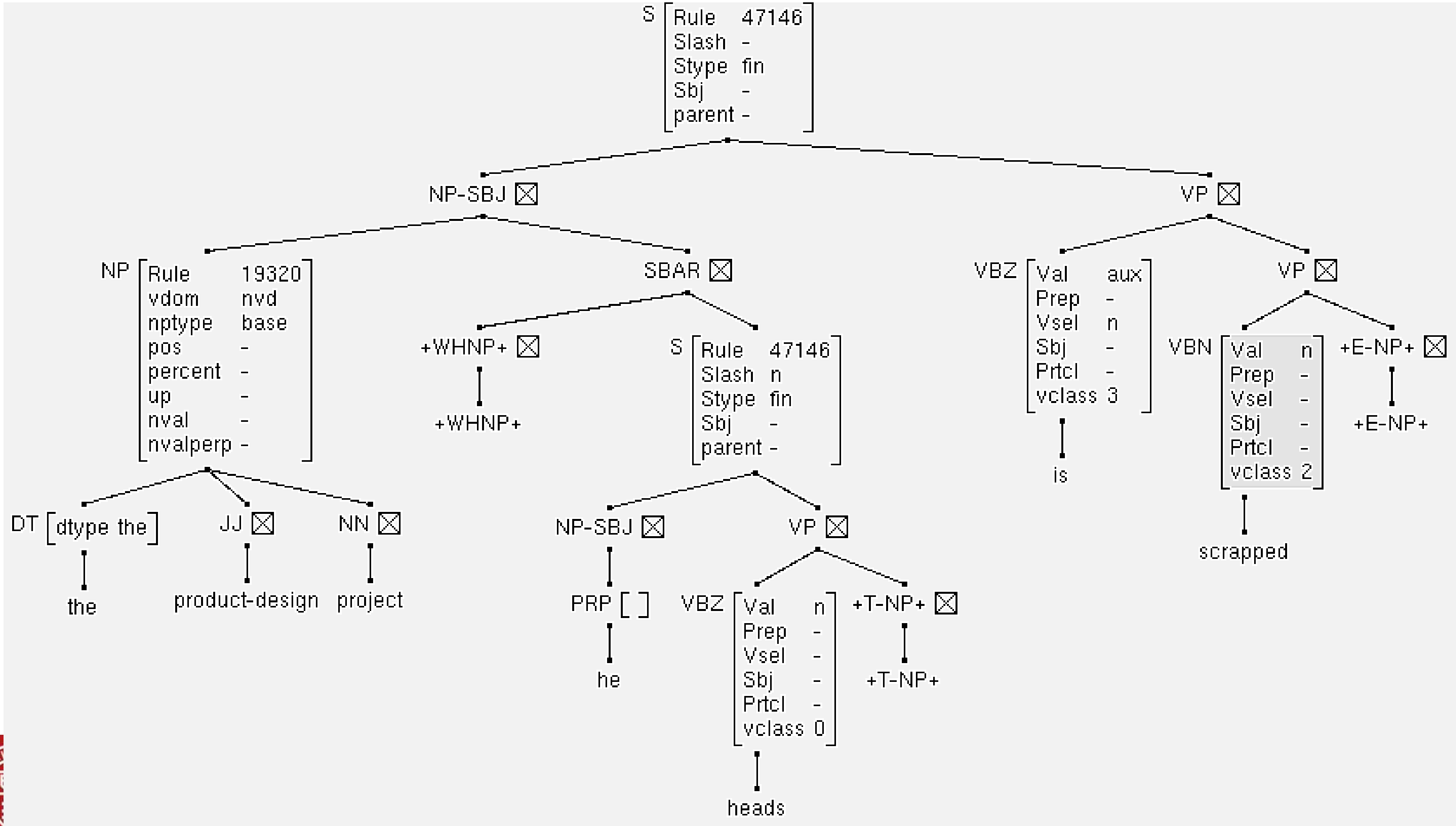
# Verbal Subcategorization

Structural information is projected onto lexical item: verbs, adverbs, nouns.



# A feature-structure Treebank Tree

The product-design project he heads is scrapped



# Treebank PCFG

- Frequencies collected from feature-annotated treebank database.
- Rule frequency table and frequency lexicon that can be used by a probabilistic parser.

# Treebank grammar and lexicon

| | | | |
|---|---|---|---|
| 29092.0 | **ROOT** | $\longrightarrow$ | **S**.fin.-.-.root |
| 14134.0 | **S**.fin.-.-.- | $\longrightarrow$ | **NP-SBJ**.nvd.base.-.-.- **VP**.fin.-.- |
| 13057.0 | **NP-SBJ**.nvd.base.-.-.- | $\longrightarrow$ | **PRP** |
| 13050.0 | **PP**.nvd.of.np | $\longrightarrow$ | **IN**.of **NP**.nvd.base.-.-.-.- |

| | |
|---|---|
| tried | **VBD**.s.e.to.- 32.0 **VBN**.s.e.to.- 11.0 **VBN**.n.-.-.- 5.0 |
| | **VBD**.z.-.-.- 1.0 **VBD**.n.-.-.- 1.0 **VBD**.s.e.g.- 1.0 |
| | **VBN**.z.-.-.- 1.0 |
| admired | **VBD**.n.-.- 1.0 |
| admit | **VB**.z.-.- 1.0 **VB**.n.-.- 1.0 **VB**.b.-.- 3.0 |
| | **VBP**.z.-.- 1.0 **VBP**.p.-.- 1.0 **VBP**.b.-.- 2.0 |
| admonishing | **VBG**.s.-.to 1.0 |

# Treebank PCFG

- PCFG of variable granularity, based on attributes incorporated into the PCFG symbols.

| PTB Sec 23 | No Prepositions | Prep. on verbs | Prep. on nouns |
|---|---|---|---|
| Labeled Recall | 86.5 | 86.11 | 85.98 |
| Labeled Precision | 86.7 | 86.50 | 86.3 |
| Labeled F-score | 86.6 | 86.31 | 86.14 |

Number of features on all categories: 19

Some structural features, mostly linguistic features.

# Scarcity of lexical data

In training sections of Penn Treebank, ∼45000 sentences

- Total verb types: ∼ 7450, tokens ∼125000.
- ∼ 2830 verb types with occurrence freq 1: 38% of all types, 2.37% of all tokens.

| | |
|---|---|
| admired | **VBD**.n.-.- 1.0 |
| admit | **VB**.z.-.- 1.0 **VB**.n.-.- 1.0 **VB**.b.-.- 3.0 |
| | **VBP**.z.-.- 1.0 **VBP**.p.-.- 1.0 **VBP**.b.-.- 2.0 |
| admonishing | **VBG**.s.-.to 1.0 |
| adopted | **VBN**.aux.e.fin 2.0 **VBD**.n.-.- 15.0 **VBD**.np.-.- 1.0 |
| | **VBN**.n.-.- 16.0 |

# Unsupervised Estimation

- Inside-outside estimation over an unlabeled corpus.

# Unsupervised Estimation

- Inside-outside estimation over an unlabeled corpus.
- Treebank PCFG as starting model.

# Unsupervised Estimation

- Inside-outside estimation over an unlabeled corpus.
- Treebank PCFG as starting model.
- Focus on learning lexical parameters.

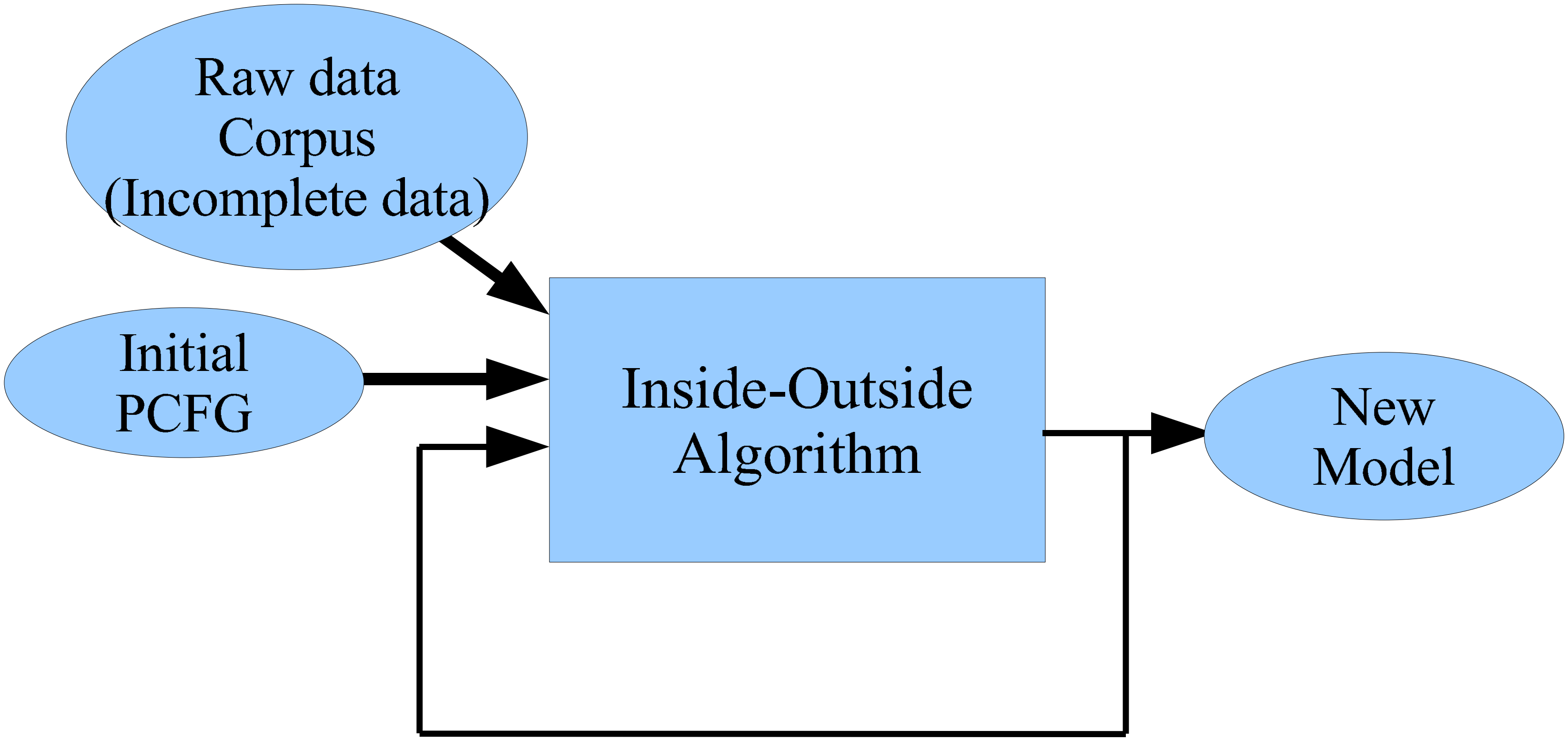
# Unsupervised Estimation

- Inside-outside estimation over an unlabeled corpus.
- Treebank PCFG as starting model.
- Focus on learning lexical parameters.
  - Lexical parameters obtained from re-estimated model and treebank.

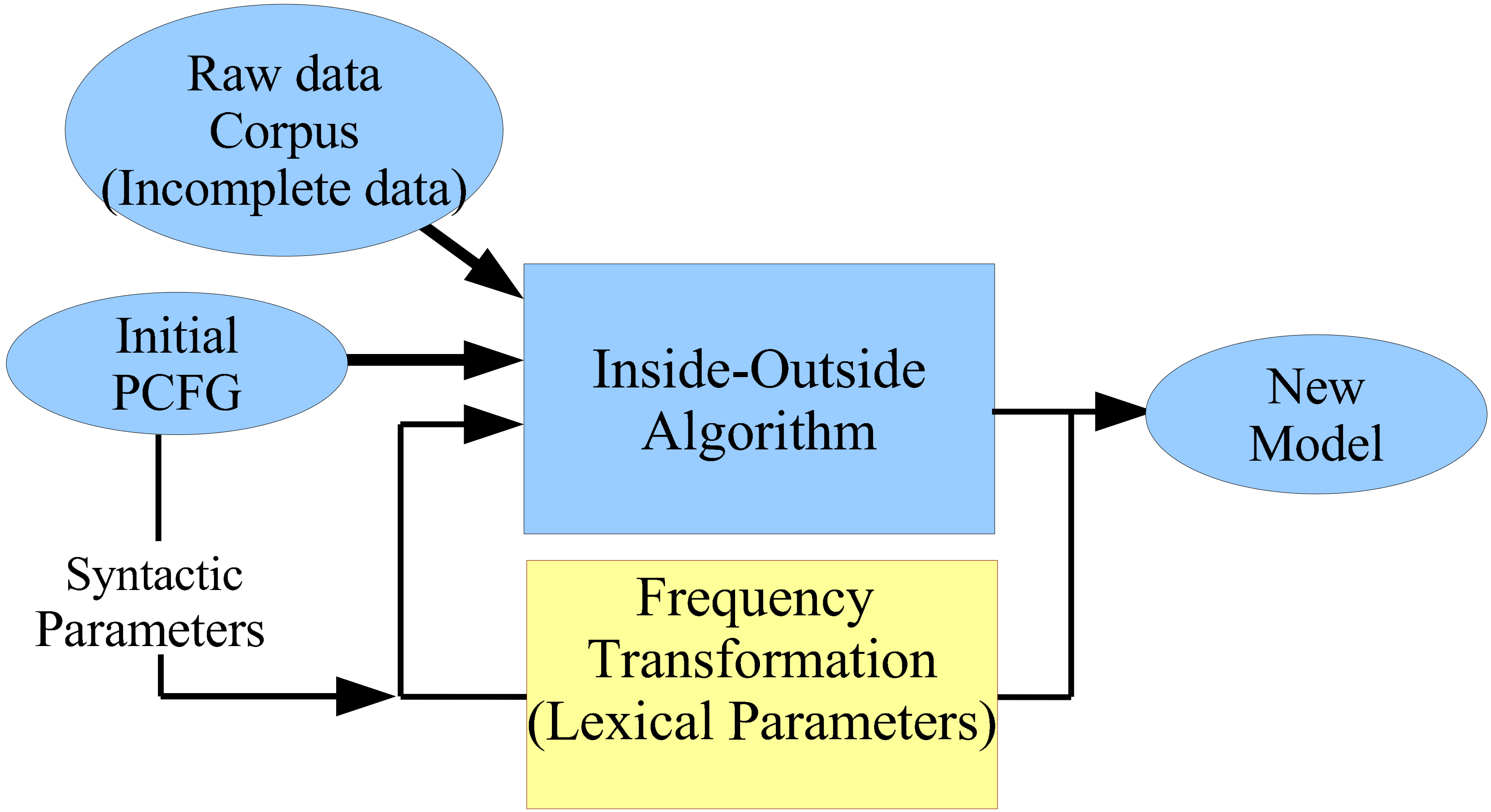
# Unsupervised Estimation

- Inside-outside estimation over an unlabeled corpus.
- Treebank PCFG as starting model.
- Focus on learning lexical parameters.
  - Lexical parameters obtained from re-estimated model and treebank.
  - Syntactic parameters obtained from treebank PCFG.

# Inside-outside Re-estimation



# Iterative Re-estimation



# Lexical Transformation

- Constraint on re-estimated lexicons.
- Ensures that re-estimated lexicons are similar to treebank lexicon.
- Linear interpolation of the treebank and the re-estimated lexicons.

(1) $$d_i(w,\tau,\iota) = (1-\lambda)t(w,\tau,\iota) + \lambda\bar{c}_i(w,\tau,\iota)$$

where

$w, \tau, \iota$ word, POS tag, incorporation sequence

Scale Corpus frequencies:

$$\bar{c}_i(w,\tau,\iota) = \frac{t(\tau,\iota)}{c_i(\tau,\iota)} c_i(w,\tau,\iota)$$

# Initial Model

- Non-novel words

# Initial Model

- Non-novel words
  - word-specific treebank distribution is maintained, but small frequency given to all possible incorporations.

# Initial Model

- Non-novel words
  - word-specific treebank distribution is maintained, but small frequency given to all possible incorporations.
- Novel words:

# Initial Model

- Non-novel words
  - word-specific treebank distribution is maintained, but small frequency given to all possible incorporations.
- Novel words:
  - Average treebank distribution for that tag.

# Initial Model

- Non-novel words
  - word-specific treebank distribution is maintained, but small frequency given to all possible incorporations.
- Novel words:
  - Average treebank distribution for that tag.
  - The re-estimation procedure is expected to acquire a word specific distribution.

# Initial Model

- Non-novel words
  - word-specific treebank distribution is maintained, but small frequency given to all possible incorporations.
- Novel words:
  - Average treebank distribution for that tag.
  - The re-estimation procedure is expected to acquire a word specific distribution.
- Words in the corpus (both novel and non-novel) get all possible incorporation values for the POS tag.

# Initial Model

- The unlabelled corpus is tagged with POS tags in Penn Treebank style (Treetagger) and tokens of words and POS tags are tabulated to obtain a frequency table $g(w, \tau)$.
- Each frequency $g(w, \tau)$ is split among possible incorporations $\iota$ in proportion to a ratio of marginal frequencies in $t_0$

$$g(w, \tau, \iota) = \frac{t_0(\tau, \iota)}{t_0(\tau)} g(w, \tau) \tag{2}$$

The tagged corpus is merged with the treebank corpus

$$t(w, \tau, \iota) = (1 - \lambda_{\tau,\iota}) t_0(w, \tau, \iota) + \lambda_{\tau,\iota} g(w, \tau, \iota) \tag{3}$$

# Experimental Setup

- Re-estimation: 4 Million words of unannotated Wall Street Journal Corpus (year 1994), sentence-length $<$ 25 words
- Each iteration results in a corresponding model.

# Experimental Setup

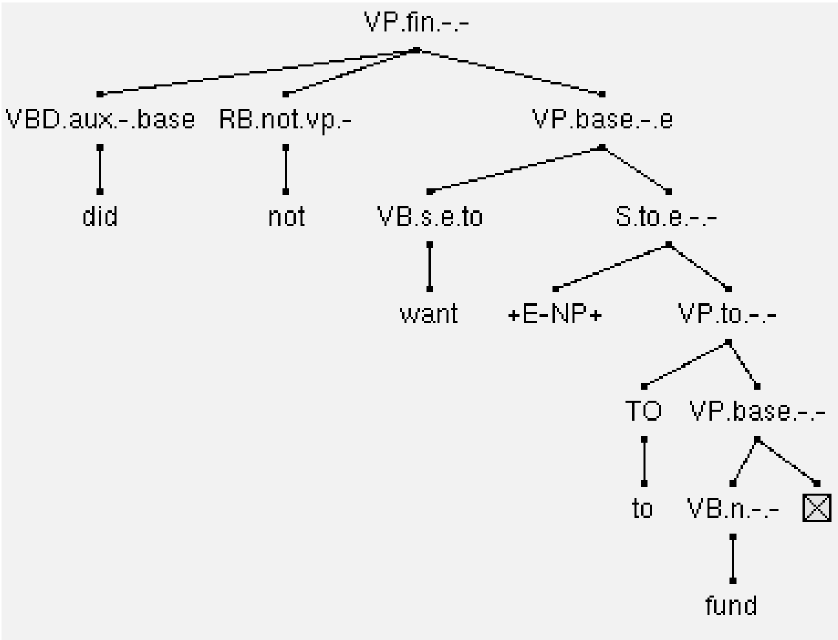
- Re-estimation: 4 Million words of unannotated Wall Street Journal Corpus (year 1994), sentence-length $<$ 25 words
- Each iteration results in a corresponding model.

Evaluation: Acquiring subcategorization frames of novel verbs.

- 1360 tokens of 117 verb types: all occurrences heldout from treebank training data.
- Tokens of test verbs : preterminal (tag + incorporation sequence) extracted from Viterbi parse.
- Gold standard is the transformed treebank.

# Subcat Frames

- Fine-grained subcategorization frames (81 subcategories)
- Intransitive, transitive, ditransitive, clausal, prepositional, etc.
- For clausal frames, the type and subject of clause.



# Subcat. error % for Novel verbs

| Iteration $i$ | Interleaved Procedure | Standard Procedure |
|---|---|---|
| $t_0$ | 33.36 | 33.36 |
| 1 | *24.40 | 28.69 |
| 2 | *23.45 | 25.56 |
| 3 | *23.05 | 27.86 |
| 4 | *22.89 | 28.41 |
| 5 | *22.81 | - |
| 6 | 22.83 | - |

10.55% absolute improvement and 31.6% error reduction

# Evaluation

Overall Error reduction: 8.97% (16.8% overall error)

# Evaluation

Overall Error reduction: 8.97% (16.8% overall error)

Incorporating Prepositions into frame

| Iteration $i$ | Subcat Error (No Prep.) | Subcat Error (Prep. on verbs) |
|---|---|---|
| $t_0$ | 33.47 | 34.98 |
| 1 | 24.40 | *25.52 |
| 2 | 23.45 | *25.04 |

# Conclusions

- Framework for adding features to Treebank PCFG: features of interest can be added.
- PCFG formalism simple, and estimation methods well defined.
- Using a Treebank-aligned grammar makes standard and reliable evaluations of re-estimated grammars possible.
- Lexical information induced for novel items; also useful for low frequency items.

# Noun Valence

- Three valences s, sbar, p
- NN and NNS (common nouns)

| Iteration $i$ | Noun valence Error |
|---|---|
| 0 | 23.13 |
| 1 | *20.35 ($p < 0.0001$) |
| 2 | 21.49 |

Table 1: Noun Valence errors, with 4M words of training data.

# Labeled Bracketing Evaluation

| Iteration $i$ | Interleaved Procedure f-score | Standard Procedure f-score |
|---|---|---|
| $t_0$ | 86.55 | 86.55 |
| 1 | 86.83 | 86.96 |
| 2 | *86.93 | 85.93 |
| 3 | *86.92 | 84.87 |
| 4 | *86.92 | 83.77 |
| 5 | 86.92 | - |
| 6 | 86.86 | - |

# Larger Training Data

| Iteration $i$ | Subcat Error 4M words | Subcat Error 8 M words |
|---|---|---|
| 0 | 33.47 | 33.47 |
| 1 | 24.40 | 24.64 |
| 2 | 23.45 | *22.26 ($>$ 95% conf.) |
| 3 | 23.05 | 22.34 |
| 4 | 22.89 | 23.05 |
| 5 | 22.81 | - |
| 6 | 22.83 | - |