# Unsupervised Learning for Natural Language Processing

Dan Klein
Computer Science Division
University of California, Berkeley

# Learning Language

Supervised NLP

Unsupervised NLP

# Unsupervised NLP

- Goal: induce linguistic structure not in the data

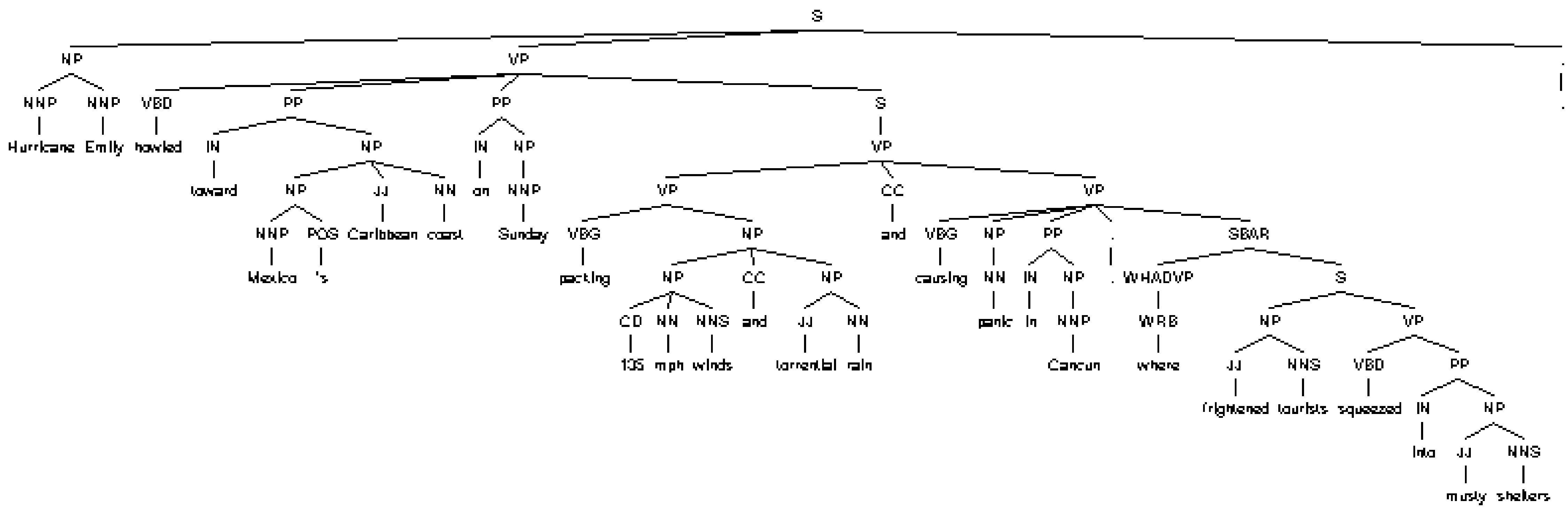
- Problem Characteristics
  - Complex linguistic phenomena
  - Rich, interacting, combinatorial structures
  - Lots of data

- Solution Characteristics
  - Incremental / hierarchical learning
  - Careful choice of what to model
  - Careful choice of what not to model

# Outline

- Unsupervised Grammar Refinement
- Unsupervised Coreference Resolution
- Unsupervised Translation Mining

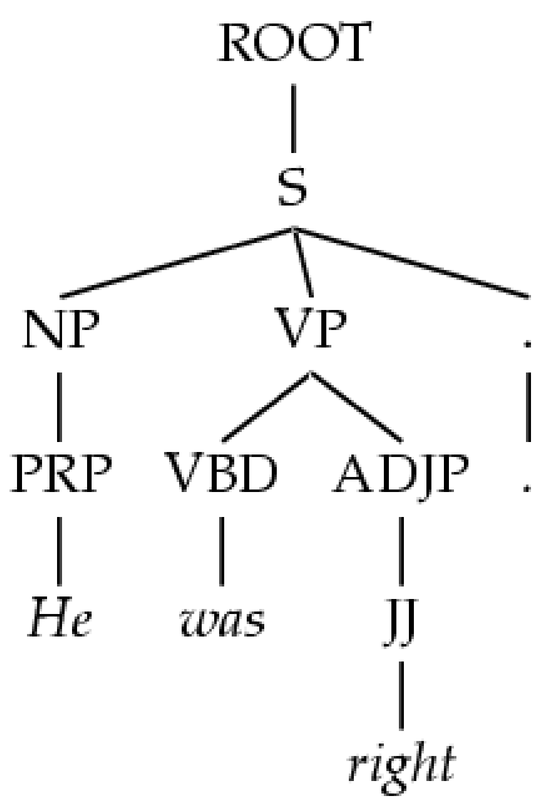
# Syntactic Analysis




Hurricane Emily howled toward Mexico 's Caribbean coast on Sunday packing 135 mph winds and torrential rain and causing panic in Cancun, where frightened tourists squeezed into musty shelters .

# Treebank PCFGs [Charniak 96]

- Use PCFGs for broad coverage parsing
- Can take a grammar right off the trees (doesn't work well):

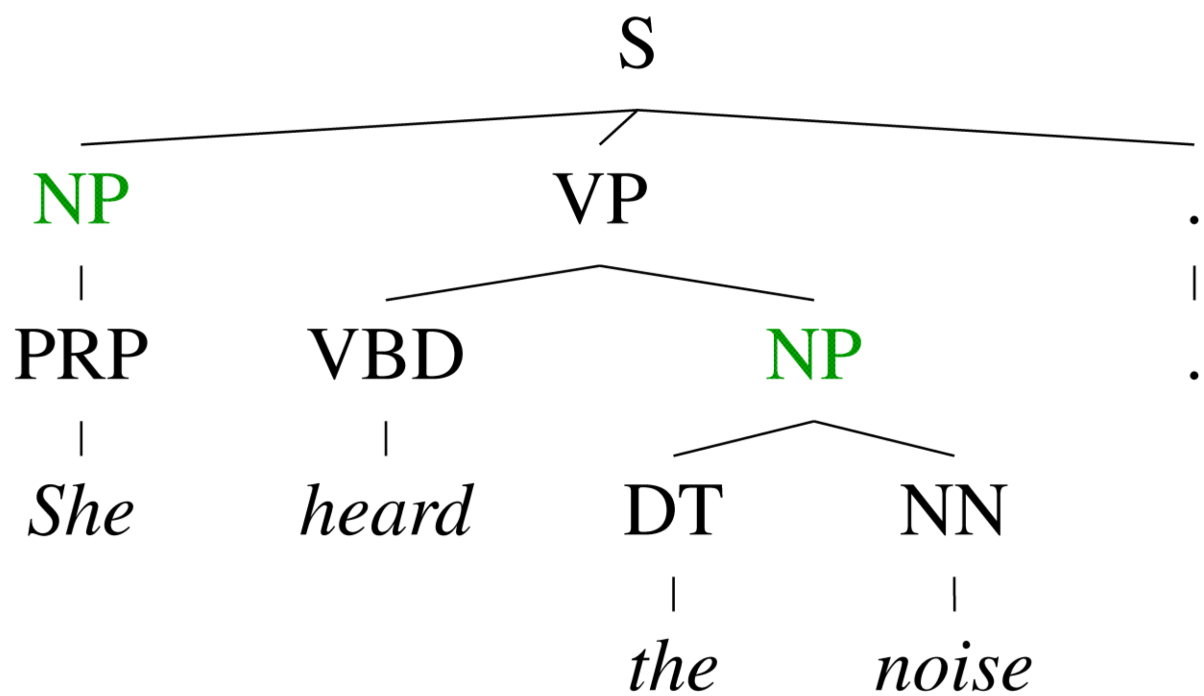



| | |
|---|---|
| ROOT → S | 1 |
| S → NP VP . | 1 |
| NP → PRP | 1 |
| VP → VBD ADJP | 1 |

…..

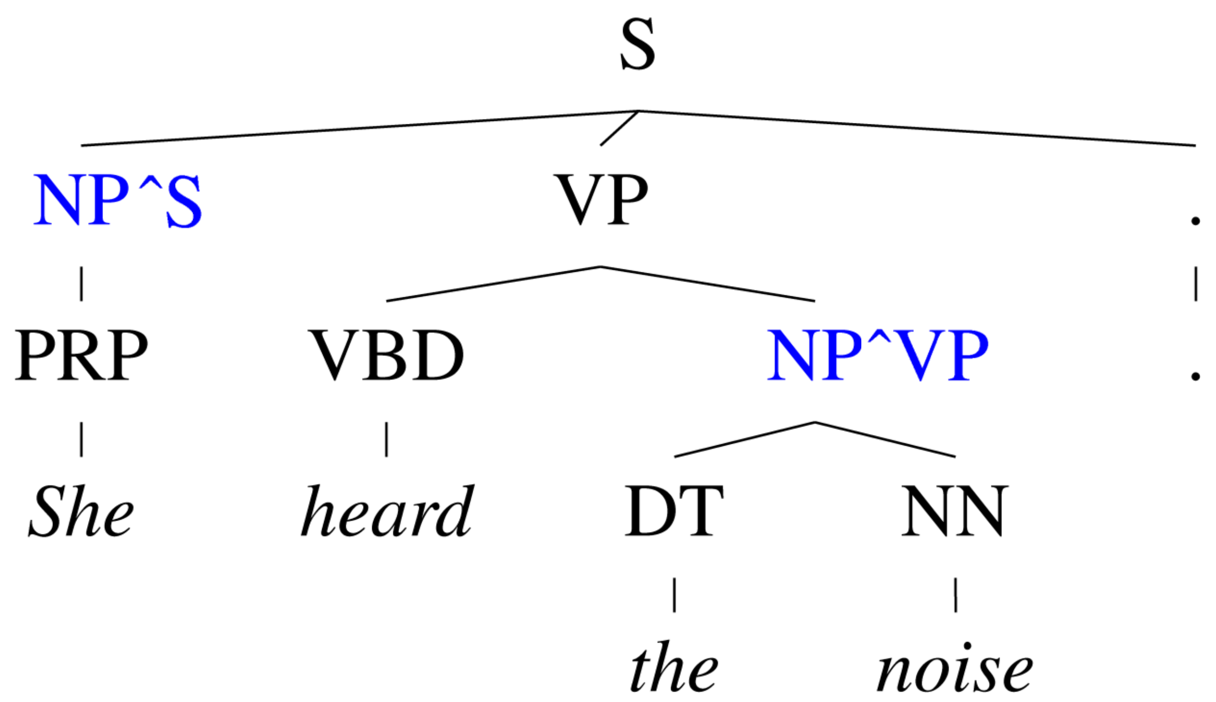
| *Model* | *F1* |
|---|---|
| Baseline | 72.0 |

# Conditional Independence?



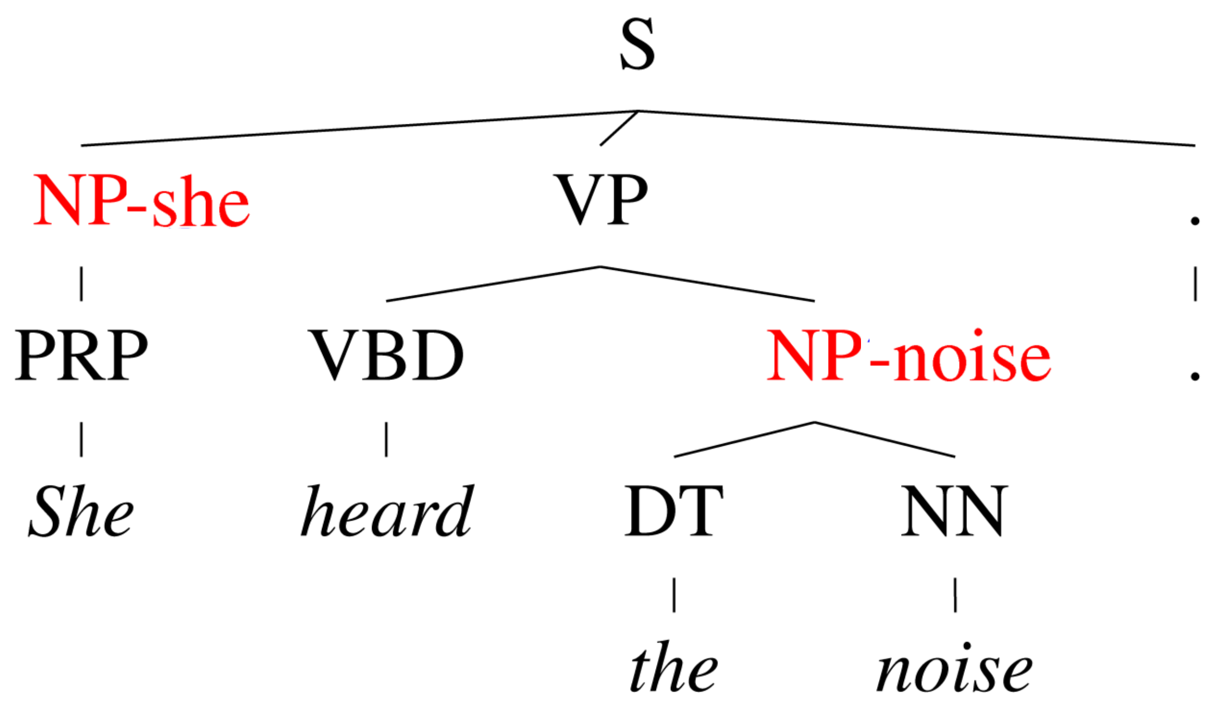

- Not every NP expansion can fill every NP slot
  - A grammar with symbols like “NP” won’t be context-free
  - Statistically, conditional independence too strong

# Grammar Refinement



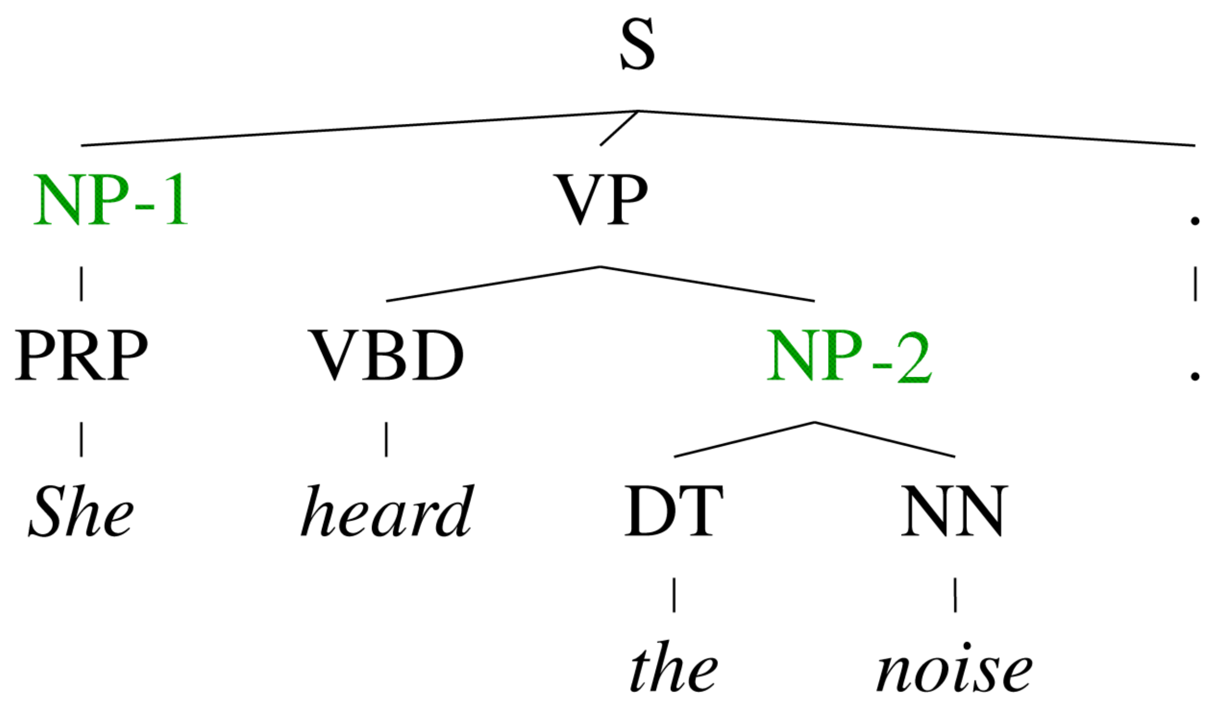

- Refining symbols improves statistical fit
  - Parent annotation [Johnson 98]

# Grammar Refinement




- Refining symbols improves statistical fit
  - Parent annotation [Johnson 98]
  - Head lexicalization [Collins 99, Charniak 00]

# Grammar Refinement

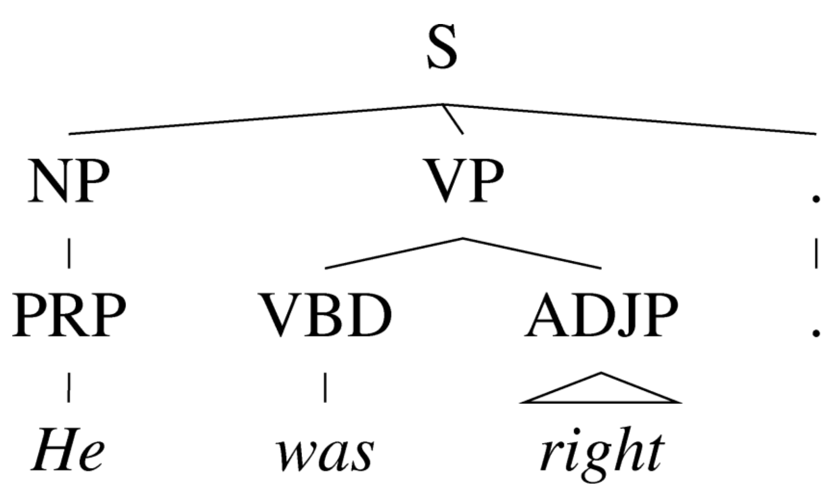


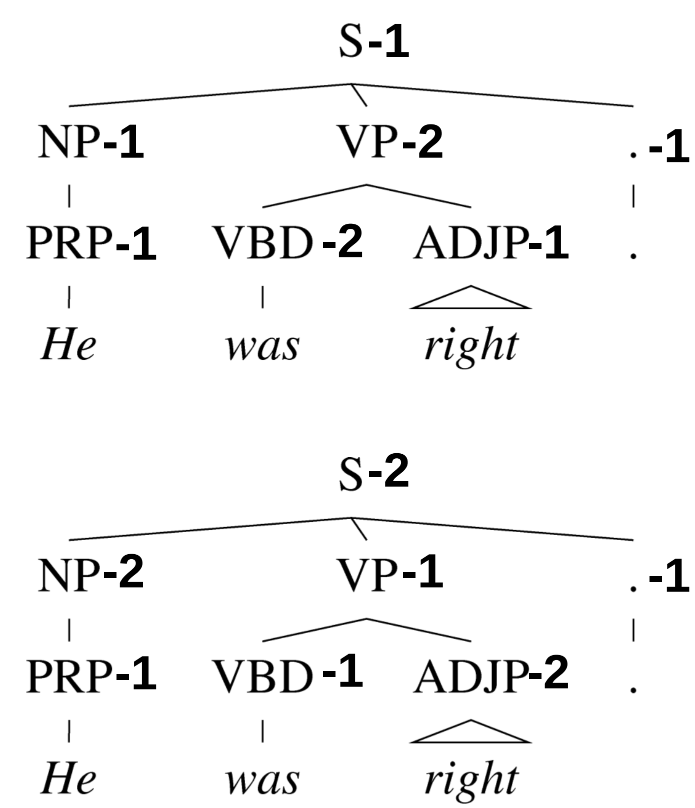

- Refining symbols improves statistical fit
  - Parent annotation [Johnson 98]
  - Head lexicalization [Collins 99, Charniak 00]
  - Automatic clustering [Petrov and Klein 06]

# Parses and Derivations

## Derivations

## Parses




Parses (T) now have multiple derivations (t)

# Training Objectives

[Matsuzaki et. al ’05, Prescher ’05]

- One option: maximum likelihood using EM
- Want derivation parameters which maximize parse likelihood

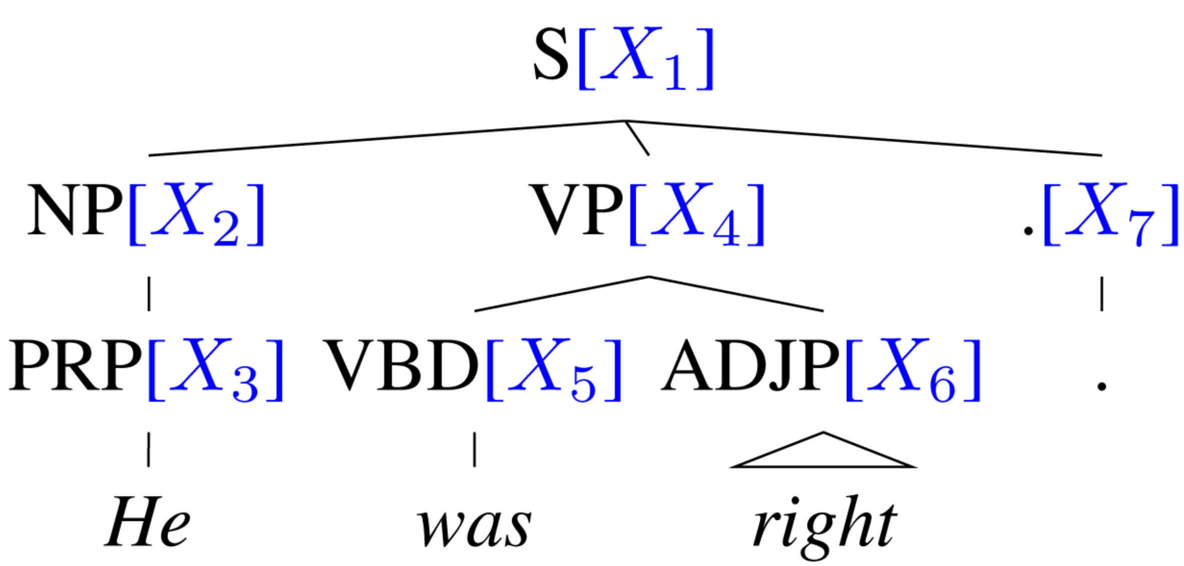
$$\max_\theta \sum_{t \in T} P(t|\theta)$$

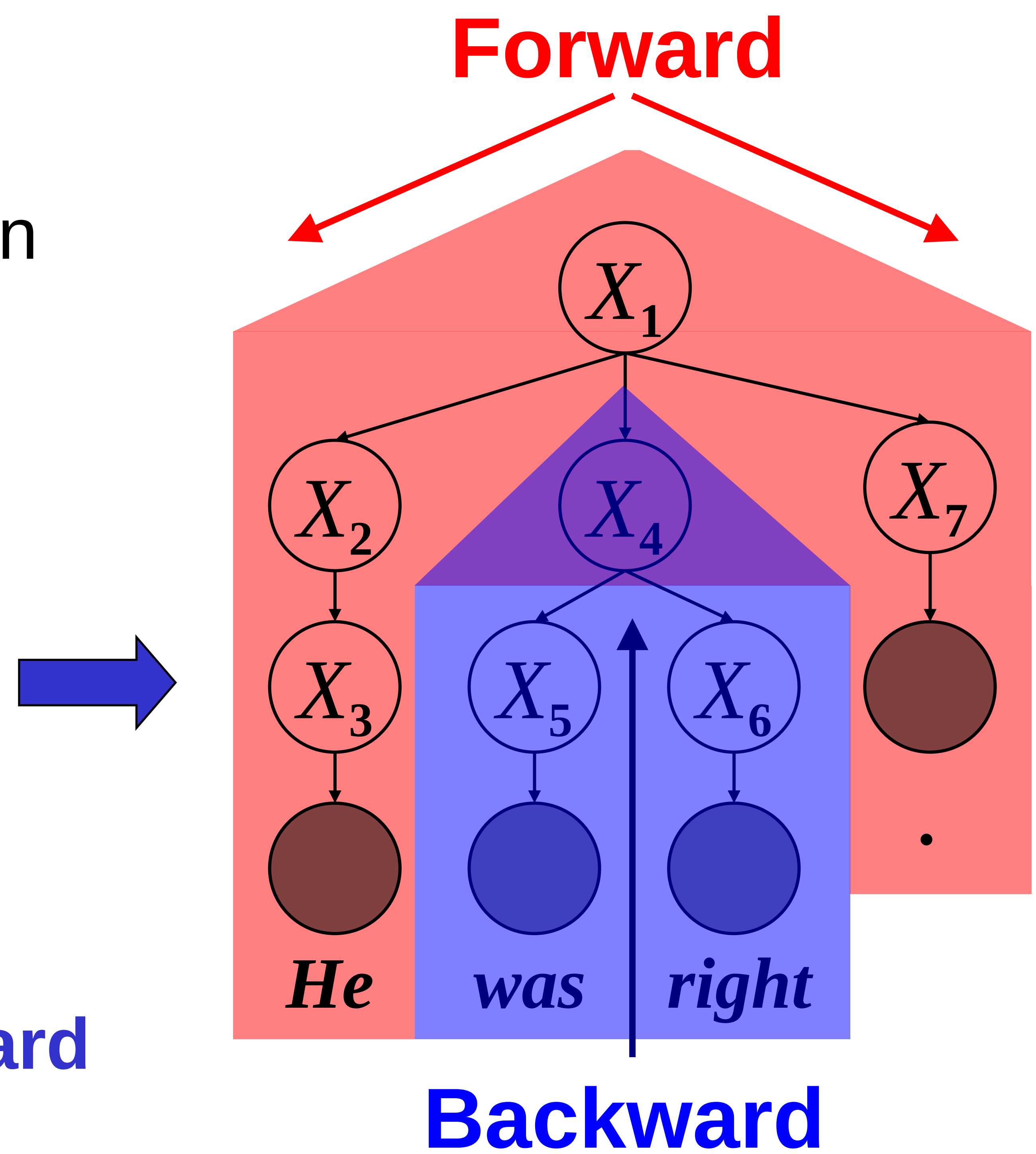
- Other options possible:
  - Variational inference [Liang et al. 07]
  - Conditional likelihood [Petrov and Klein 08]

# Learning Latent Grammars

## EM algorithm:

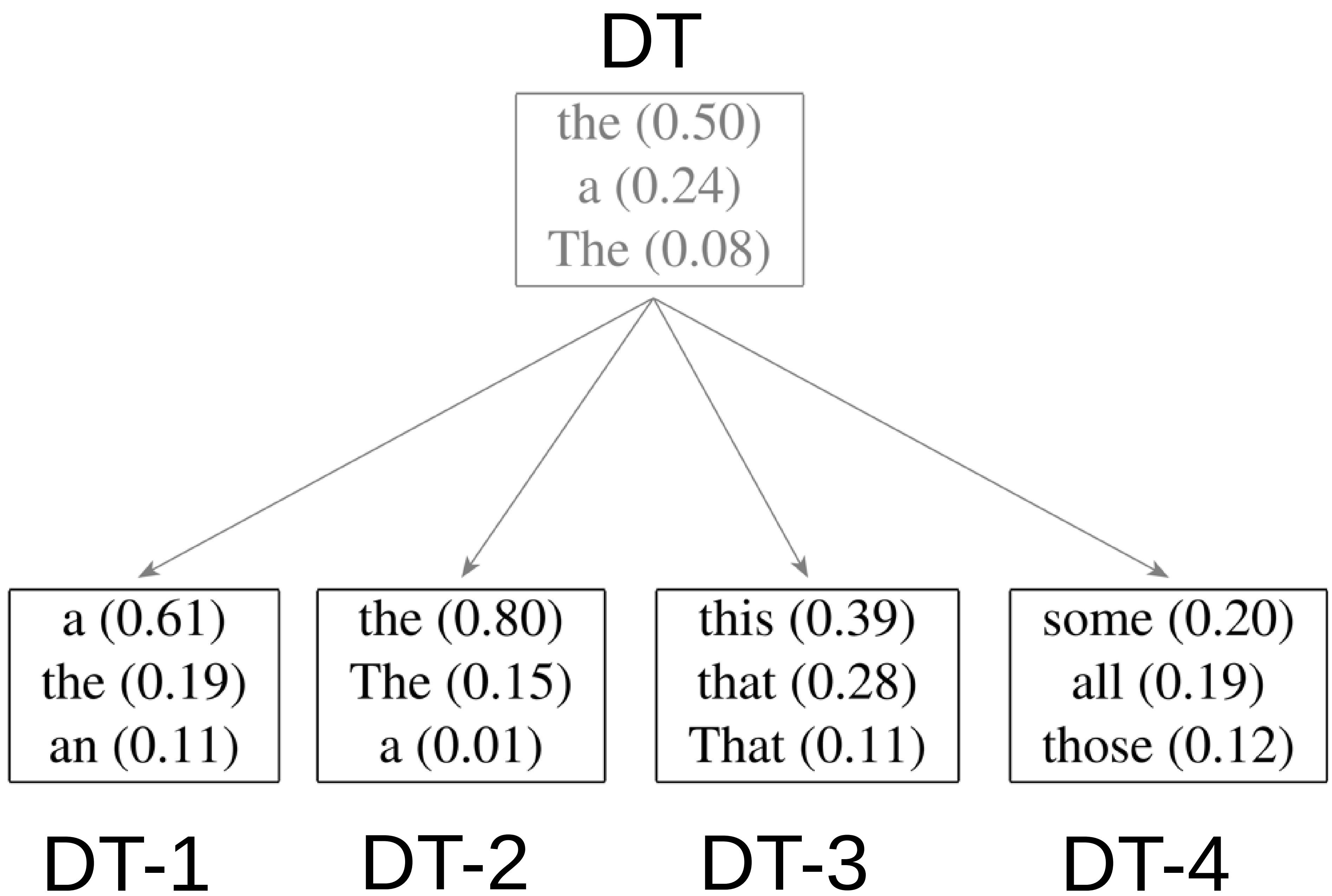
- Brackets are known
- Base categories are known
- Only induce subsymbols



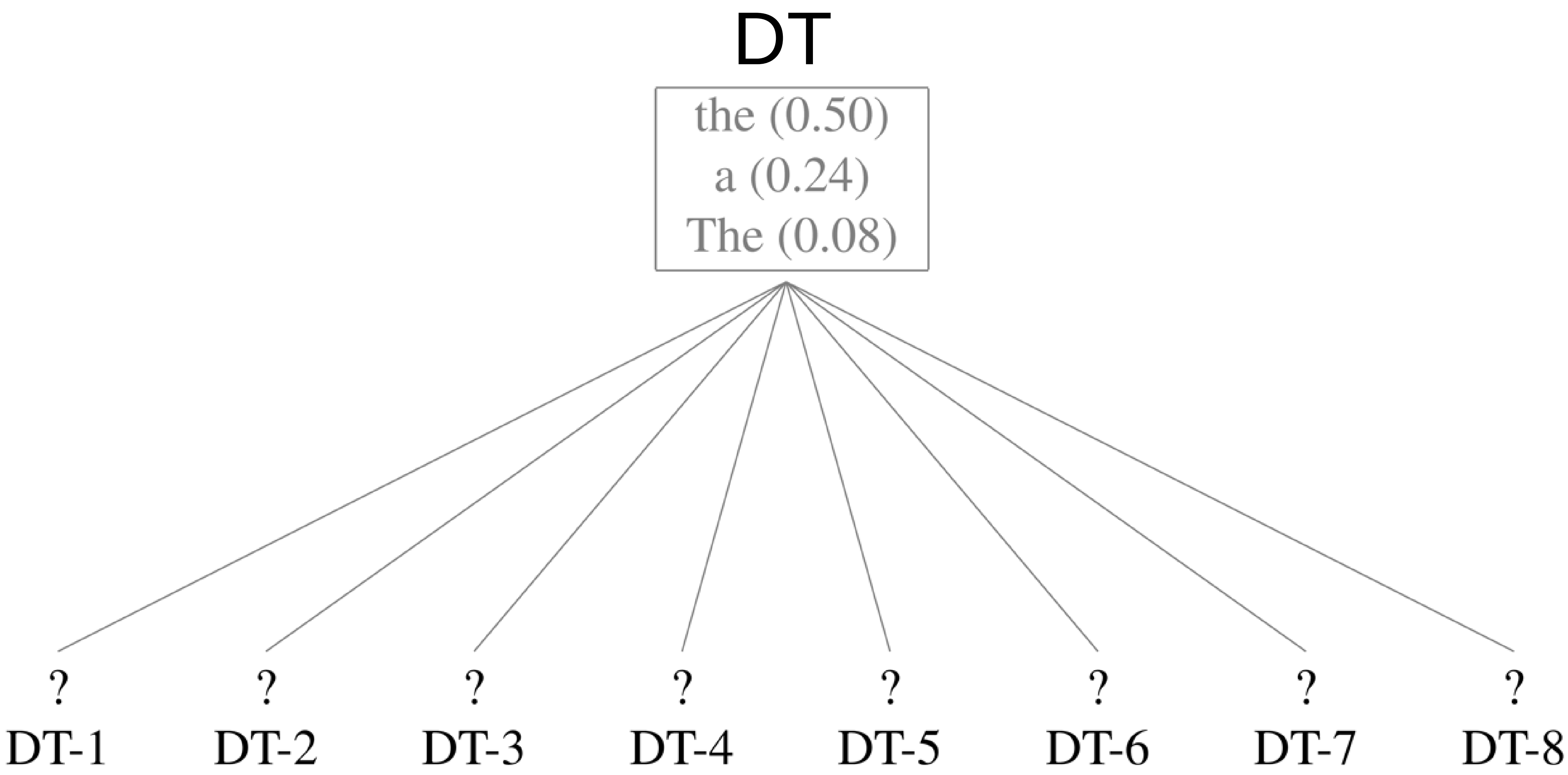




**Just like Forward-Backward for HMMs.**

# Refinement of the DT tag



# Refinement of the DT tag

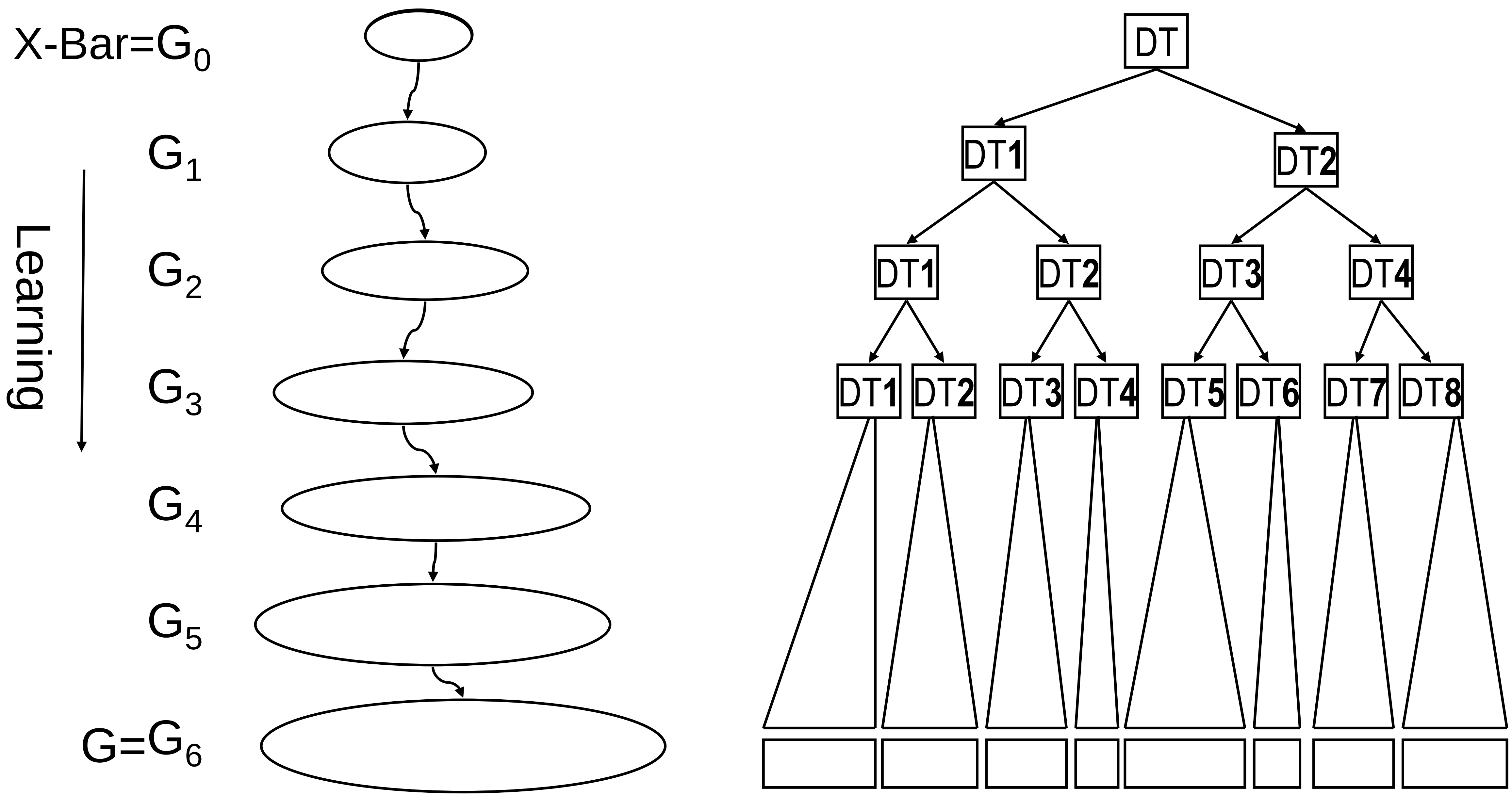


# Hierarchical Refinement

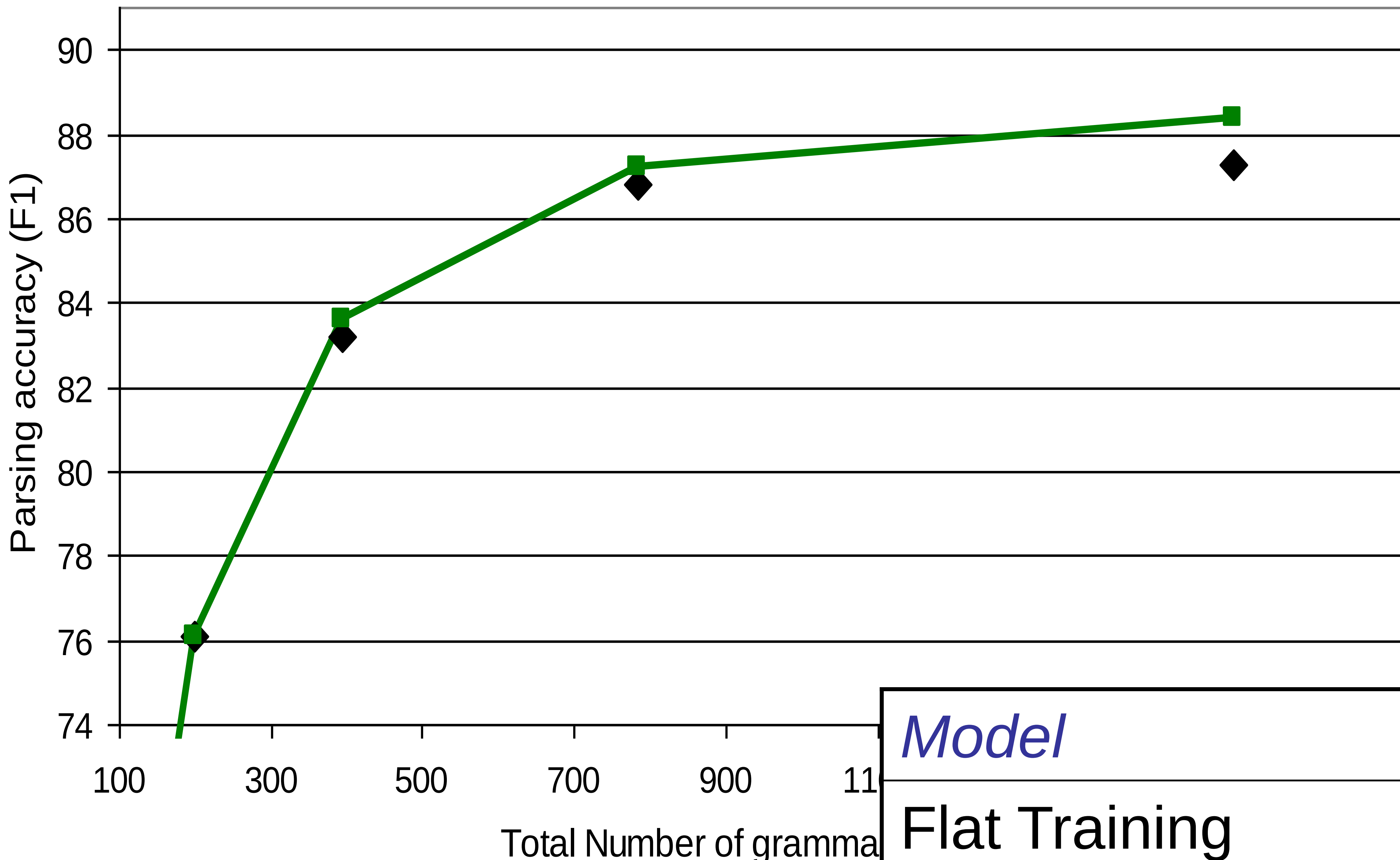
DT

| |
|---|
| the (0.50) |
| a (0.24) |
| The (0.08) |

# Grammar Ontogeny



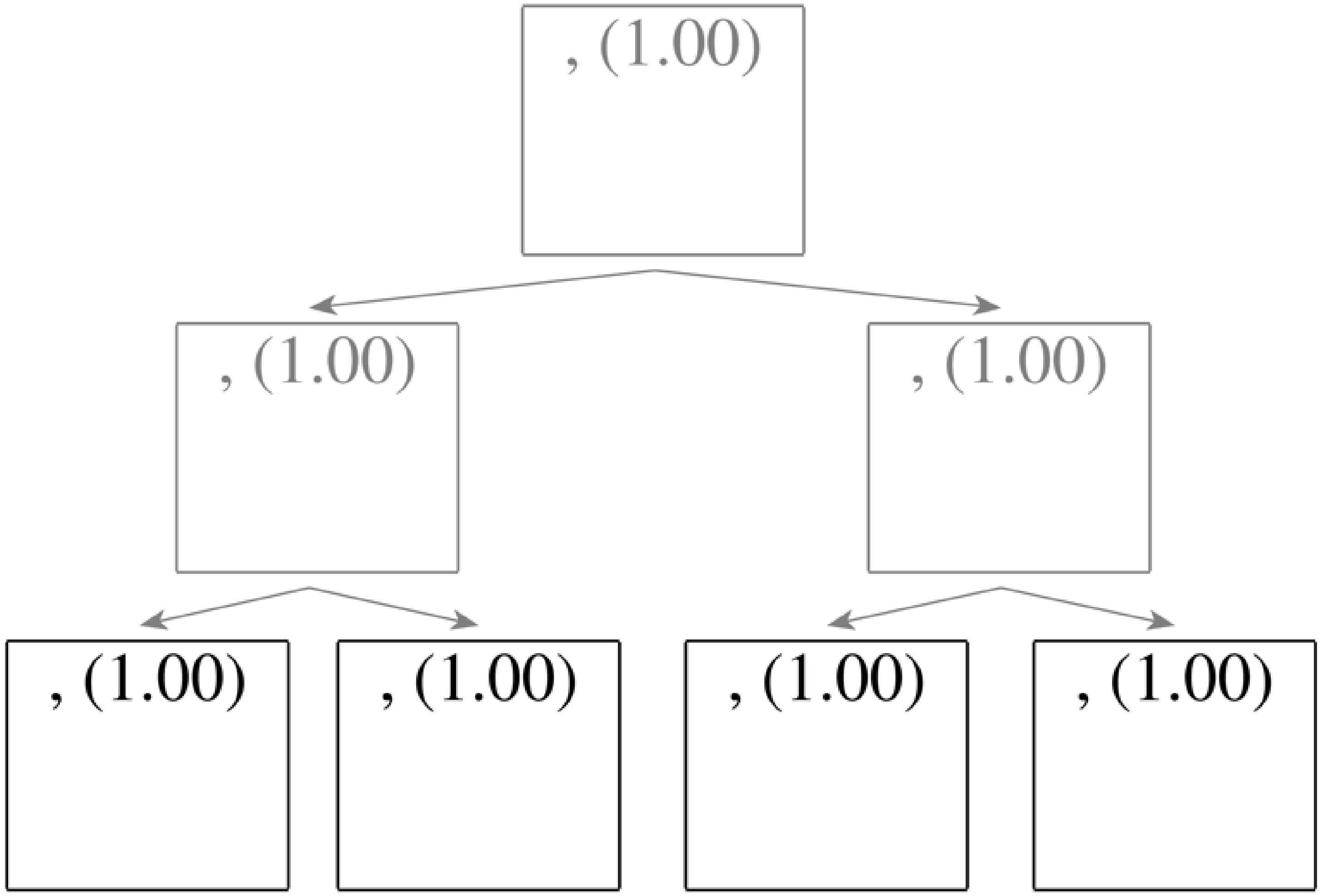
# Hierarchical Estimation Results




| *Model* | *F1* |
|---|---|
| Flat Training | 87.3 |
| Hierarchical Training | 88.4 |

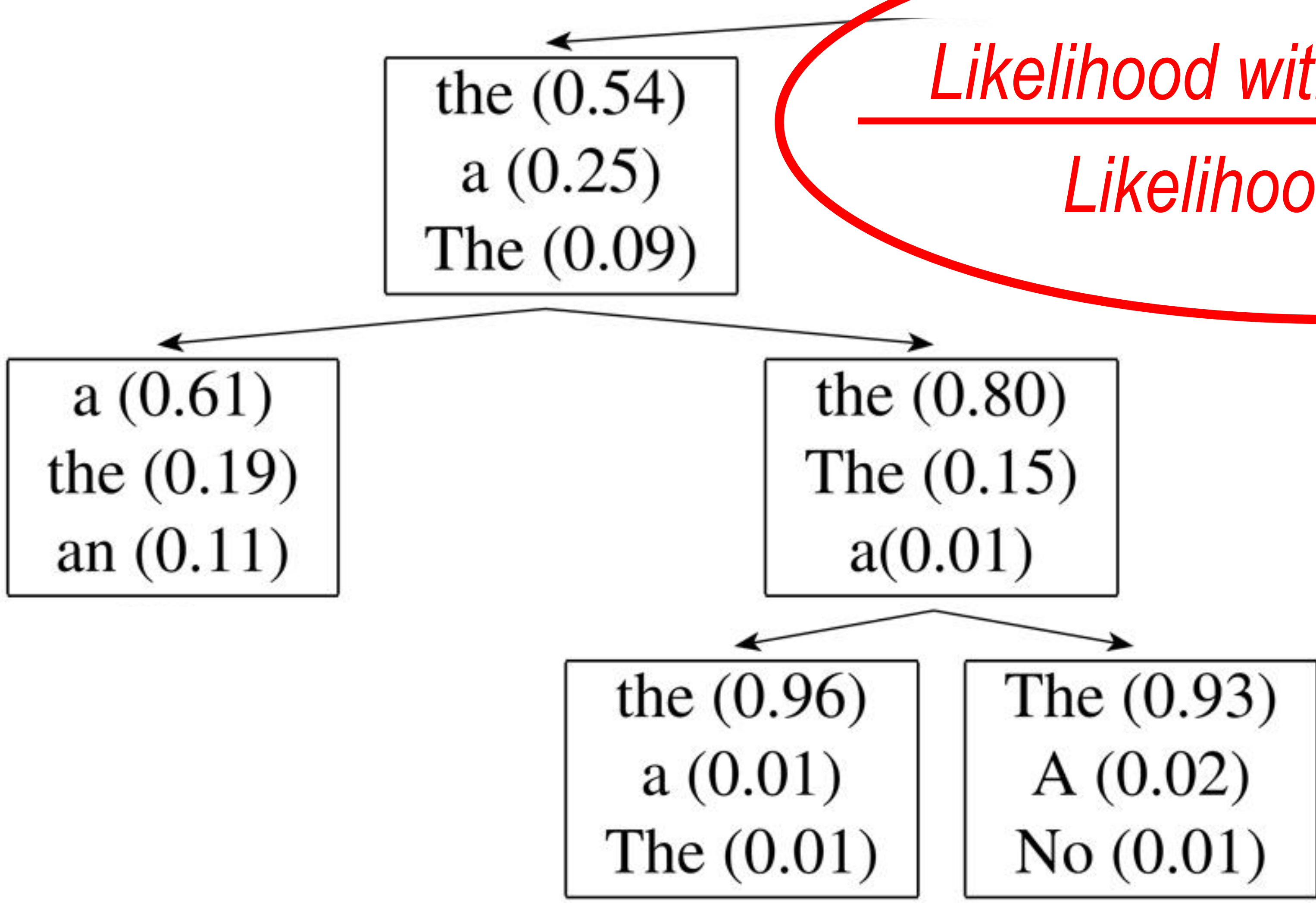
# Refinement of the , tag

- Splitting all categories equally is wasteful:



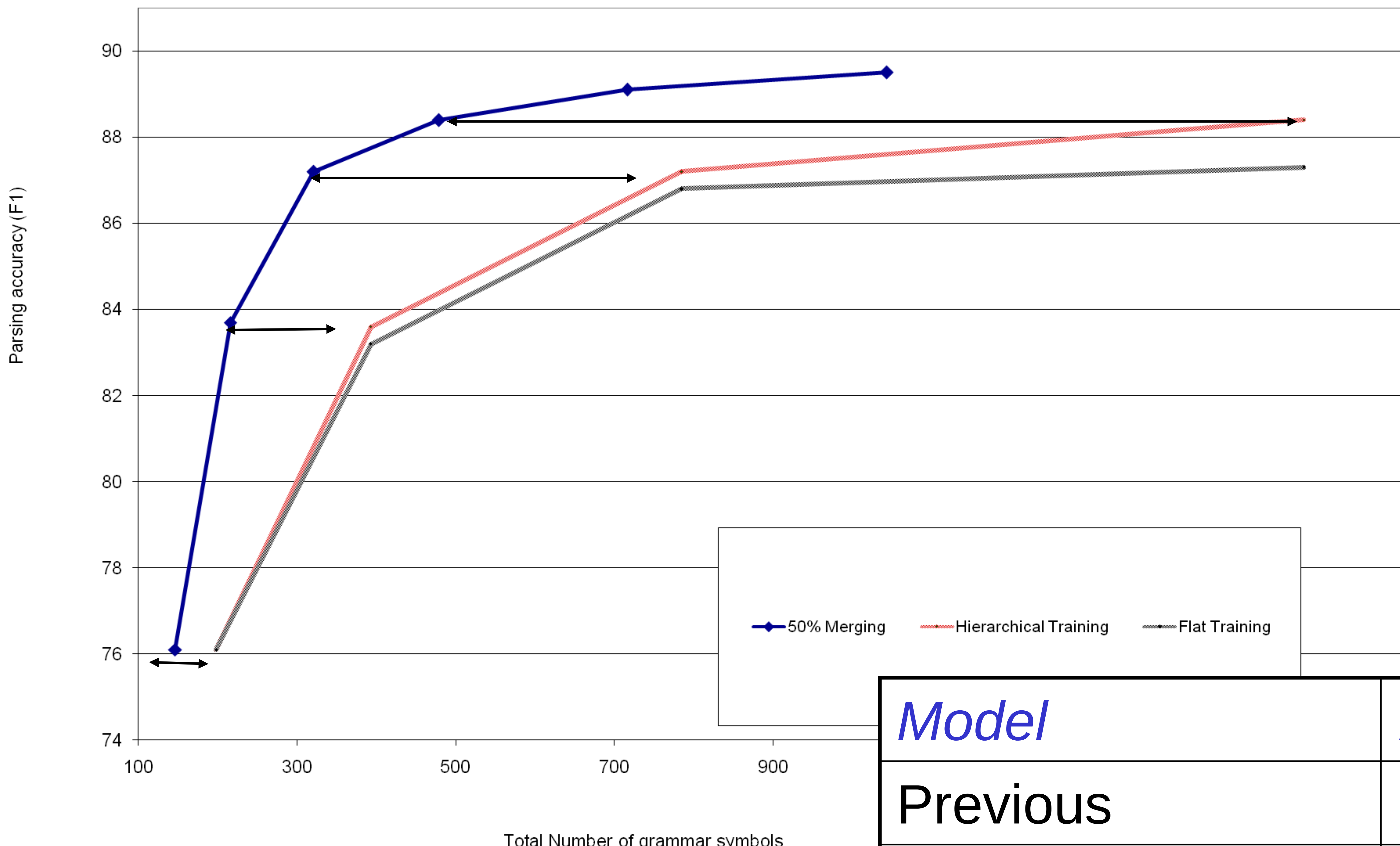
# Adaptive Splitting

- Want to split complex categories more
- Idea: split everything, roll back bad splits



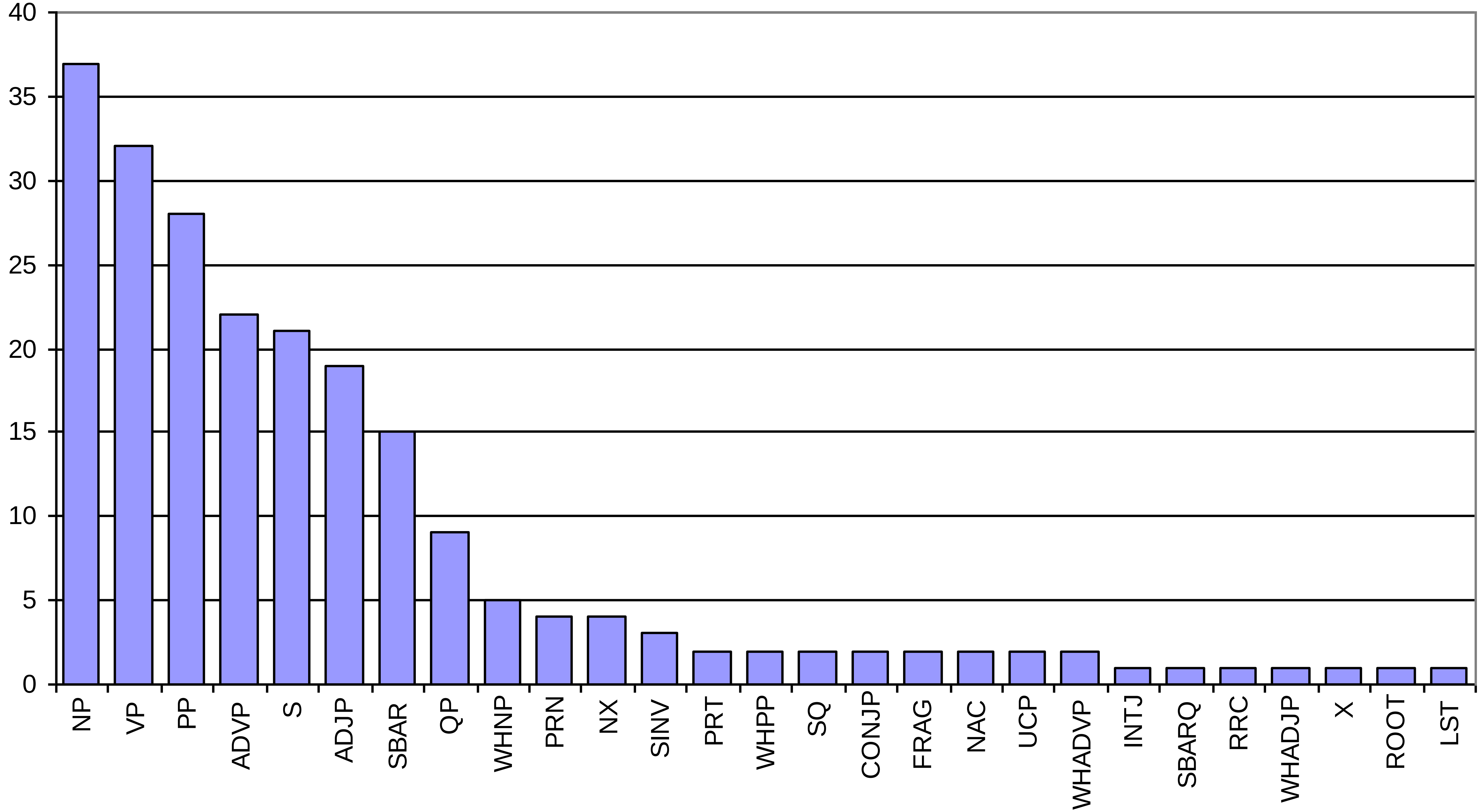

$$\frac{\textit{Likelihood with split reversed}}{\textit{Likelihood with split}}$$

# Adaptive Splitting Results

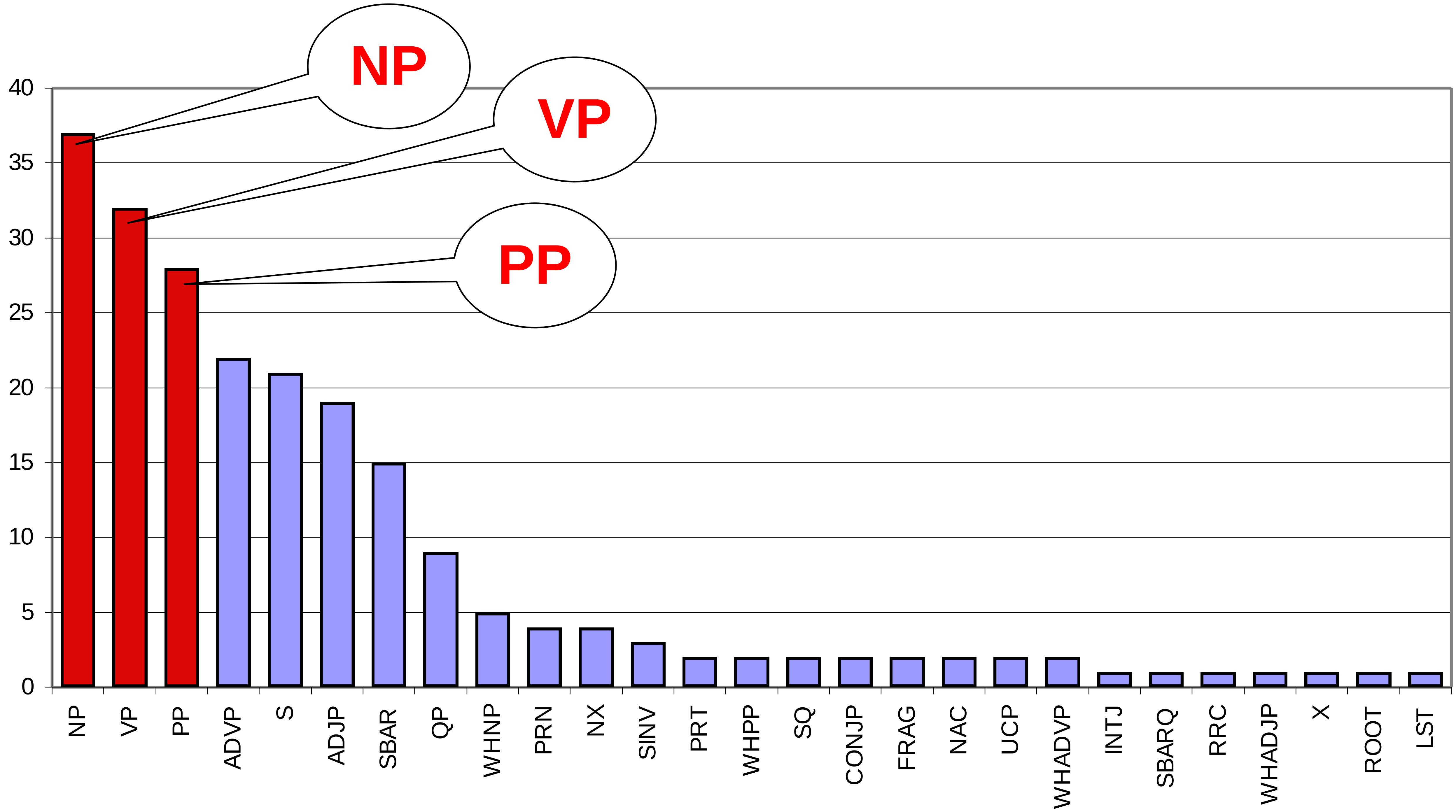



| Model | F1 |
|---|---|
| Previous | 88.4 |
| With 50% Merging | 89.5 |

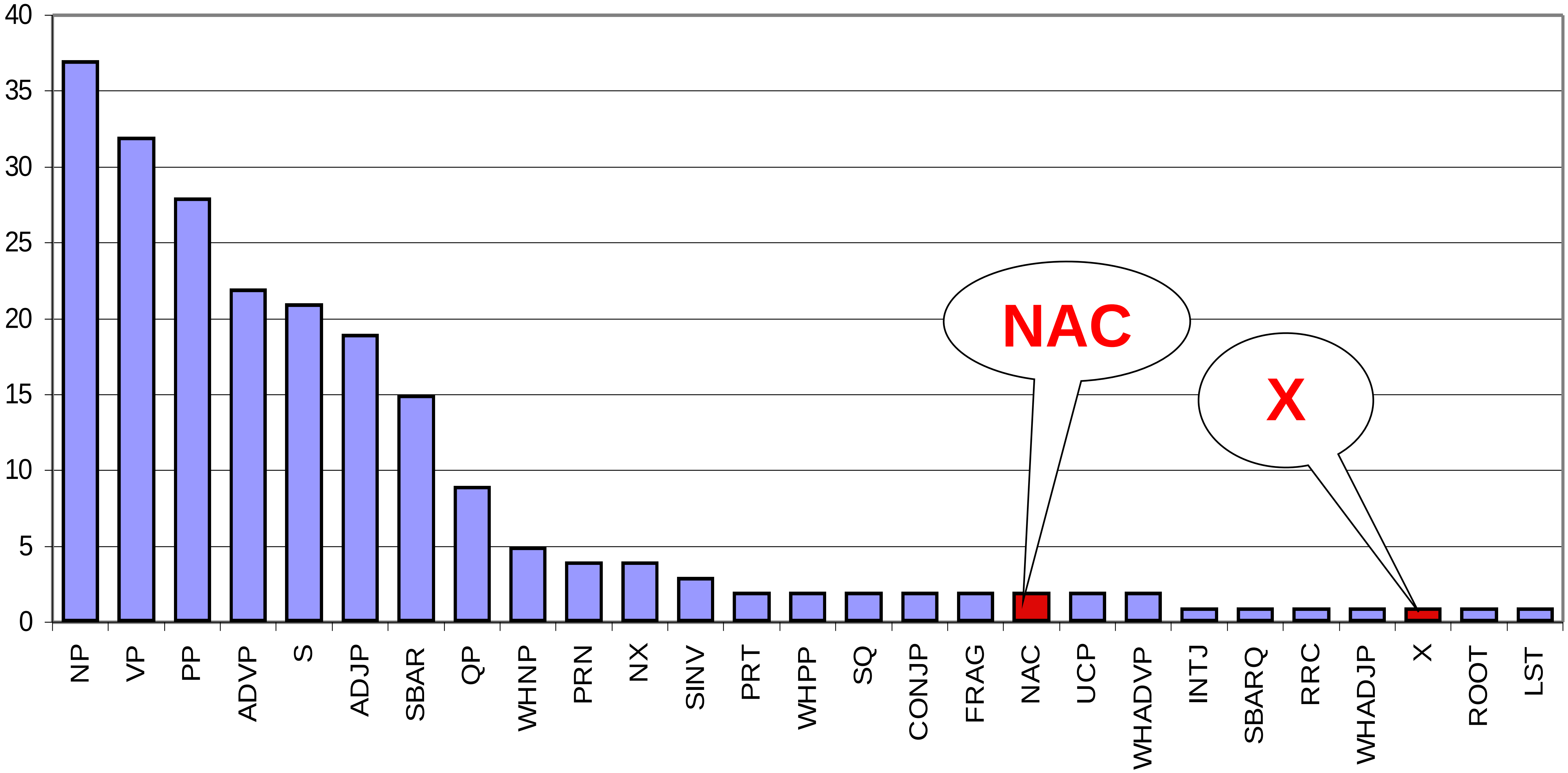
# Number of Phrasal Subcategories



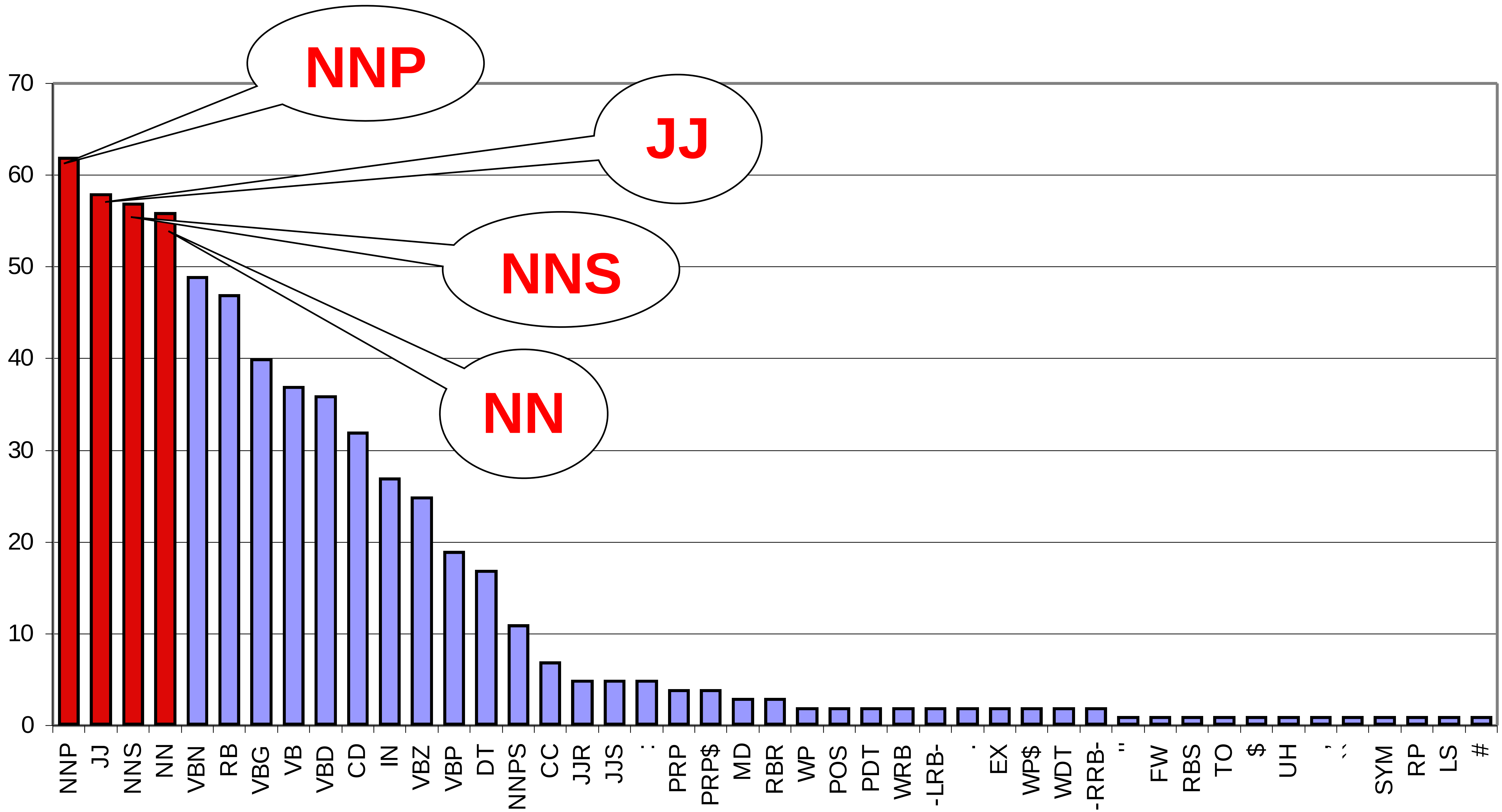
# Number of Phrasal Subcategories



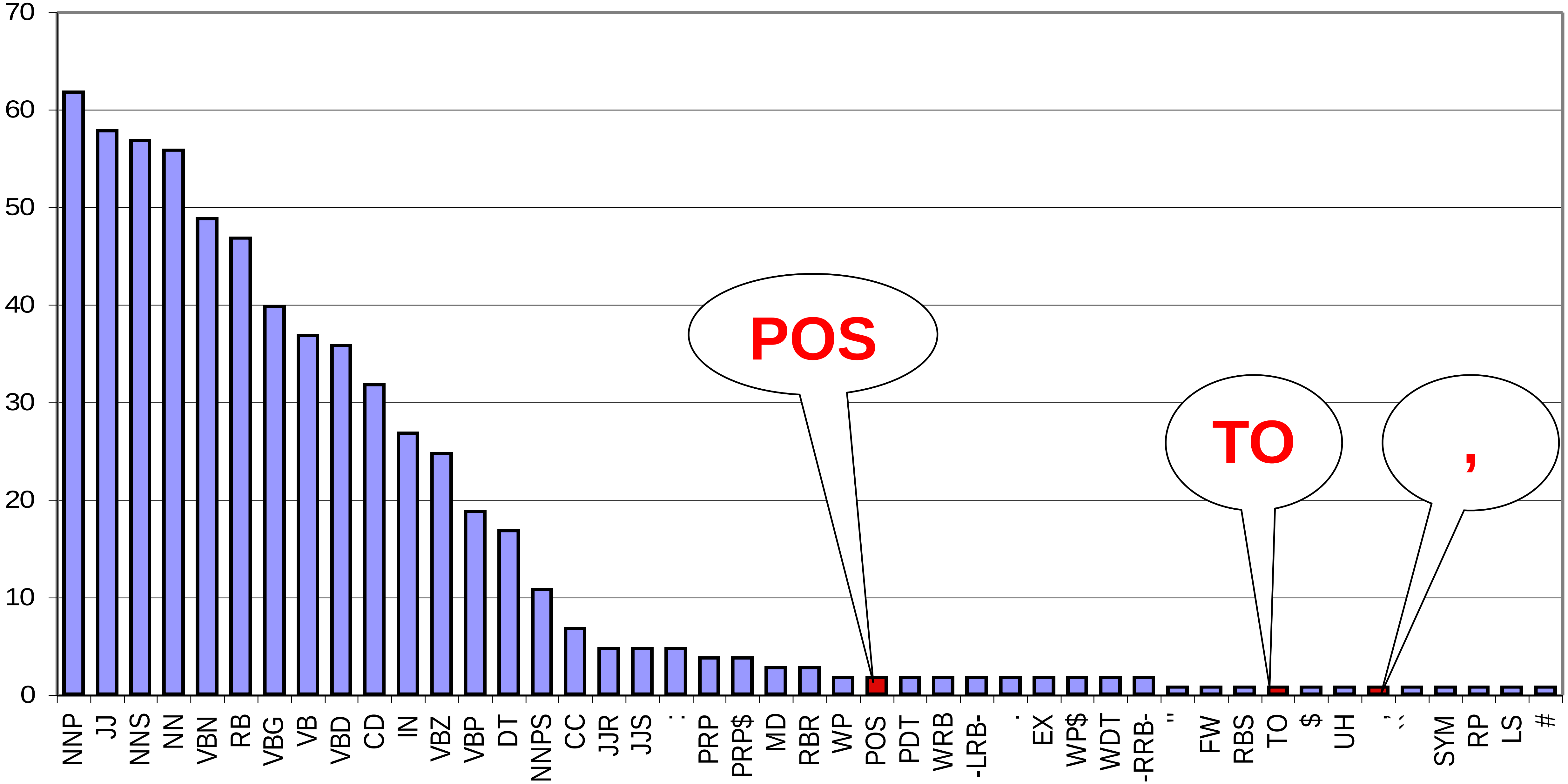
# Number of Phrasal Subcategories



# Number of Lexical Subcategories



# Number of Lexical Subcategories



# Learned Lexical Clusters

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 |

University of California C A L N L P Berkeley

# Learned Lexical Clusters

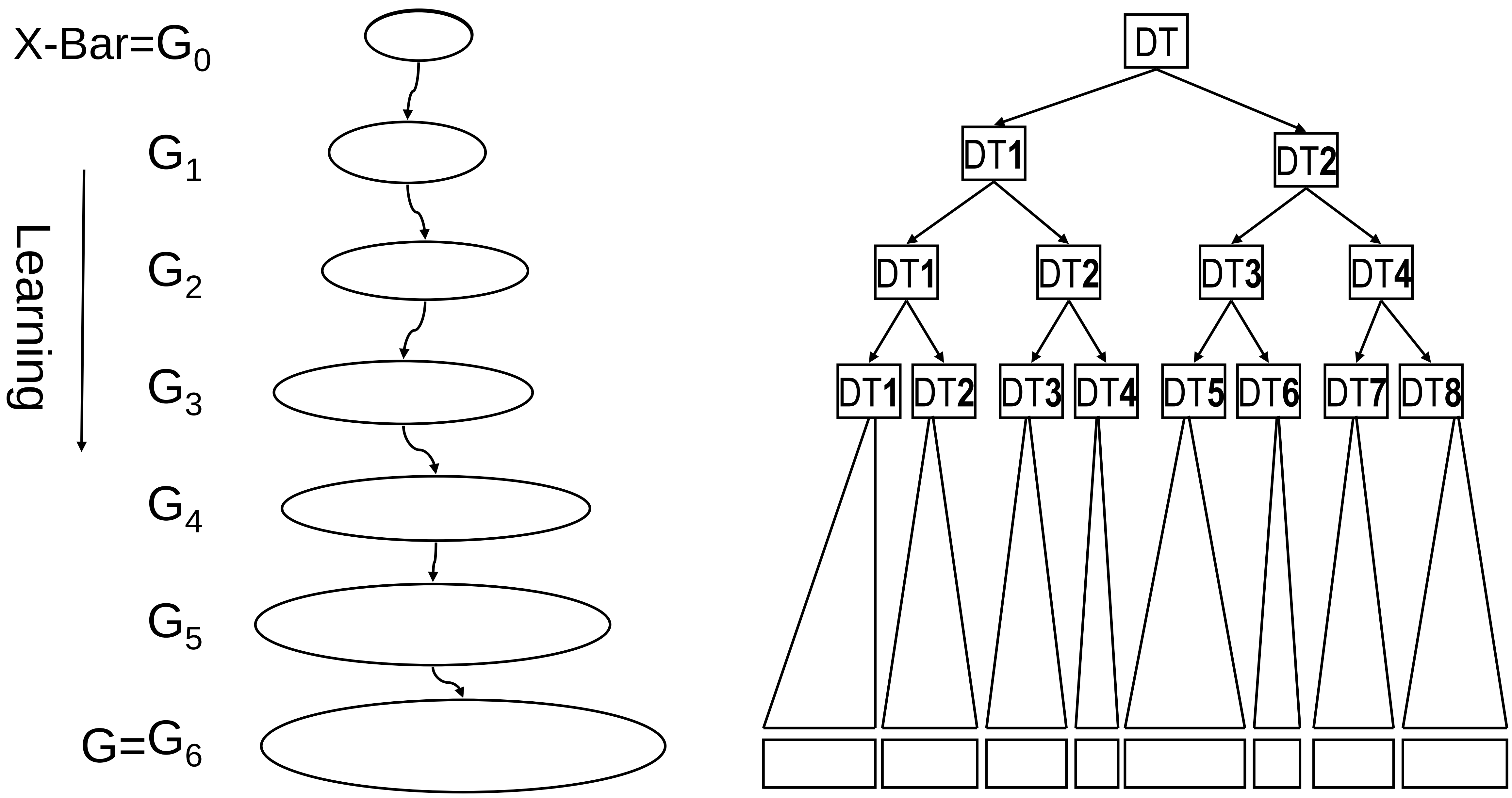
## Relative adverbs (RBR):

| | | | |
|---|---|---|---|
| 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 |

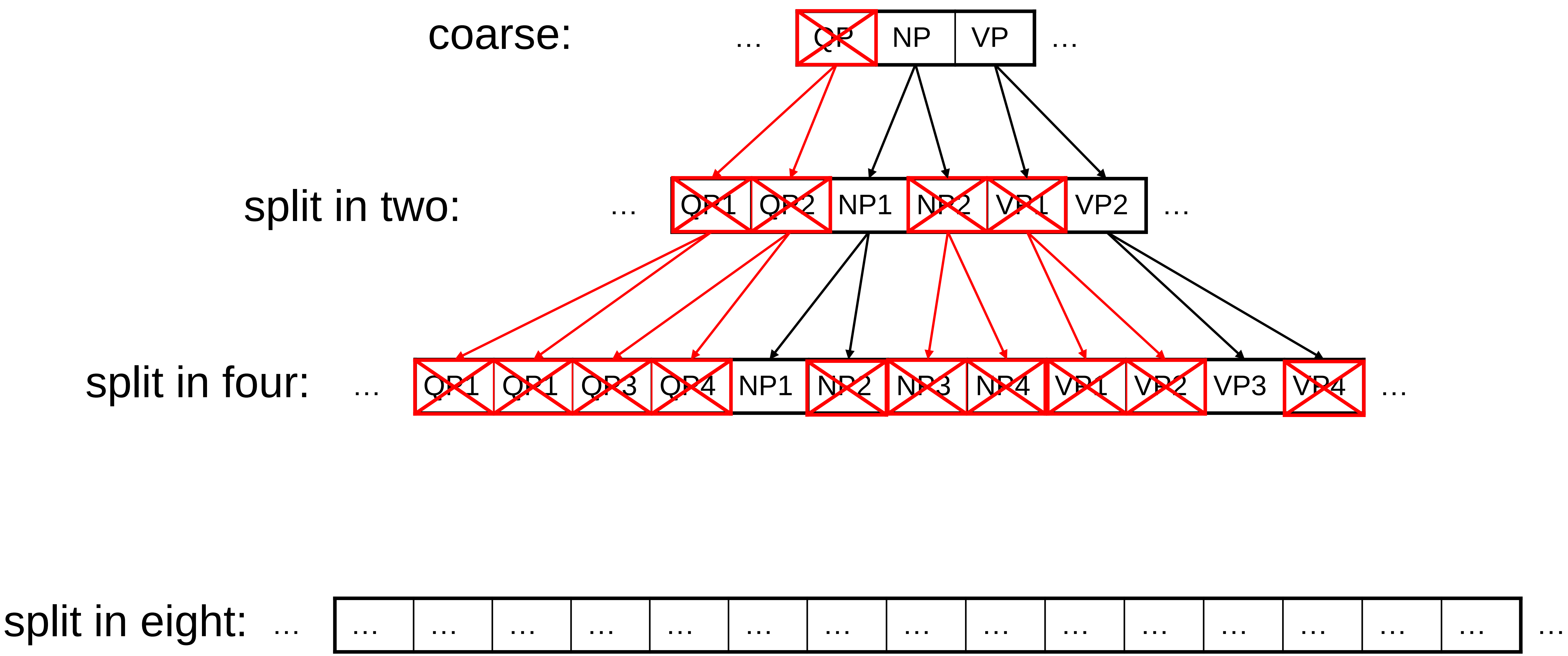
# Incremental Learning



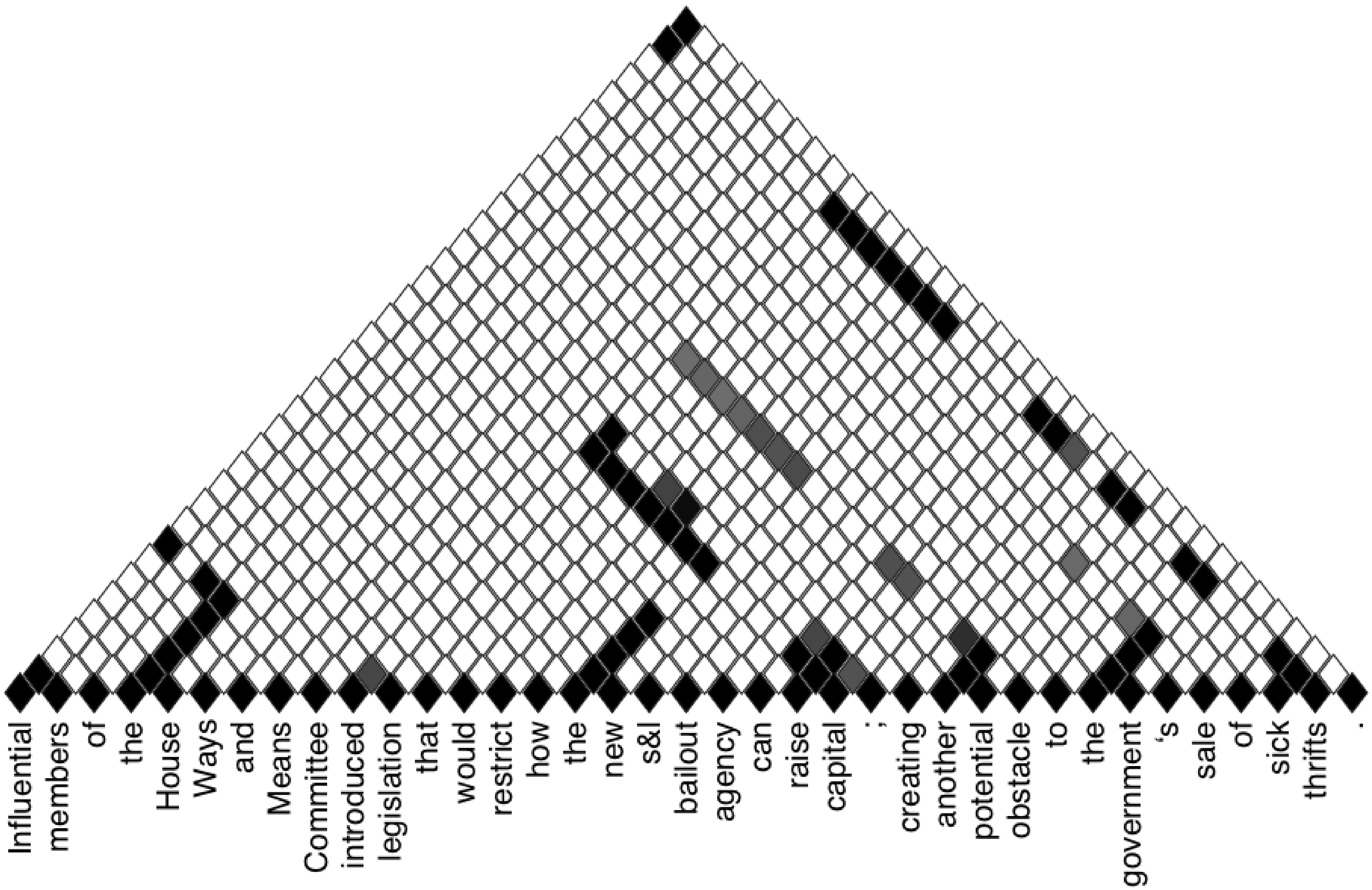
[Charniak 98, Charniak and Johnson 05, Petrov and Klein 07]

# Coarse-to-Fine Pruning

## Consider the span 5 to 12:

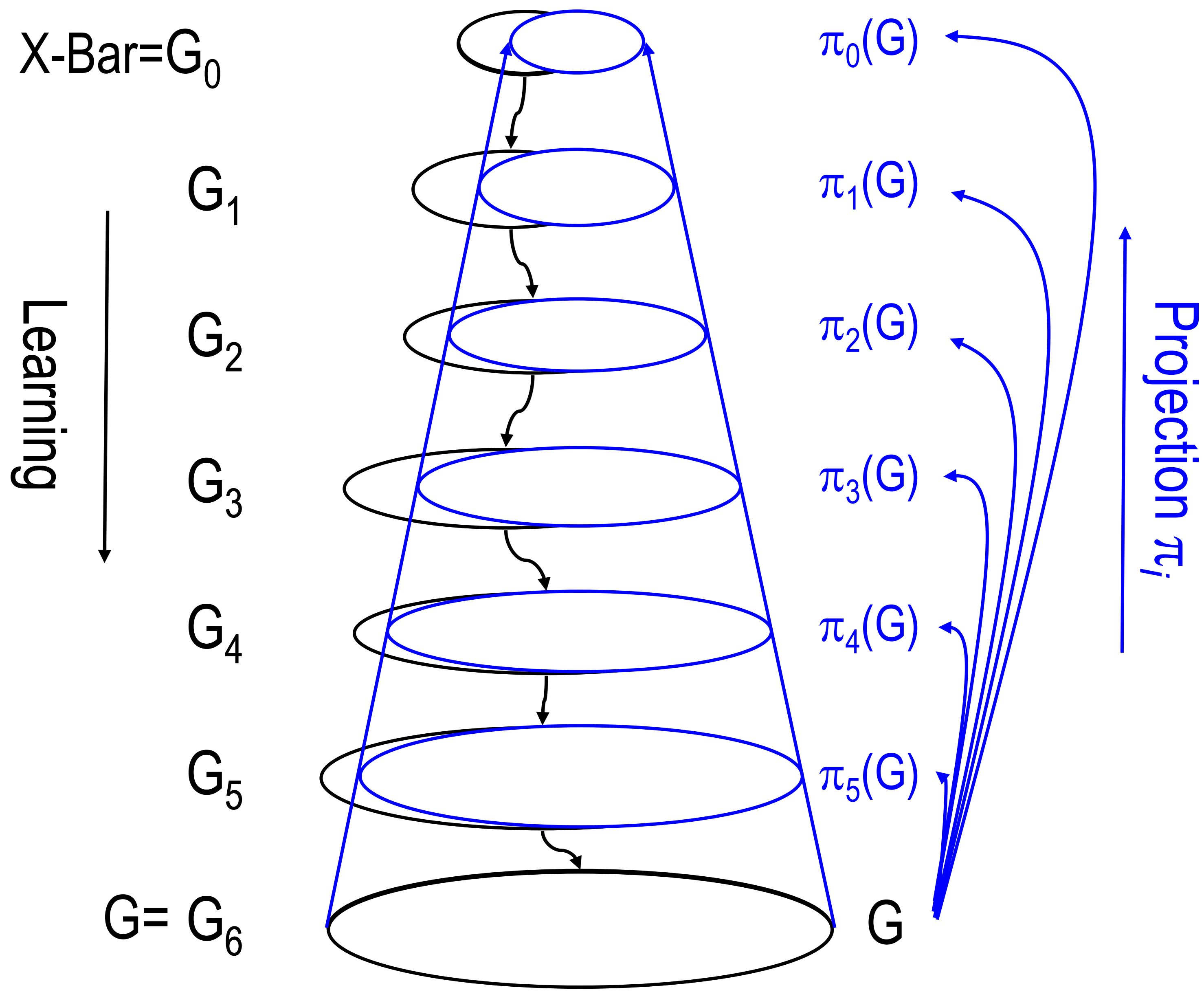


# Bracket Posteriors



[Petrov and Klein 07]

# Projected Grammars

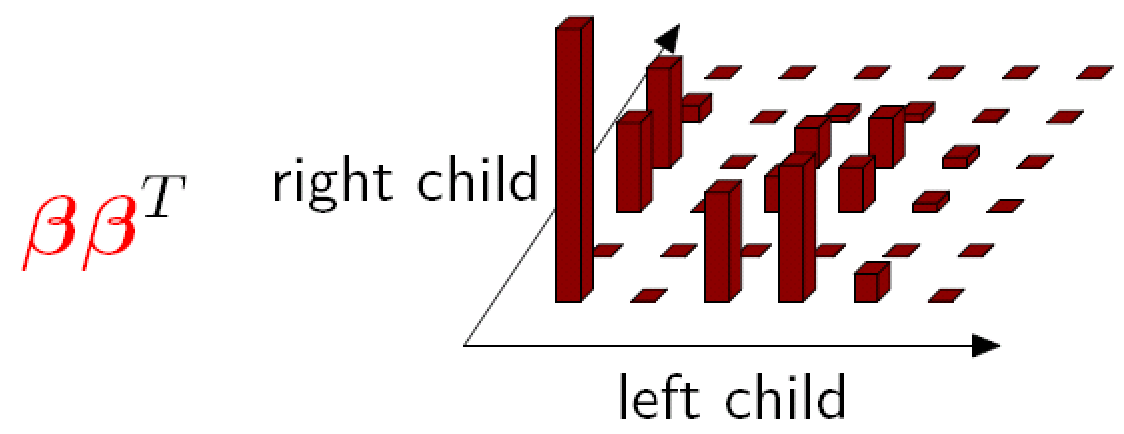


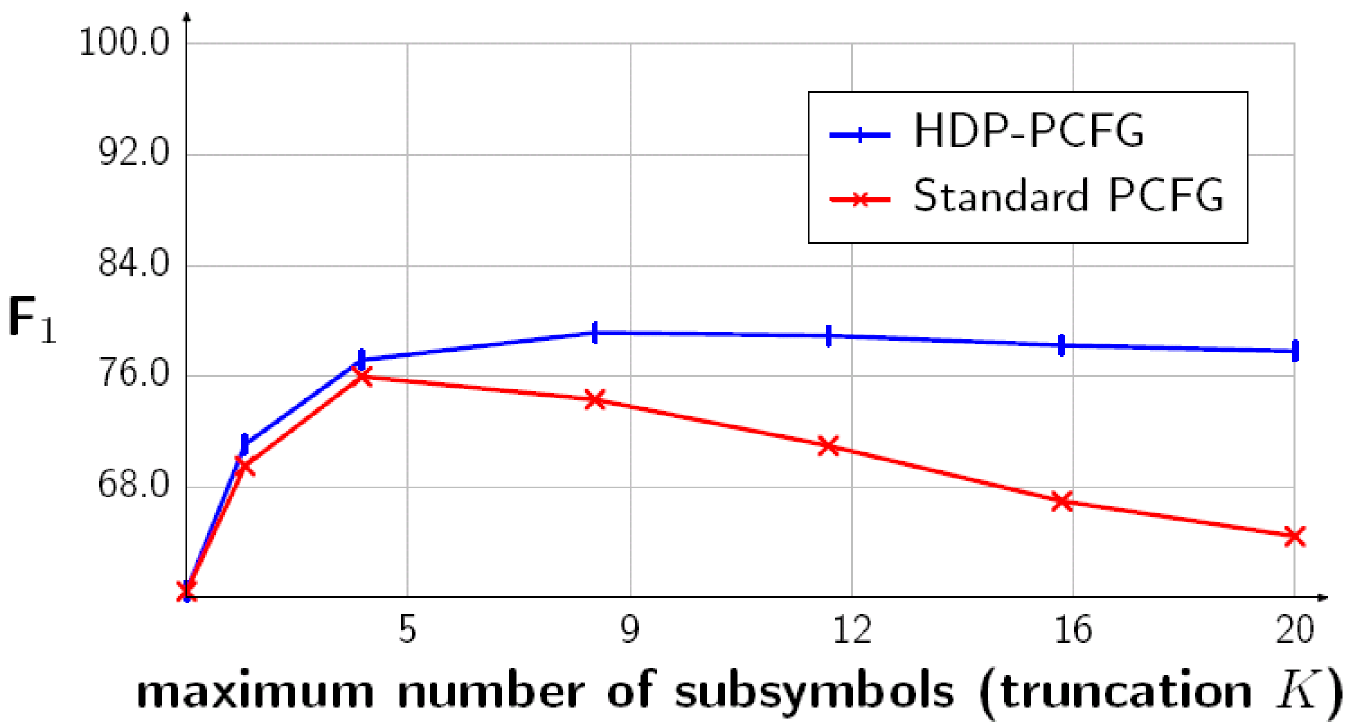
# Final Results (Accuracy)

| | | ≤ 40 words F1 | all F1 |
|---|---|---|---|
| ENG | Charniak&Johnson ‘05 (generative) | 90.1 | 89.6 |
| | **Split / Merge** | **90.6** | **90.1** |
| GER | Dubey ‘05 | 76.3 | - |
| | **Split / Merge** | **80.8** | **80.1** |
| CHN | Chiang et al. ‘02 | 80.0 | 76.6 |
| | **Split / Merge** | **86.3** | **83.4** |

# Nonparametric PCFGs

[Liang, Petrov, Jordan, & Klein '07]

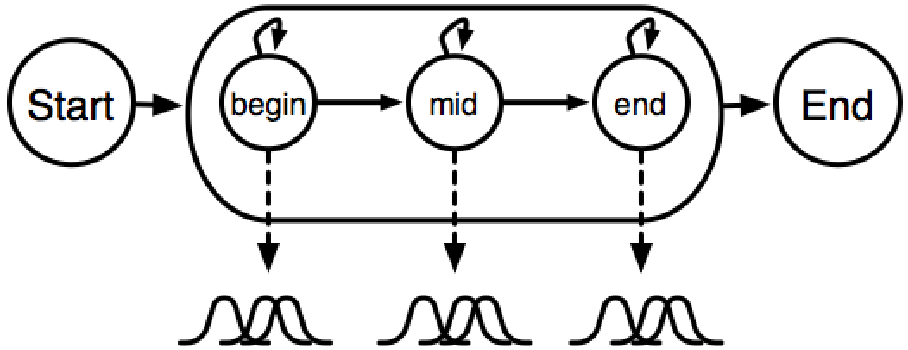


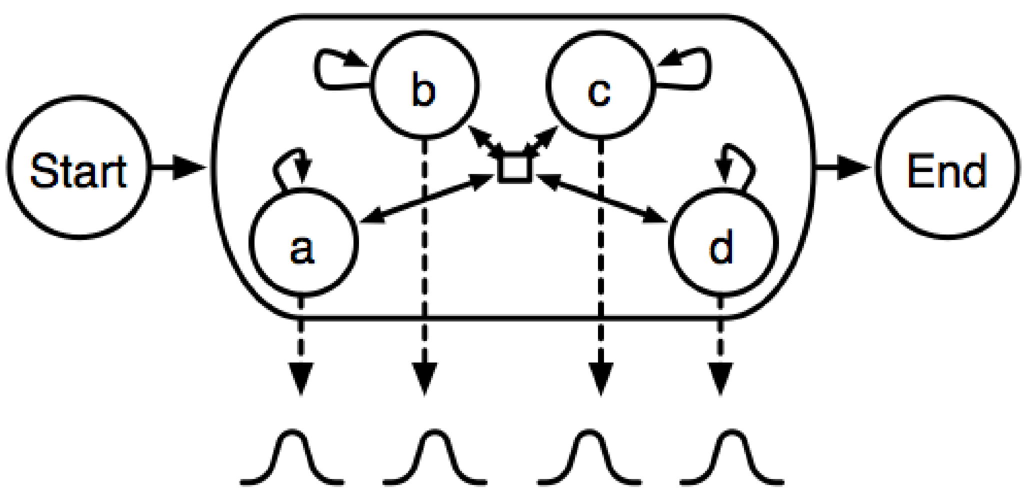
[Petrov, Pauls, & Klein '07]

# Unstructured Phone Models

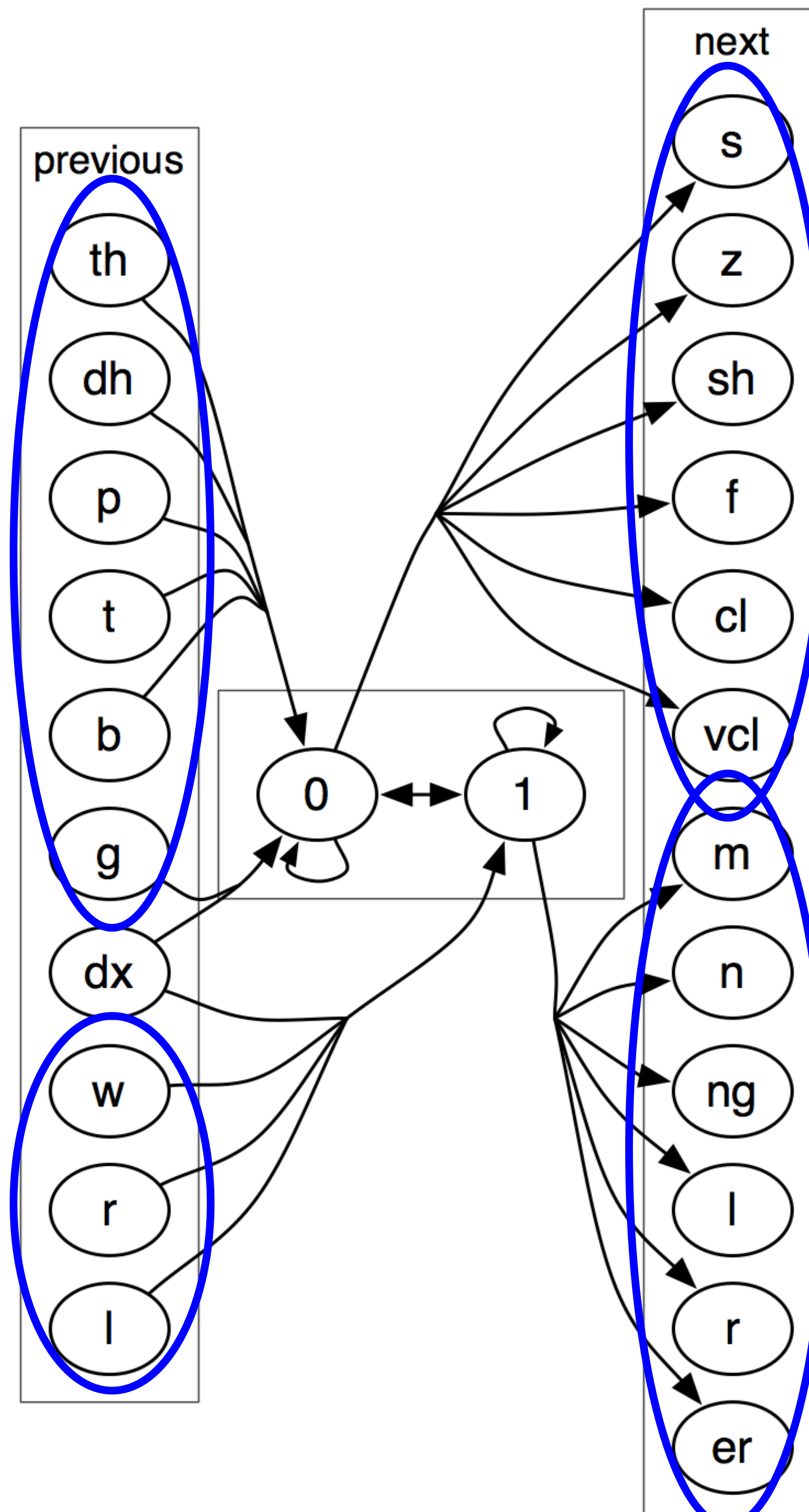
## Standard Model




## Automatic Splits




| | |
|---|---|
| HMM Baseline | 25.1% |
| **5 Split rounds** | **21.4%** |



# Summary

- Latent-variable grammar refinement
  - Automatically learns good grammar splits
  - Gives state-of-the-art parsing accuracy
  - Admits very efficient parsing algorithms
- More applications beyond parsing!

# Outline

- Unsupervised Grammar Refinement
- Unsupervised Coreference Resolution
- Unsupervised Translation Mining

# Unsupervised Coreference

Weir Group whose headquarters

U.S

corporation power plant , which

Jiangsu

[Li et al 04, Haghighi and Klein 07]

# Generative Mention Models

*Weir Group* *Weir Group* *Weir HQ*

… "Weir group" . "whose" … "headquarters" …

---

*United States* *Weir Group*

………… "U.S" ……… "corporation" ………

---

*Weir Plant* *Weir Plant* *Jiangsu*

. "power plant" . "which" ……. "Jiangsu" …….

# Generative Mention Models

*Weir Group* → … “Weir group” … *Weir Group* → “whose” … *Weir HQ* → “headquarters” …

---

*United States* → ……………… “U.S …………… *Weir Group* → corporation” …………

**Inference Time**

---

*Weir Plant* → · “power plant” · *Weir Plant* → “which” ……… *Jiangsu* → “Jiangsu” ………

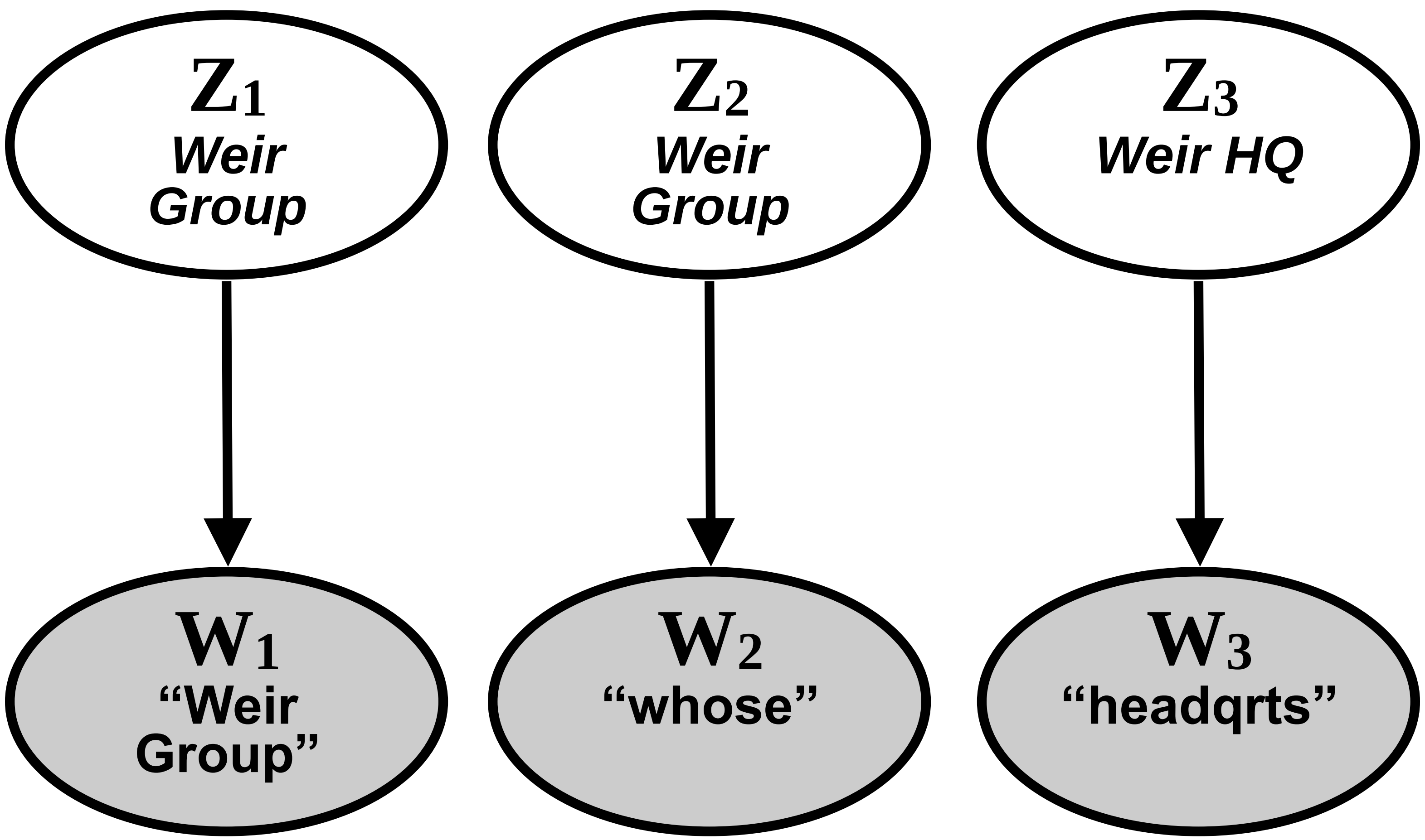
# Finite Mixture Model

Entity Distribution

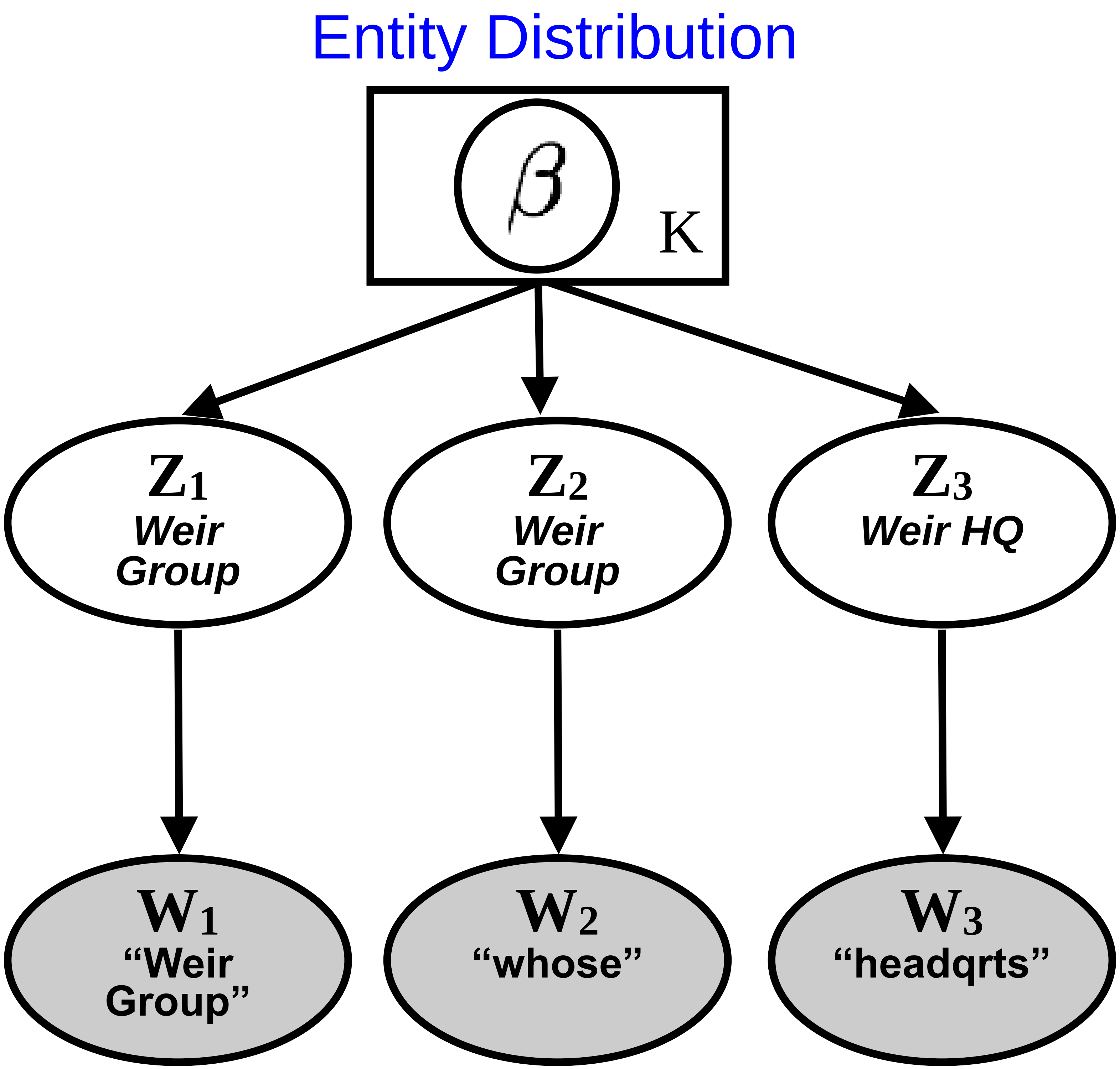
P(*Weir Group*) = 0.2,
P(*Weir HQ*) = 0.5,
.....

Mention Parameters

P(W | *Weir Group*):
"Weir Group"=0.4,
"whose"=0.2,
.......



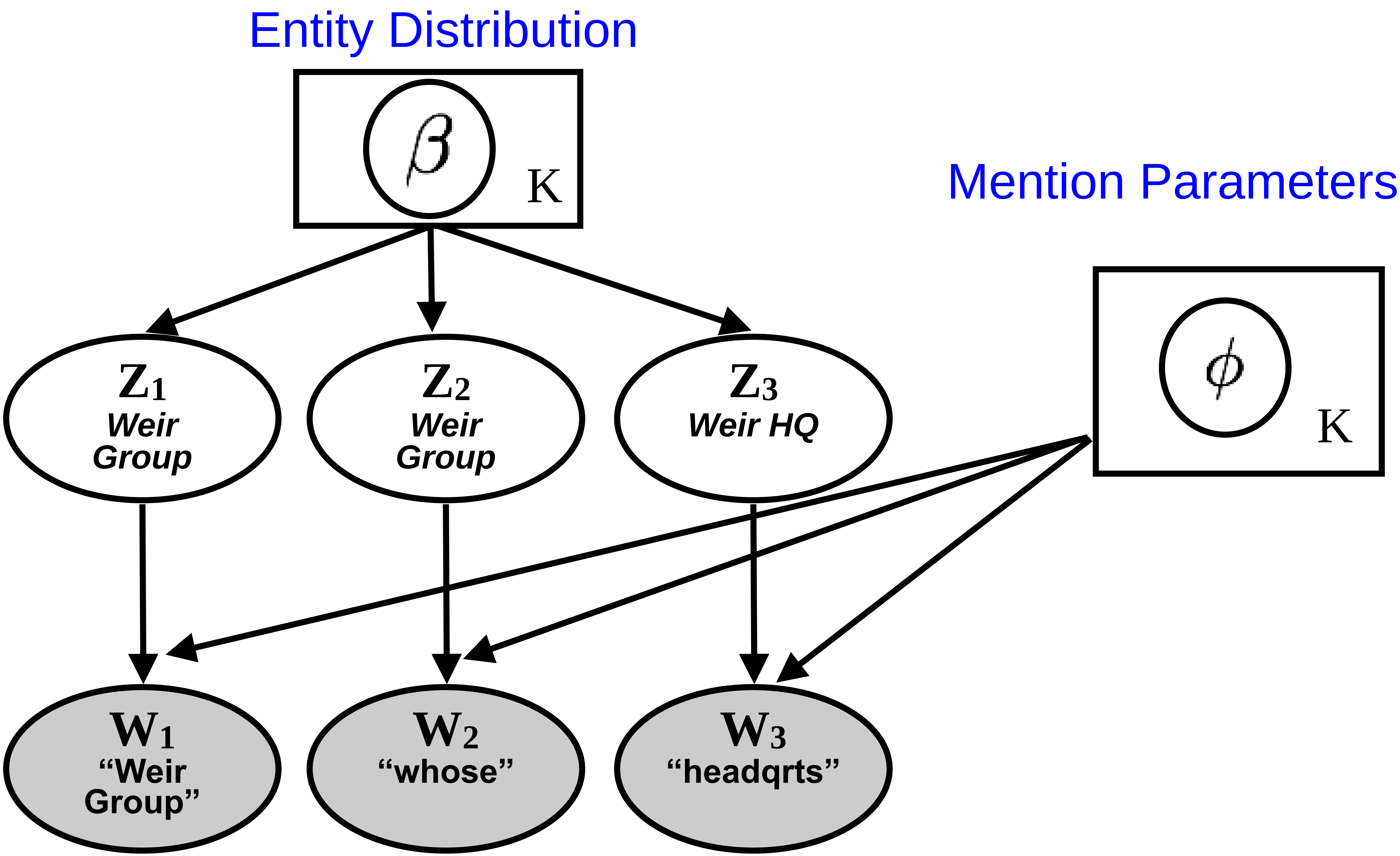
# Finite Mixture Model



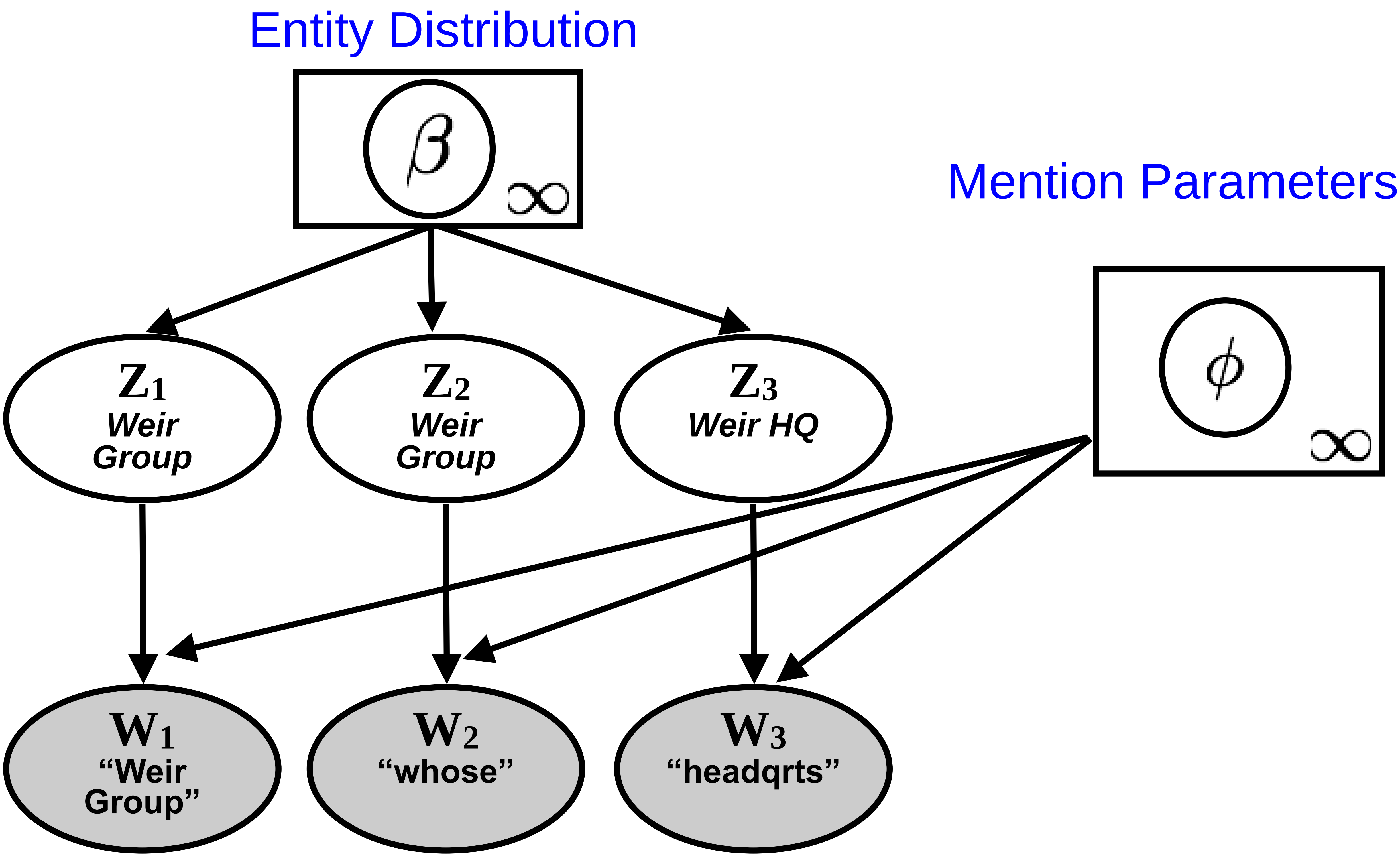

Mention Parameters

P(W | *Weir Group*):
"Weir Group"=0.4,
"whose"=0.2,
.......

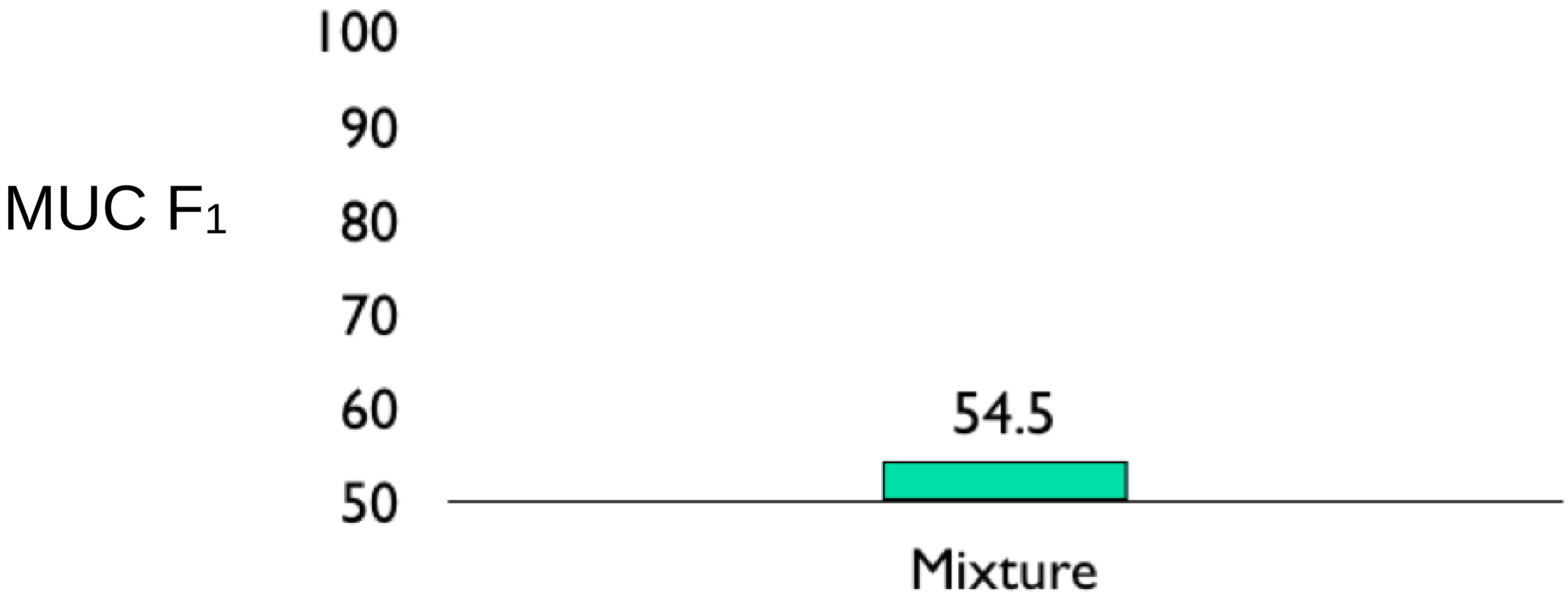
# Finite Mixture Model



# Infinite Mixture Model



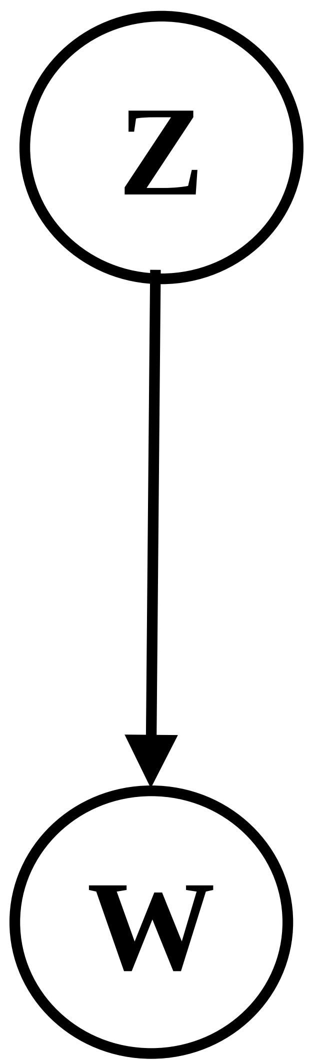
# Infinite Mixture Model




The Weir Group , whose headquarters is in the U.S is a large specialized corporation. This power plant , which , will be situated in Jiangsu, has a large generation capacity.

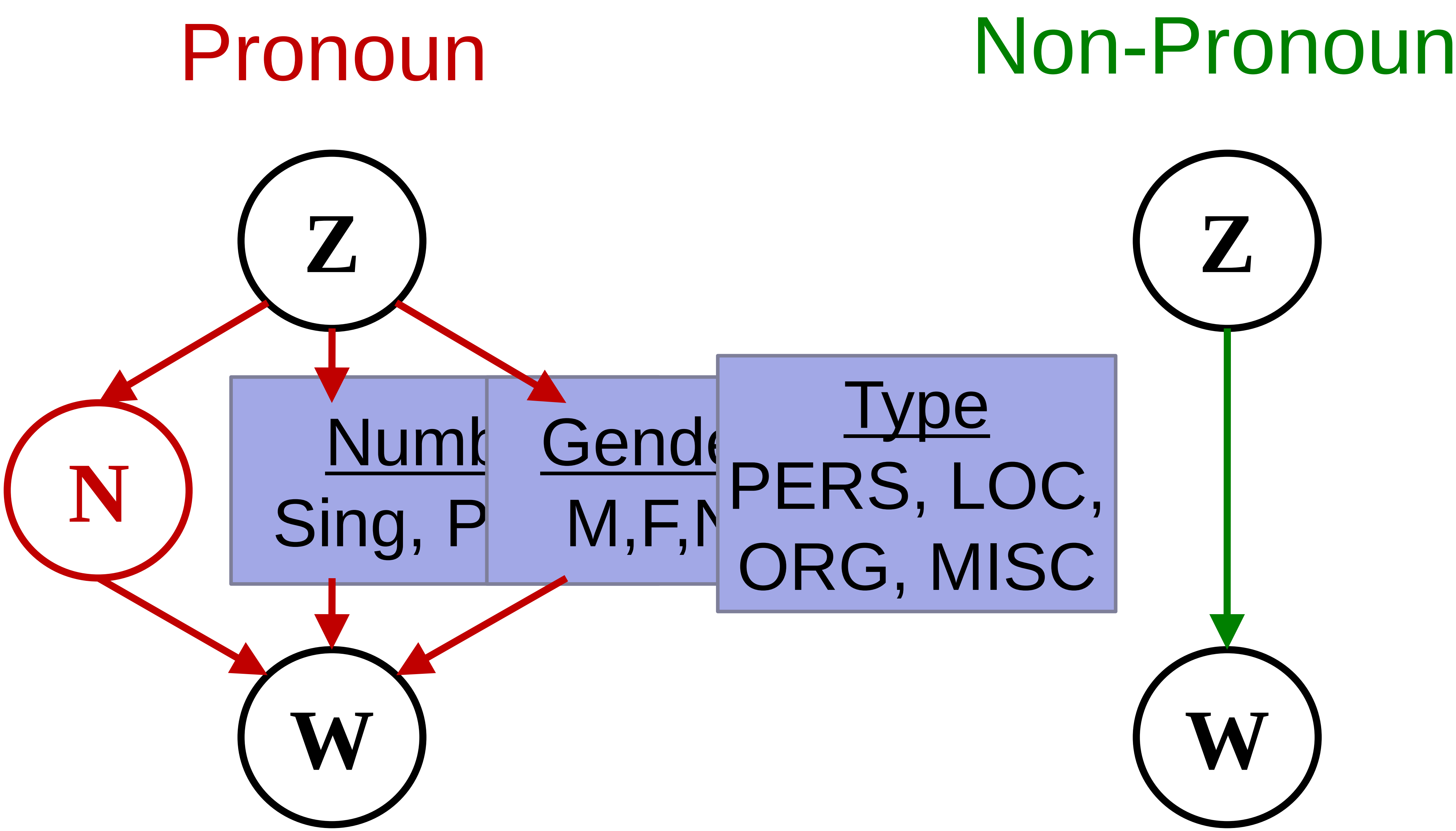
# Enriching the Mention Model

Mention Model



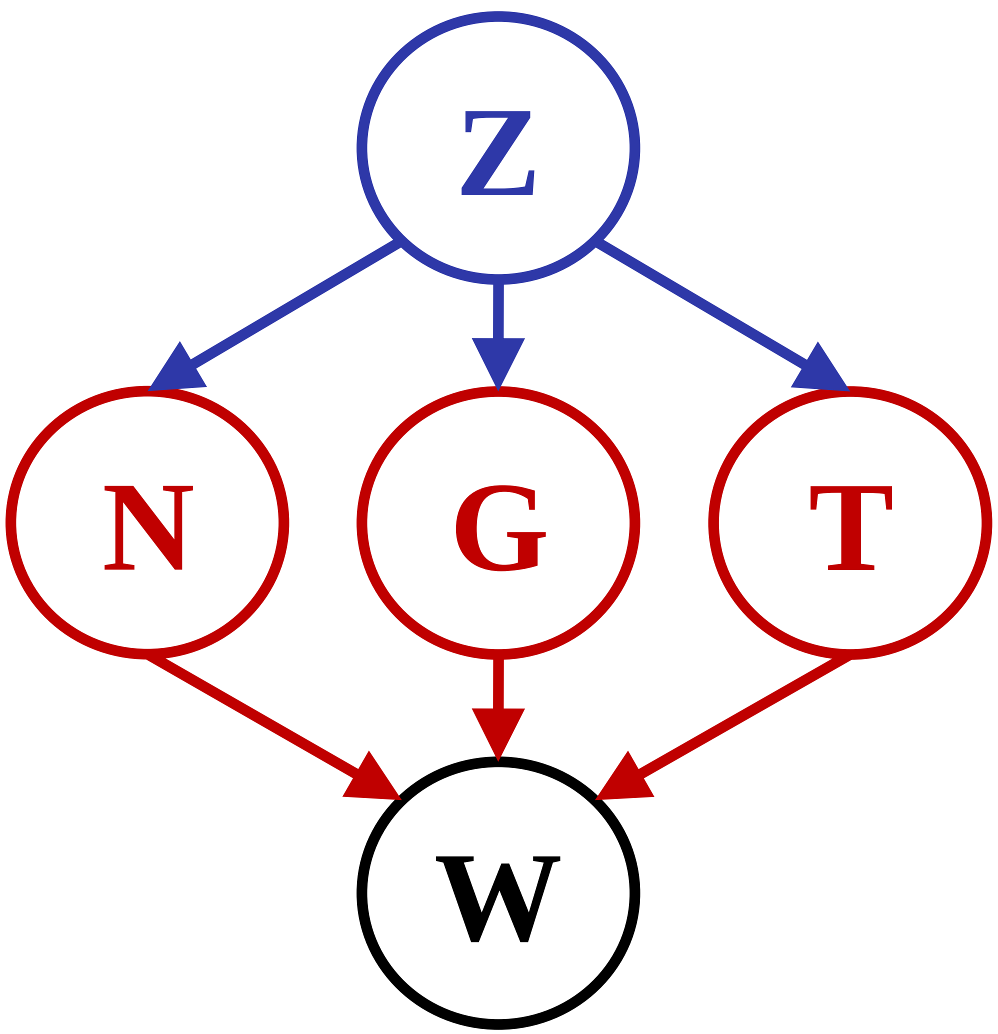

P(W | *Weir Group*):
“Weir Group”=0.4,
“whose”=0.2,
…….

# Enriching the Mention Model

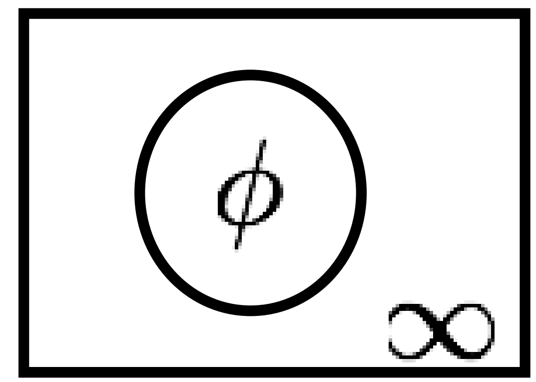


# Enriching the Mention Model

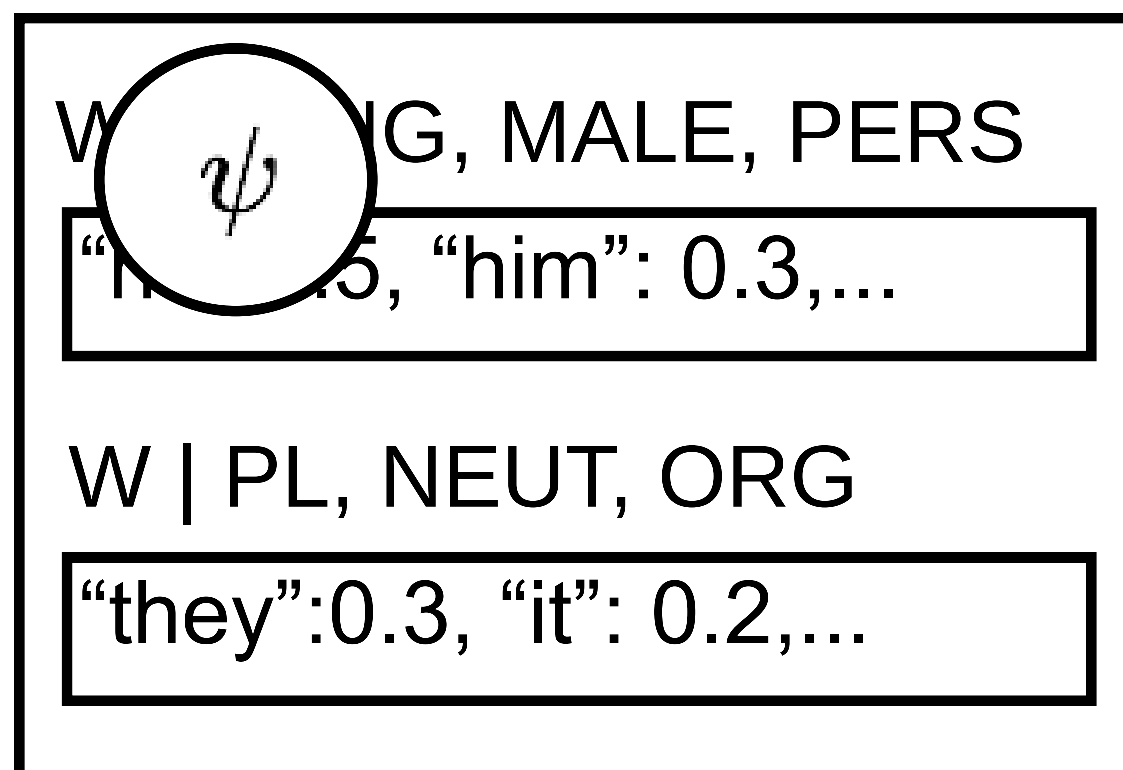
Pronoun




Entity Parameters

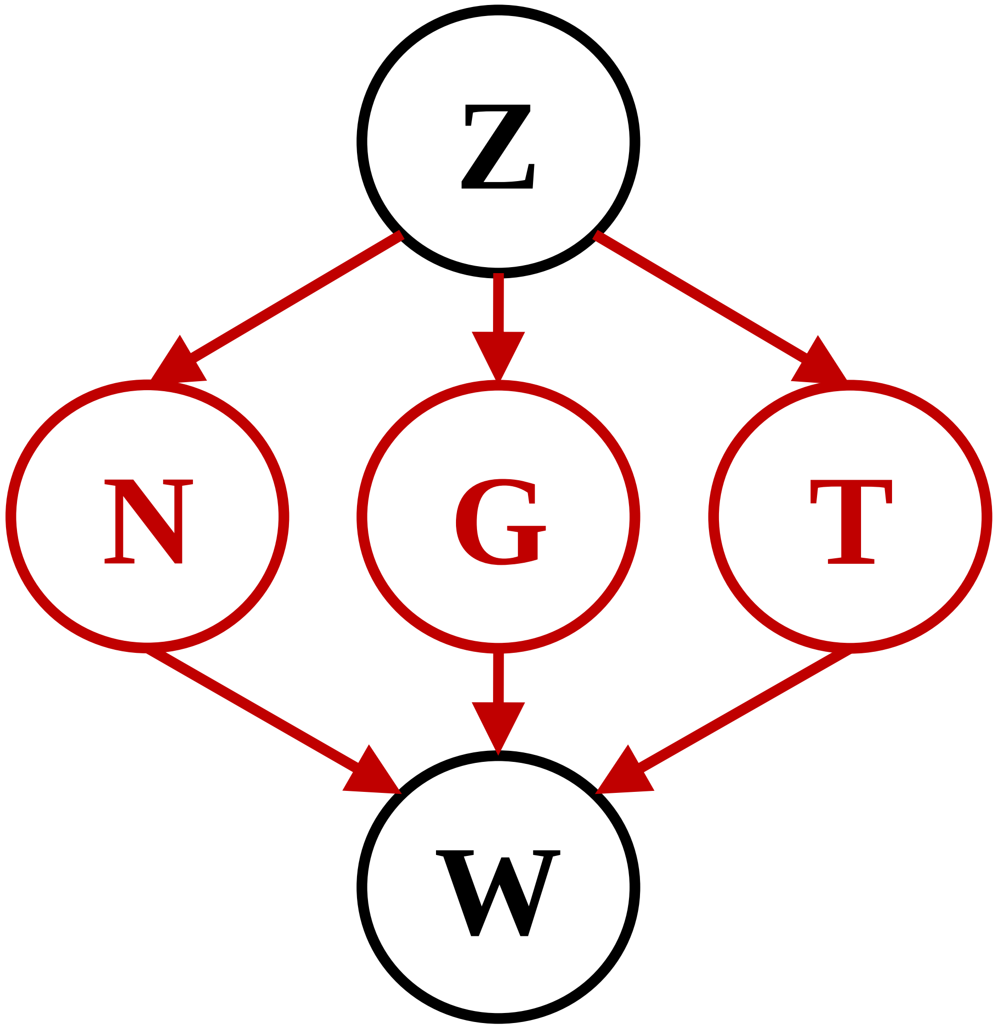



Pronoun Parameters

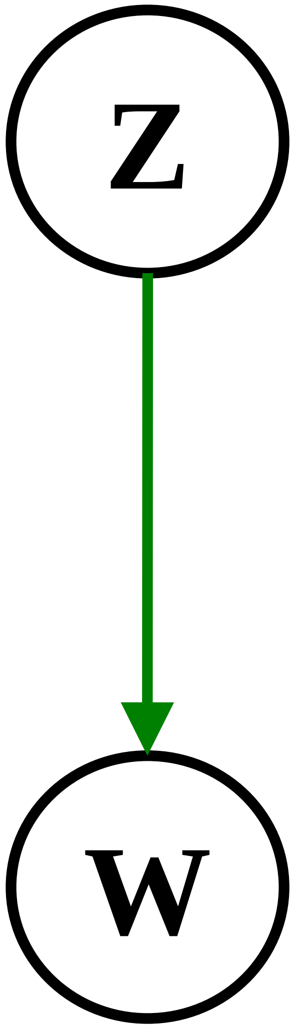


# Enriching the Mention Model

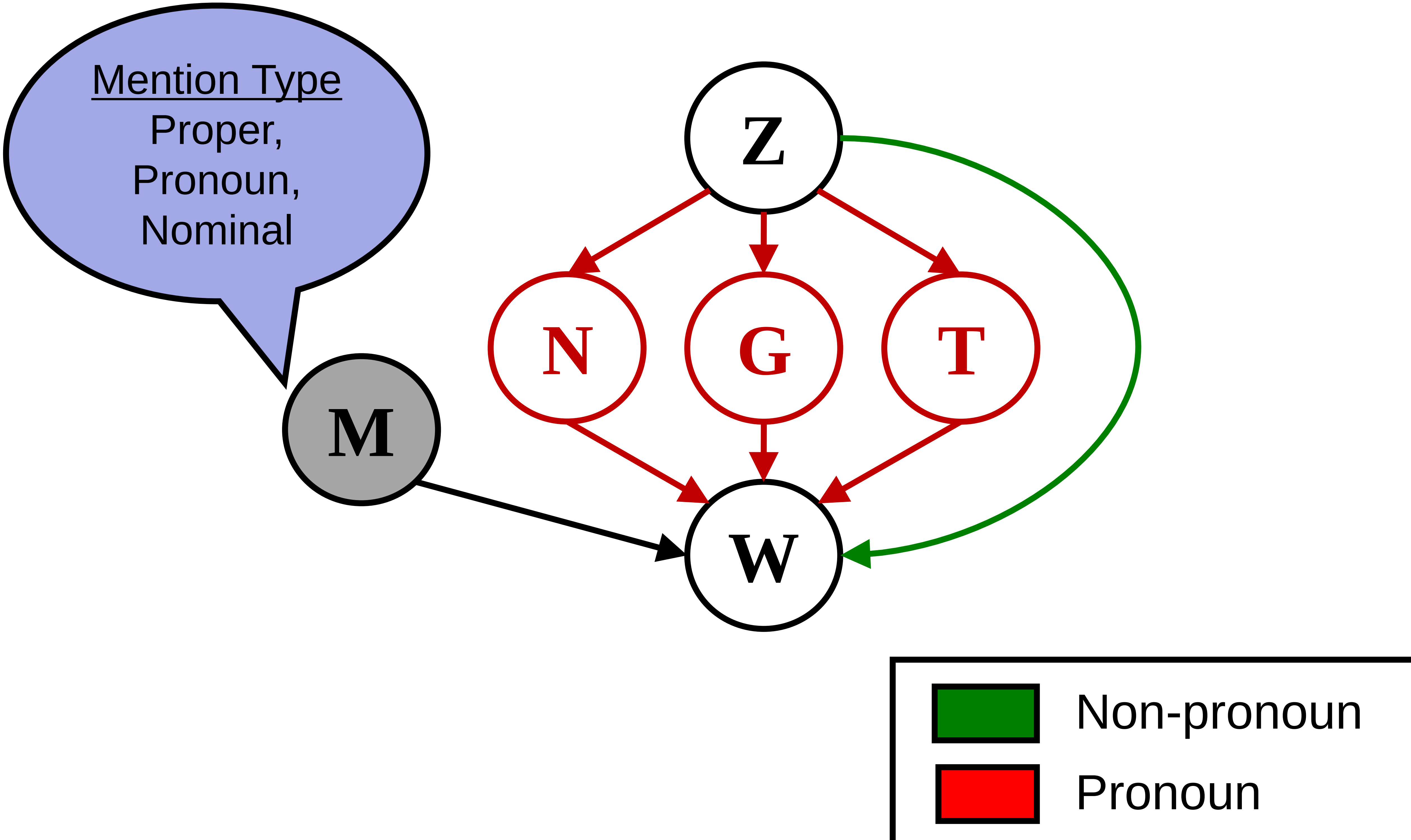
Pronoun



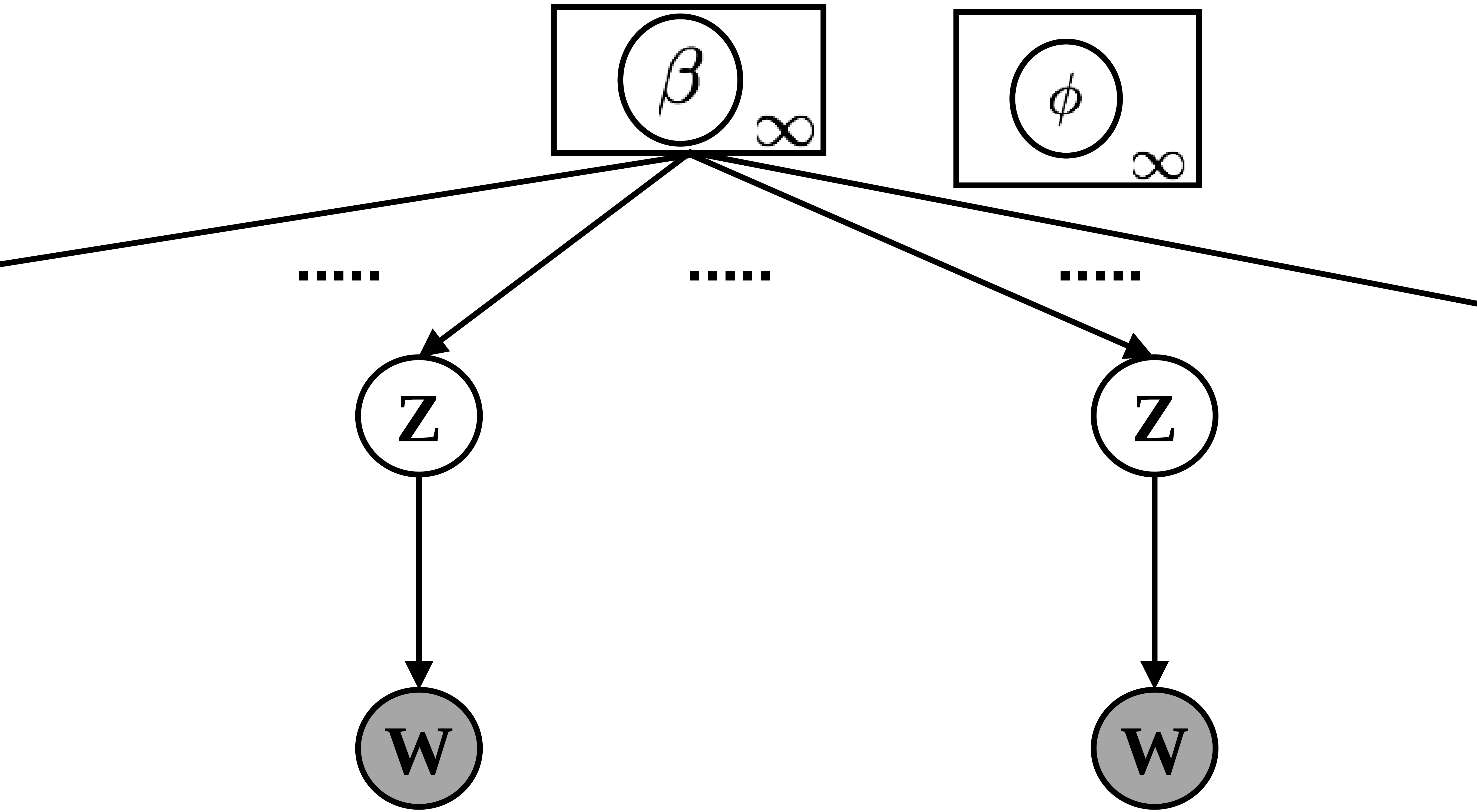

Non-Pronoun



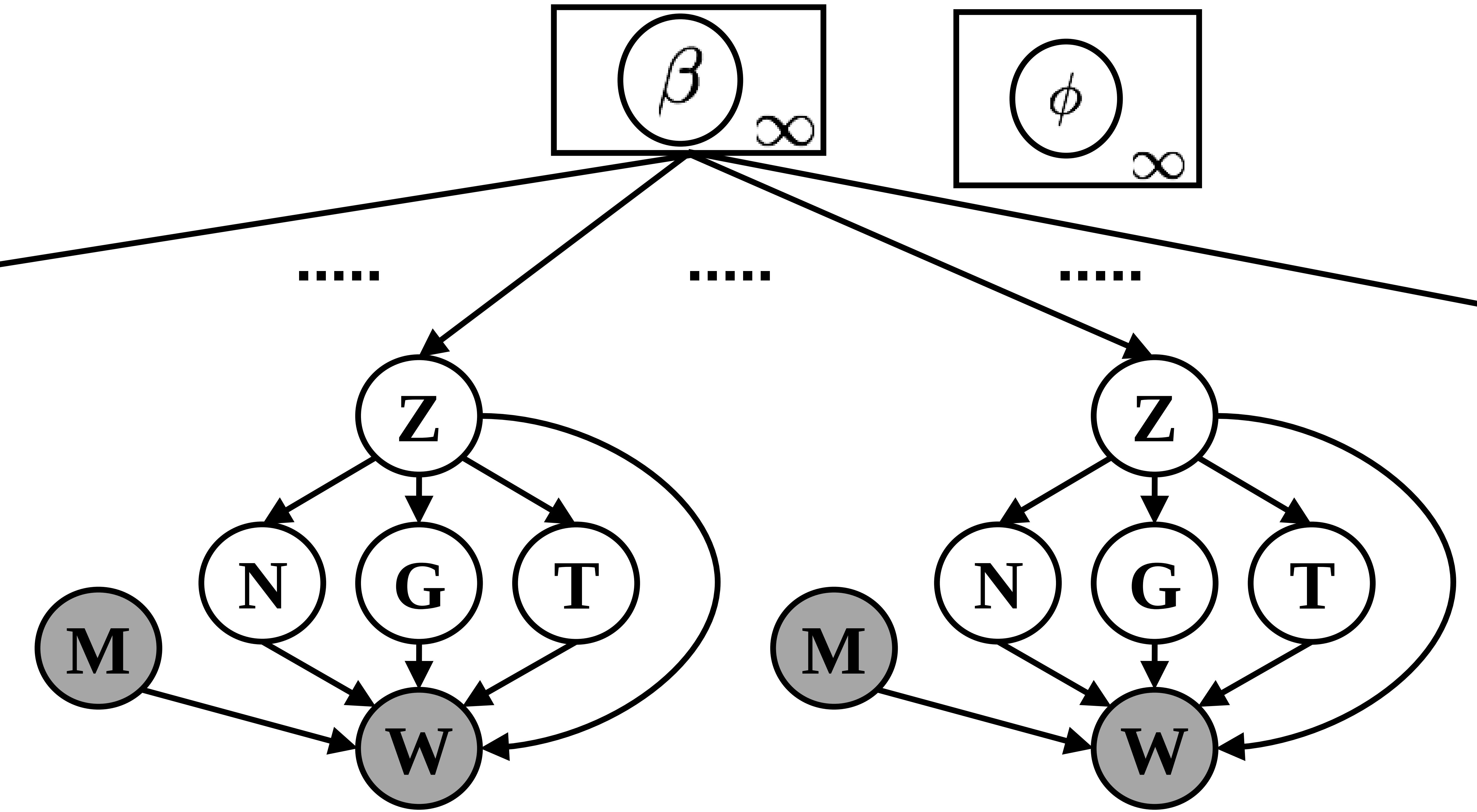
# Enriching Mention Model



# Enriching Mention Model

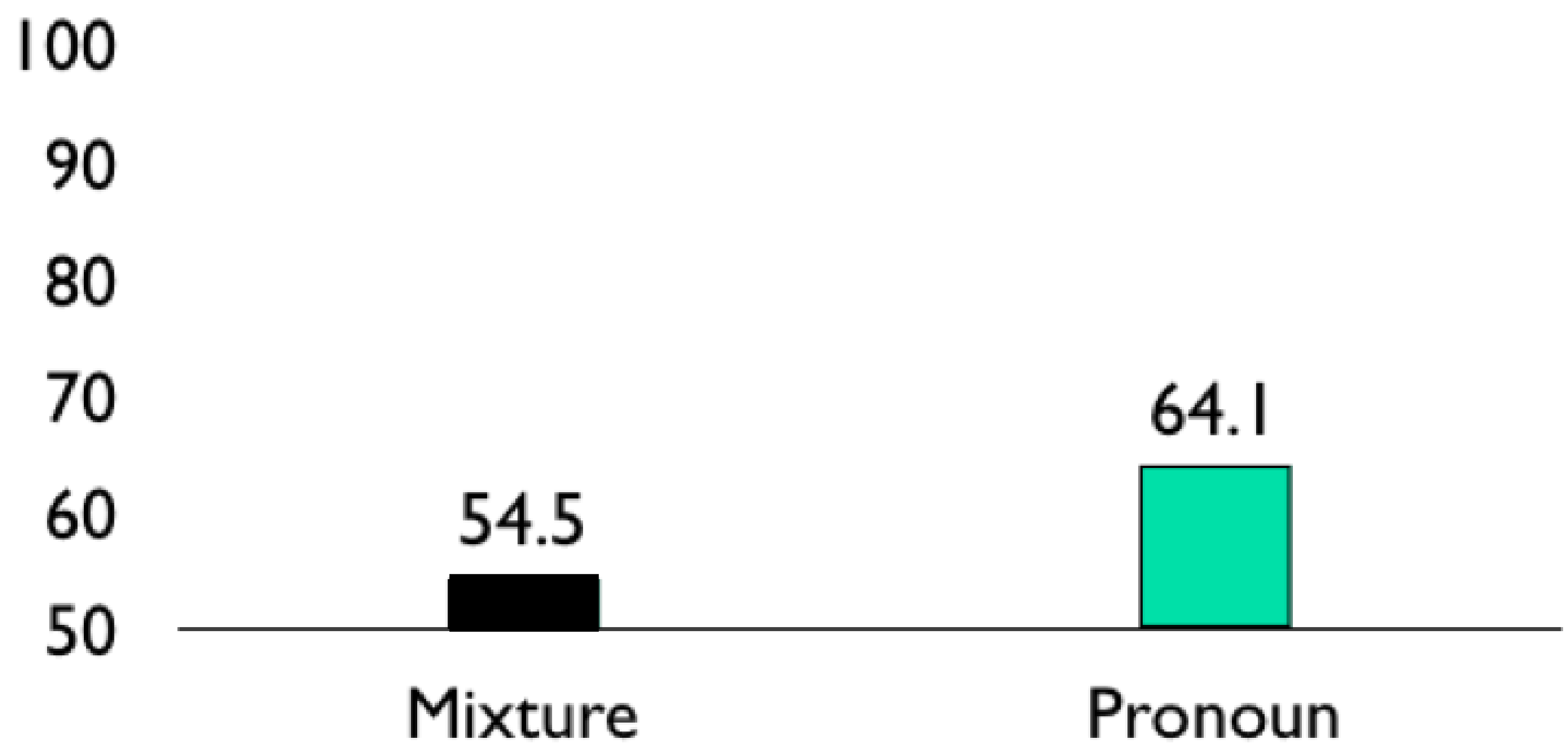


# Enriching Mention Model



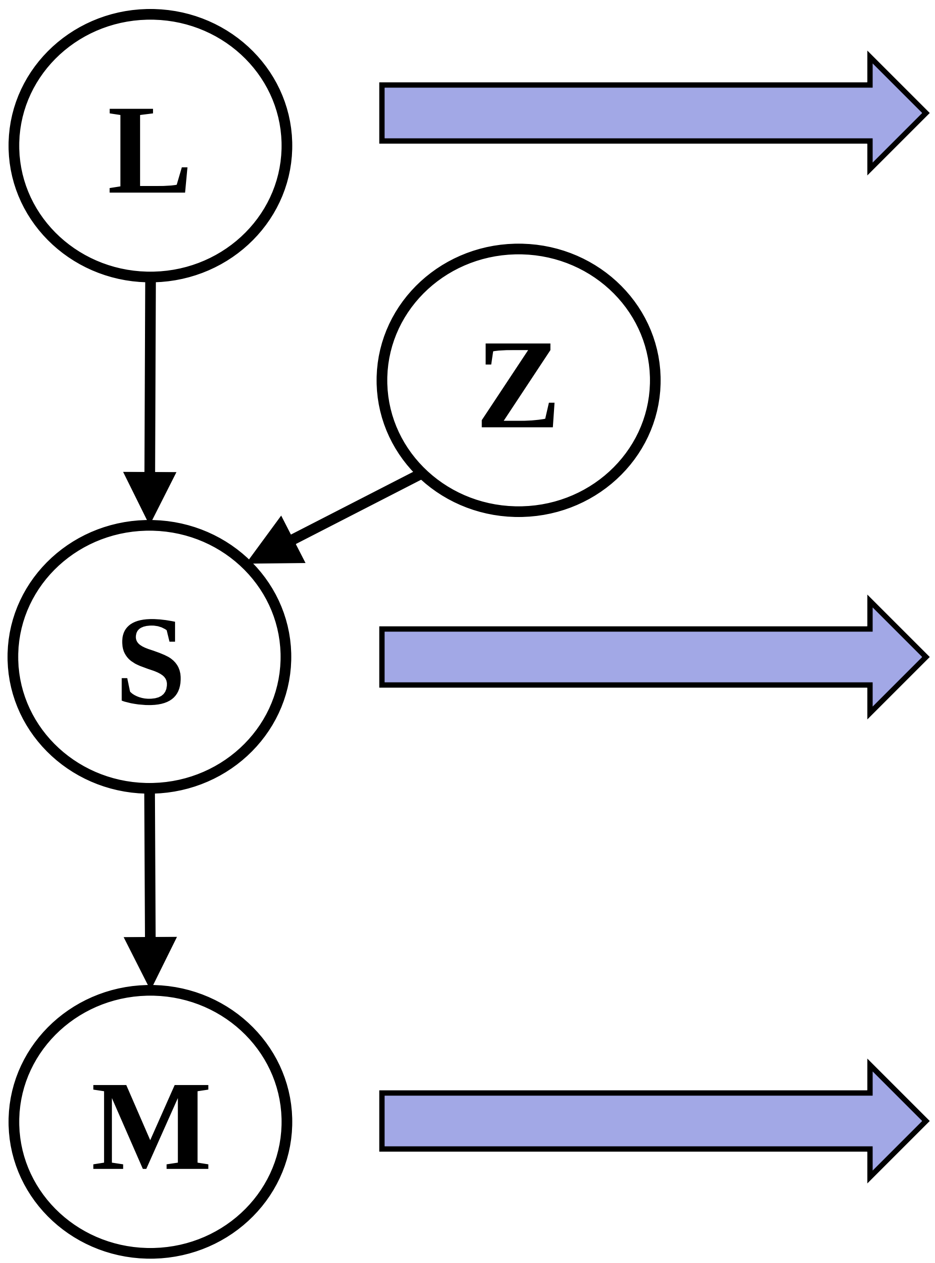
# Pronoun Model

MUC $F_1$




The Weir Group , whose headquarters is in the U.S is a large specialized corporation. This power plant , which , will be situated in Jiangsu, has a large generation capacity.

# Salience Model




| Entity | Activation |
|---|---|
| 1 | 1.0 |
| 2 | 0.0 |

| Salience Values |
|---|
| TOP, HIGH, MED, LOW, NONE |

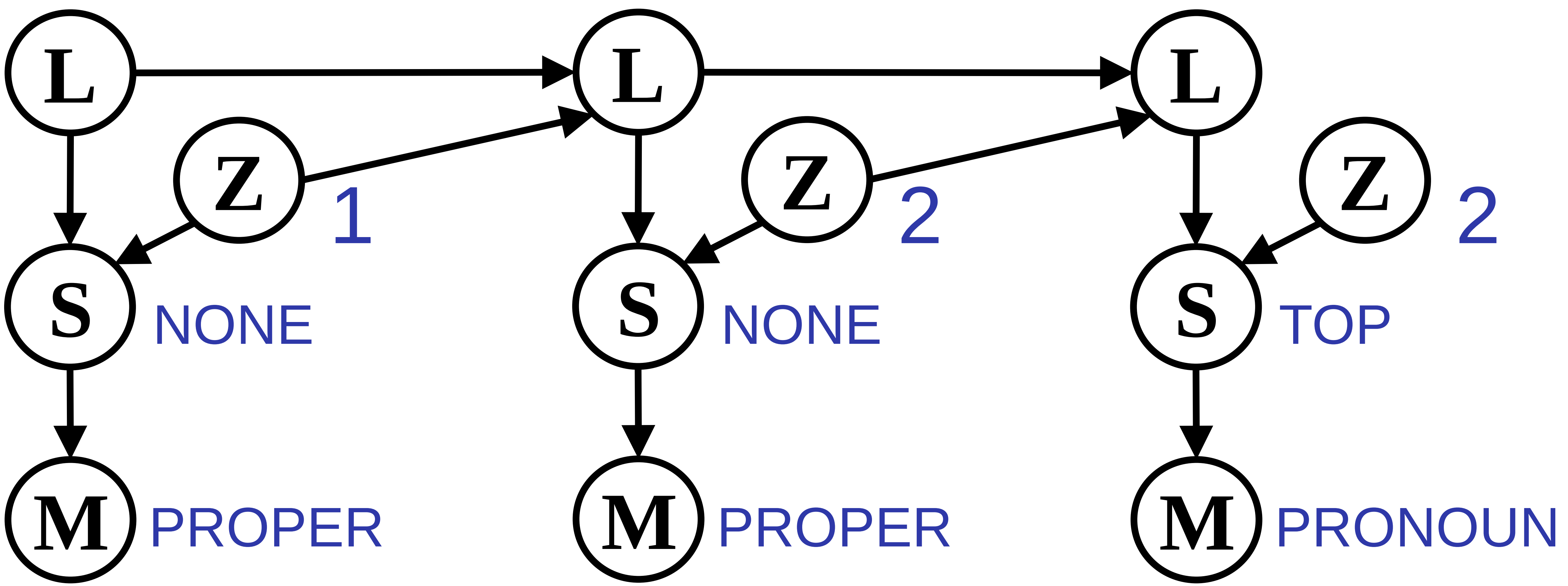
| Mention Type |
|---|
| Proper, Pronoun, Nominal |

# Salience Model

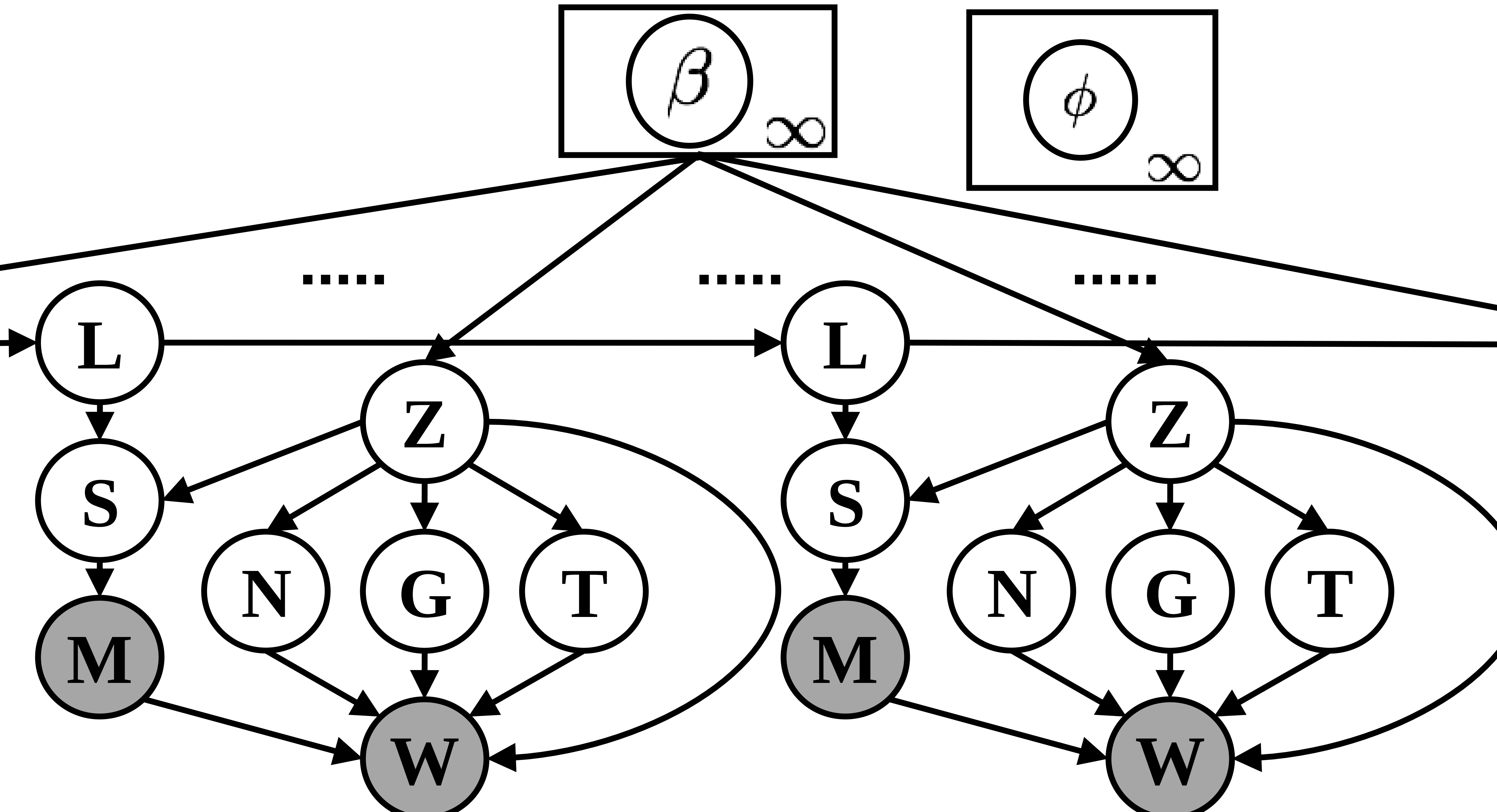
| Entity | Activation |
|---|---|
| 1 | 0.0 |
| 2 | 0.0 |

| Entity | Activation |
|---|---|
| 1 | 1.0 |
| 2 | 0.0 |

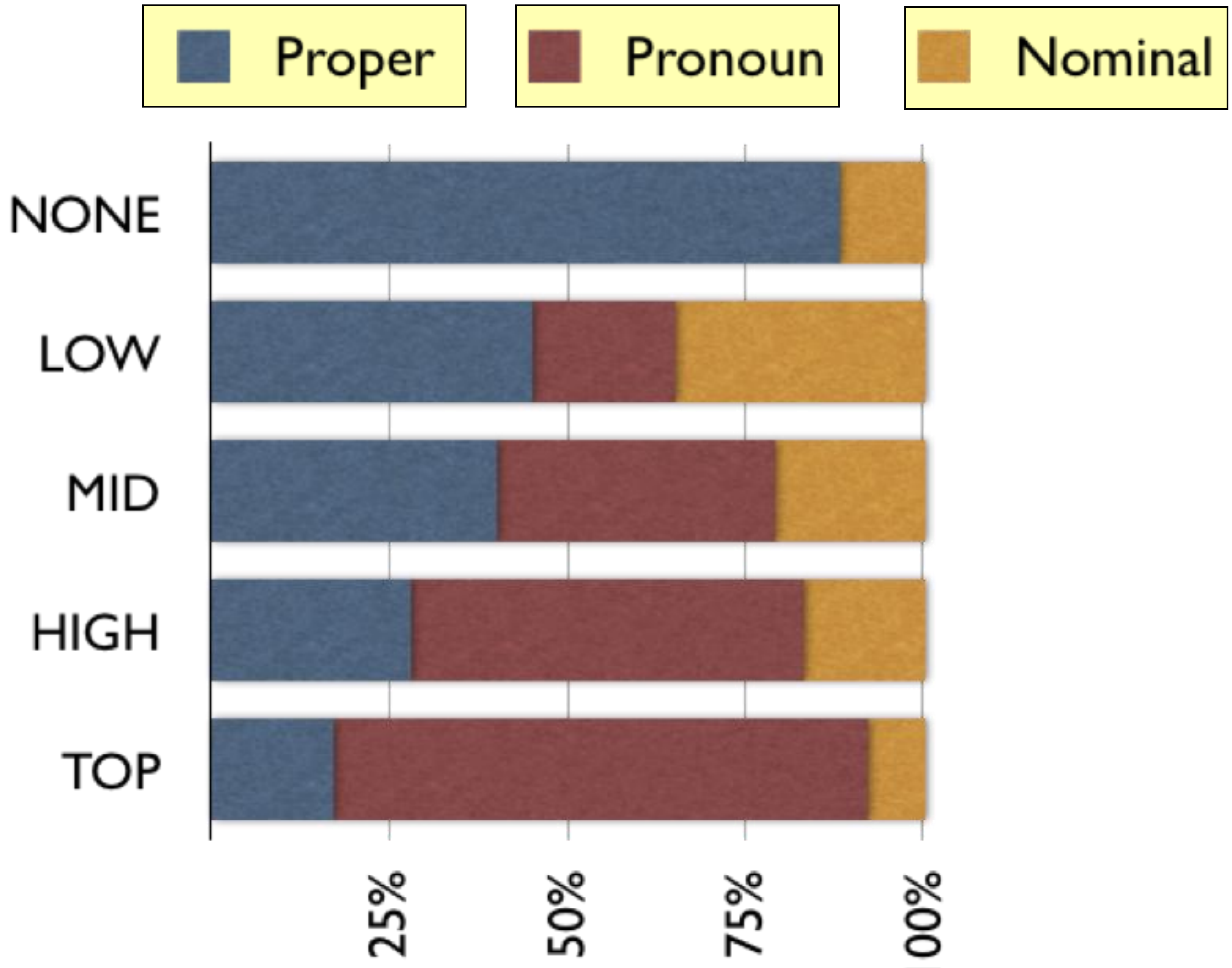
| Entity | Activation |
|---|---|
| 1 | 0.5 |
| 2 | 1.0 |



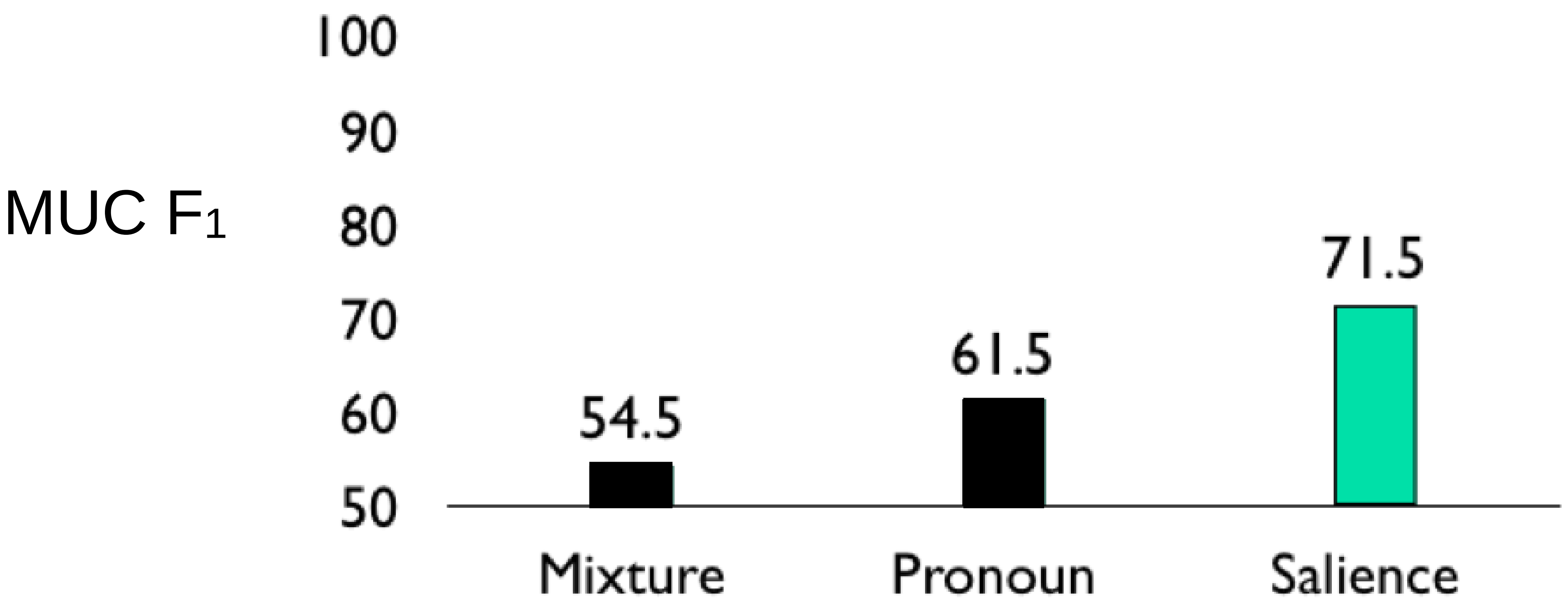
# Salience Model



# Salience Model

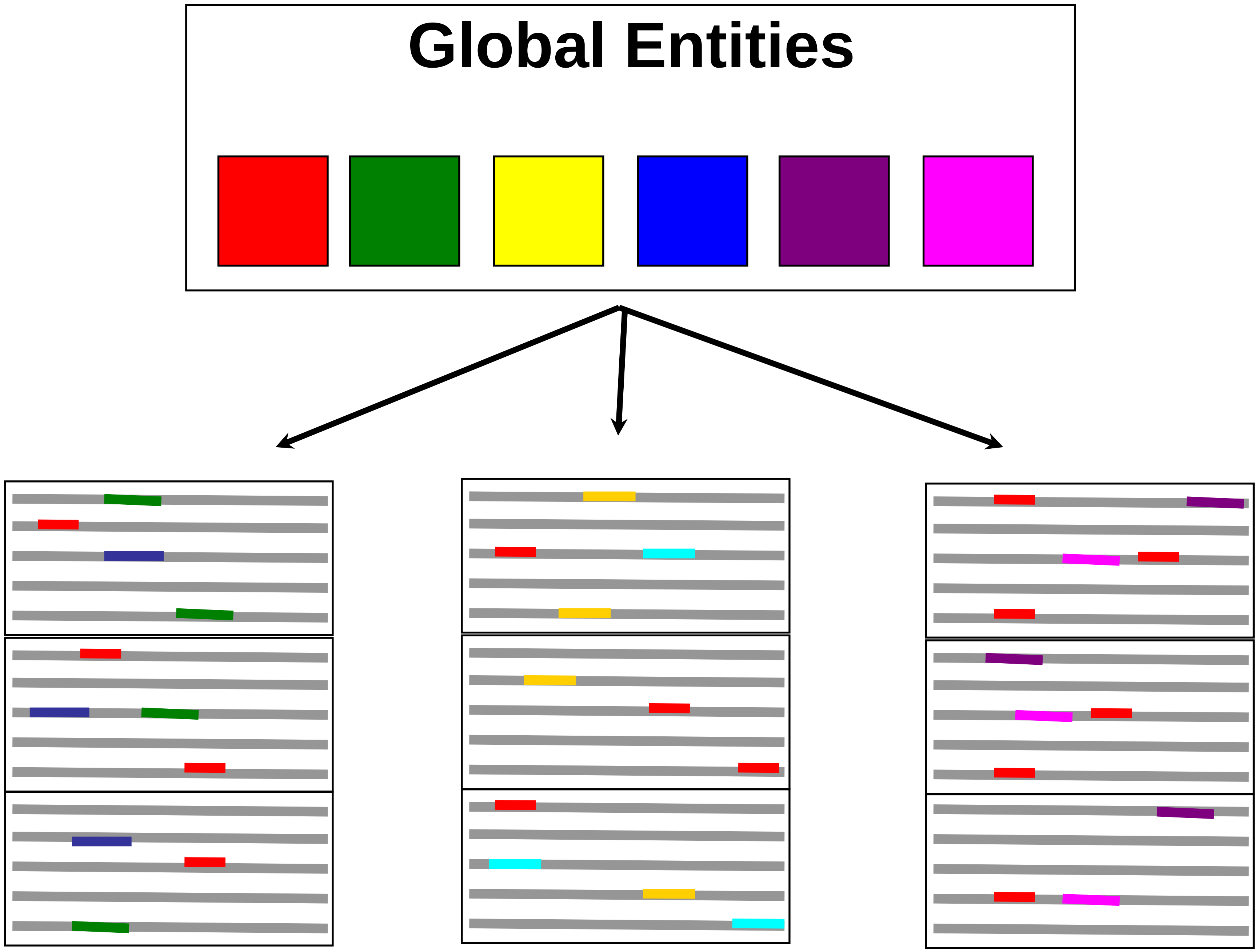


# Salience Model

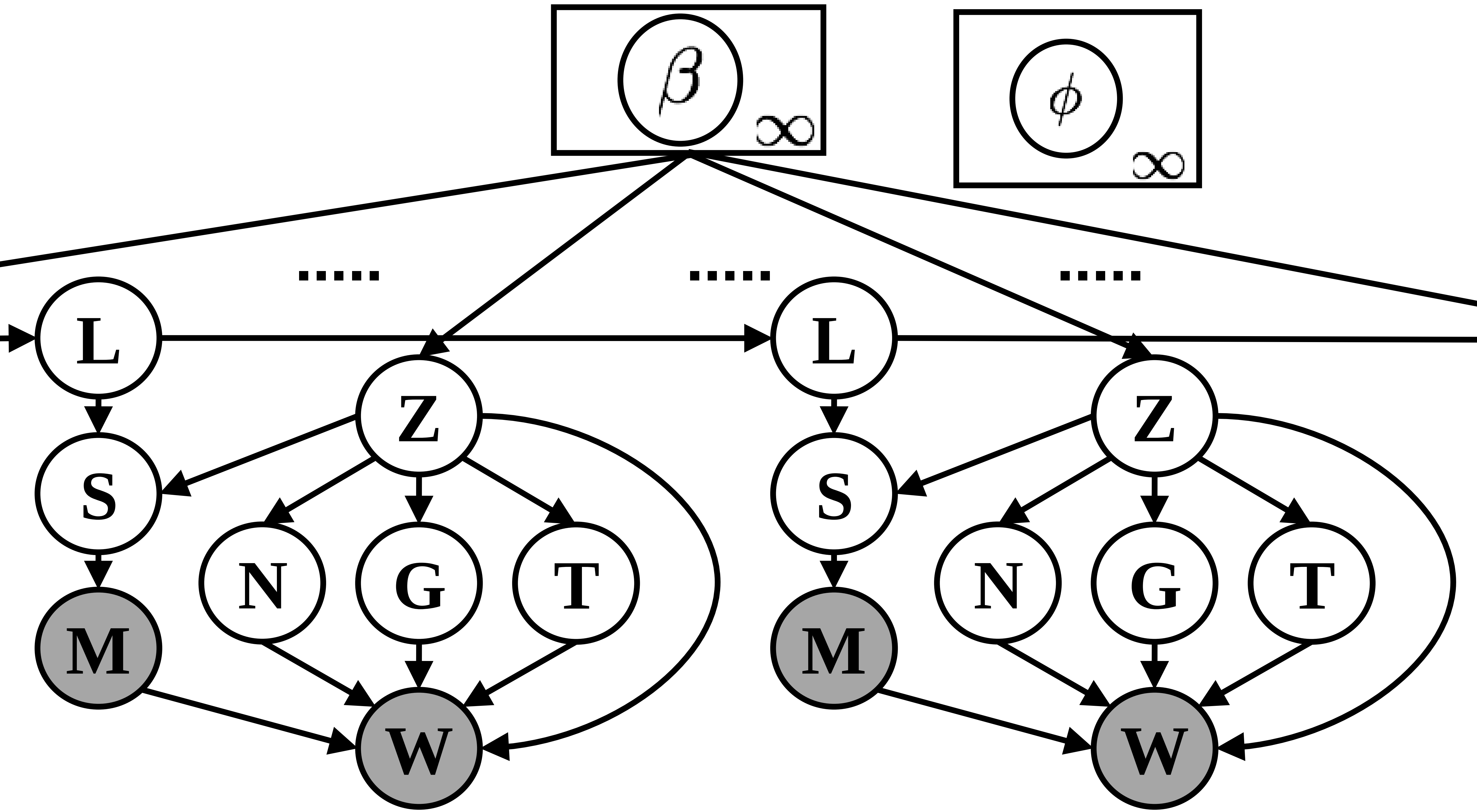



The Weir Group , whose headquarters is in the U.S is a large specialized corporation. This power plant , which , will be situated in Jiangsu, has a large generation capacity.

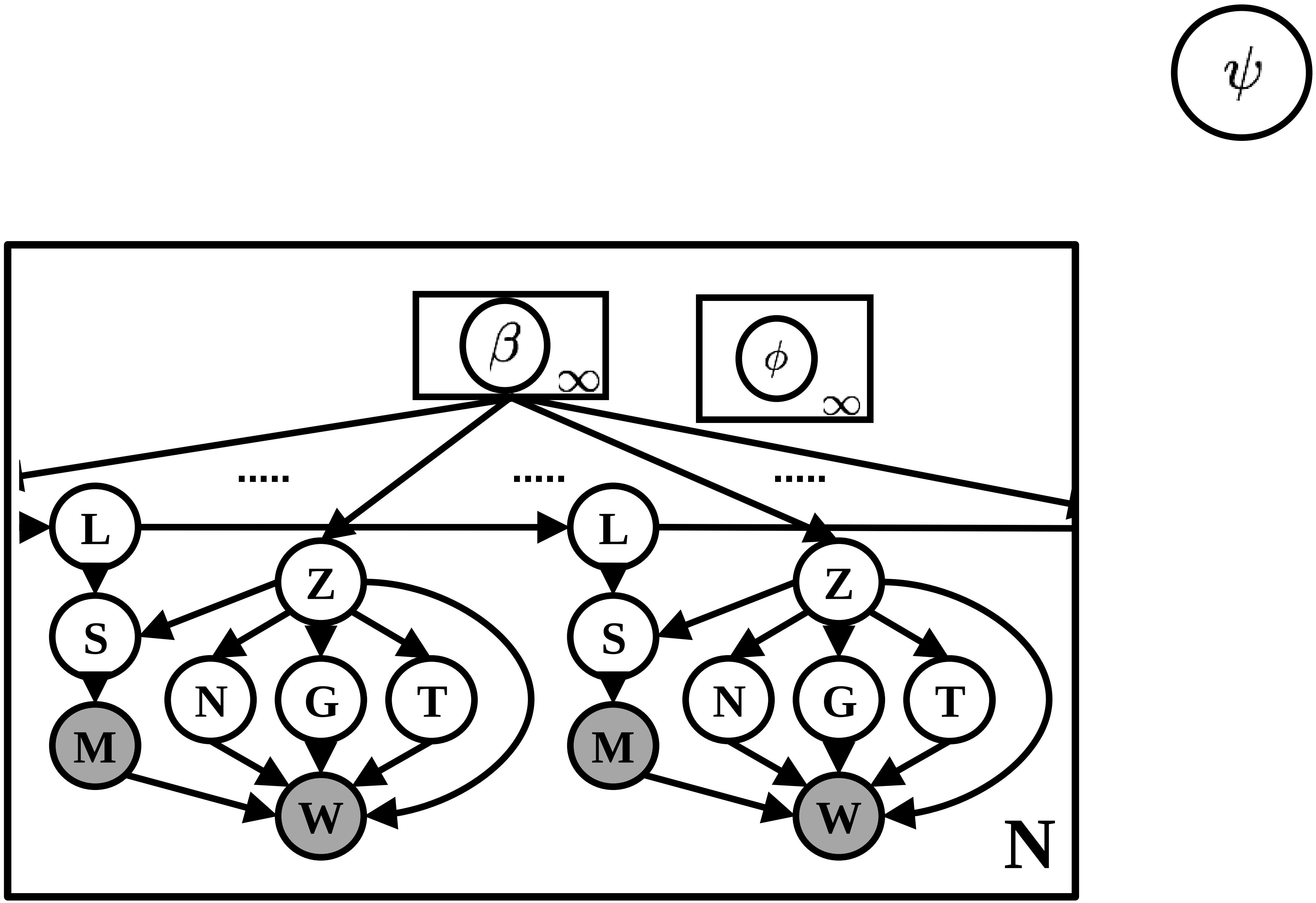
# Global Coreference Resolution



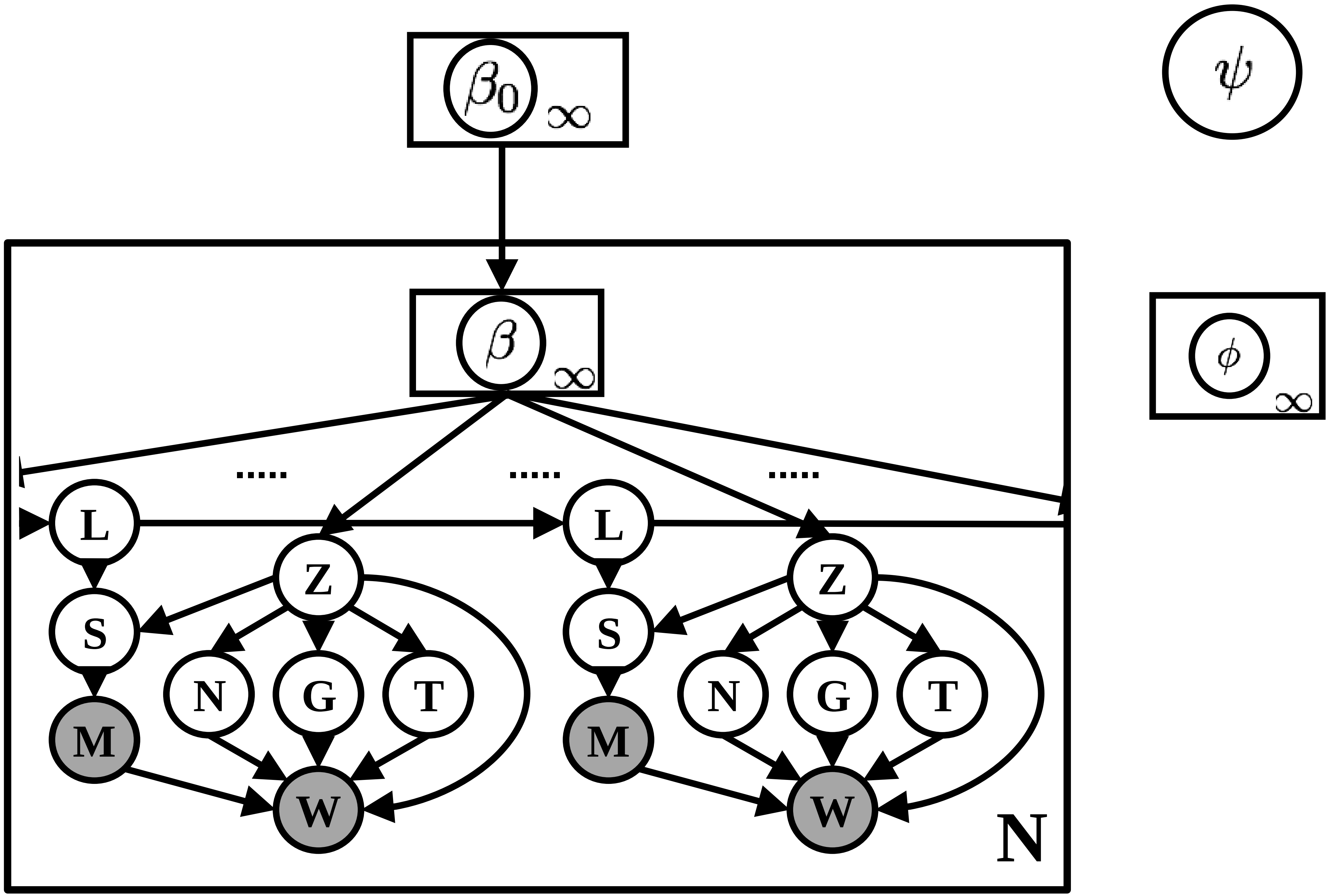
# Global Entity Model



# Global Entity Model

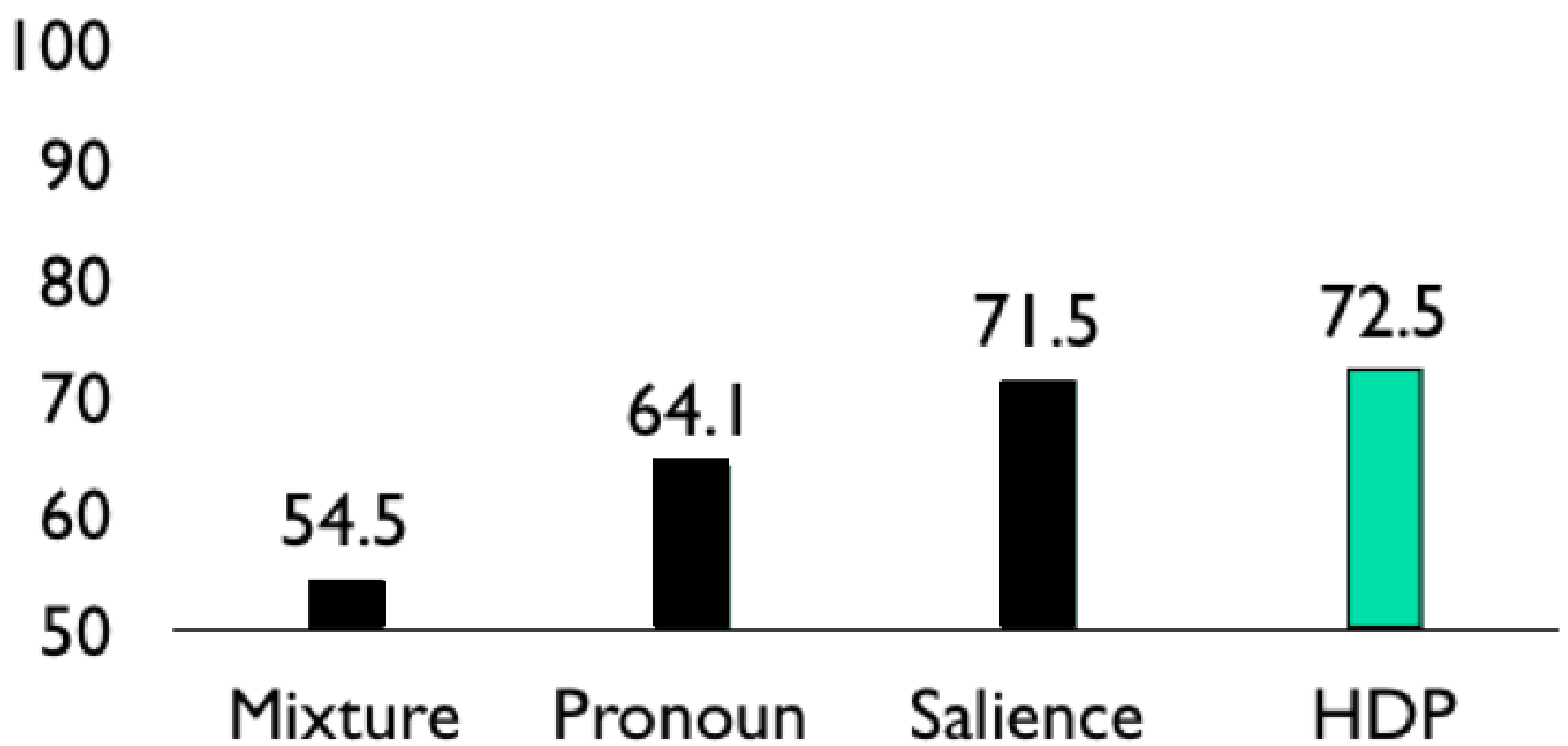


# Global Entity Model



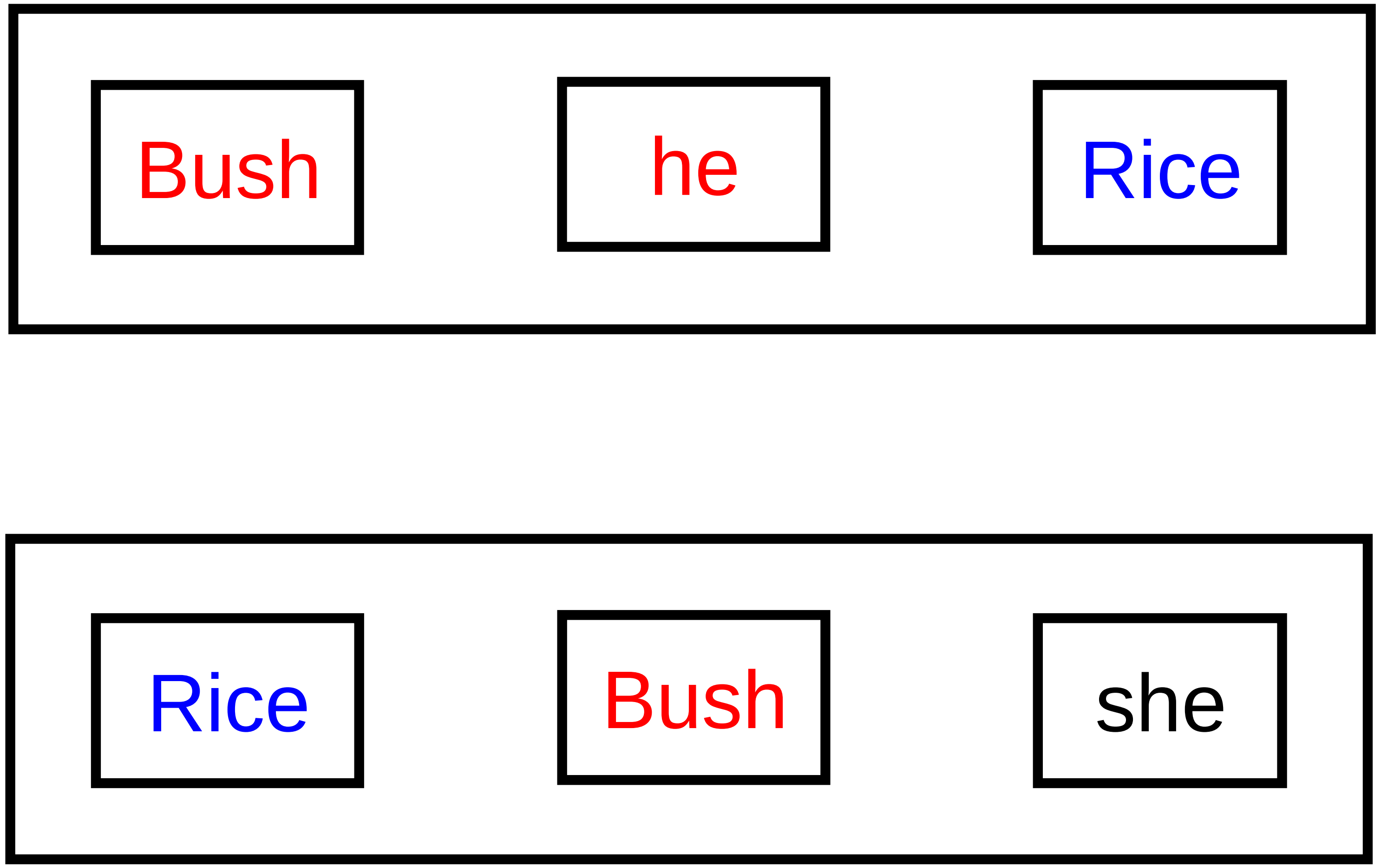
# HDP Model

MUC $F_1$




The Weir Group , whose headquarters is in the U.S is a large specialized corporation. This power plant , which , will be situated in Jiangsu, has a large generation capacity.

# Global Entity Resolution



# Experiments

- MUC6 English NWIRE (all mentions)
  - 53.6 F1* [Cardie and Wagstaff 99] Unsupervised
  - 70.3 F1 [Unsup Entity-Mention] Unsupervised
  - 73.4 F1 [McCallum & Wellner 04] Supervised
  - 81.3 F1 [Luo et al 04] Supervised++

** MUC score*

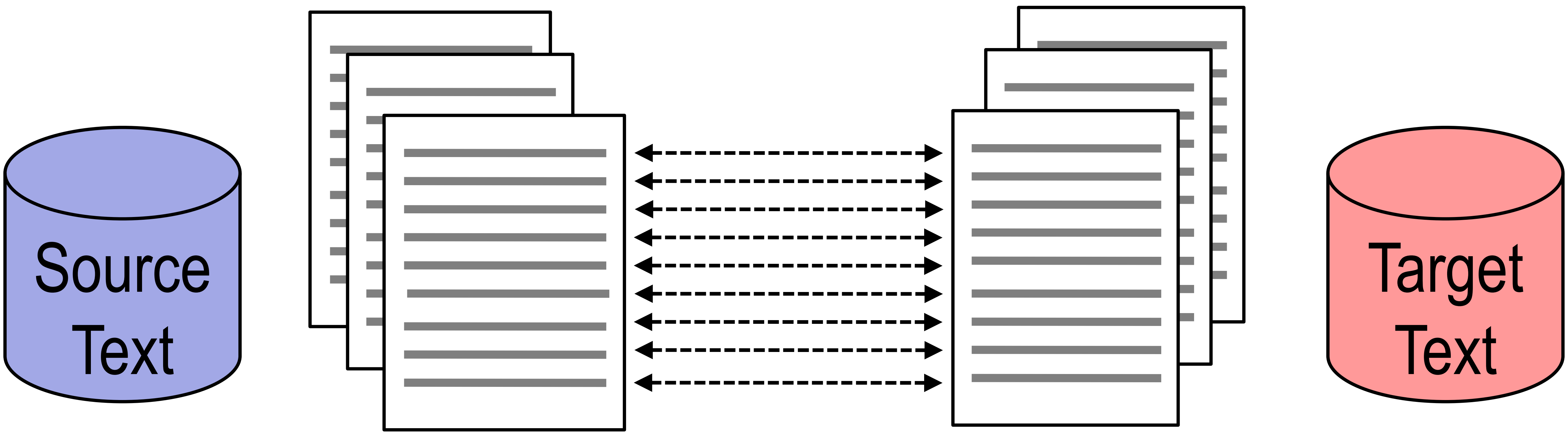
# Summary

- Fully generative unsupervised coref model
  - Basic model of pronoun structure
  - Sequential model of local attentional state
  - HDP global coreference model ties documents
- Competitive with supervised results
  - Many features not exploited
  - Still lots of room to improve!

# Outline

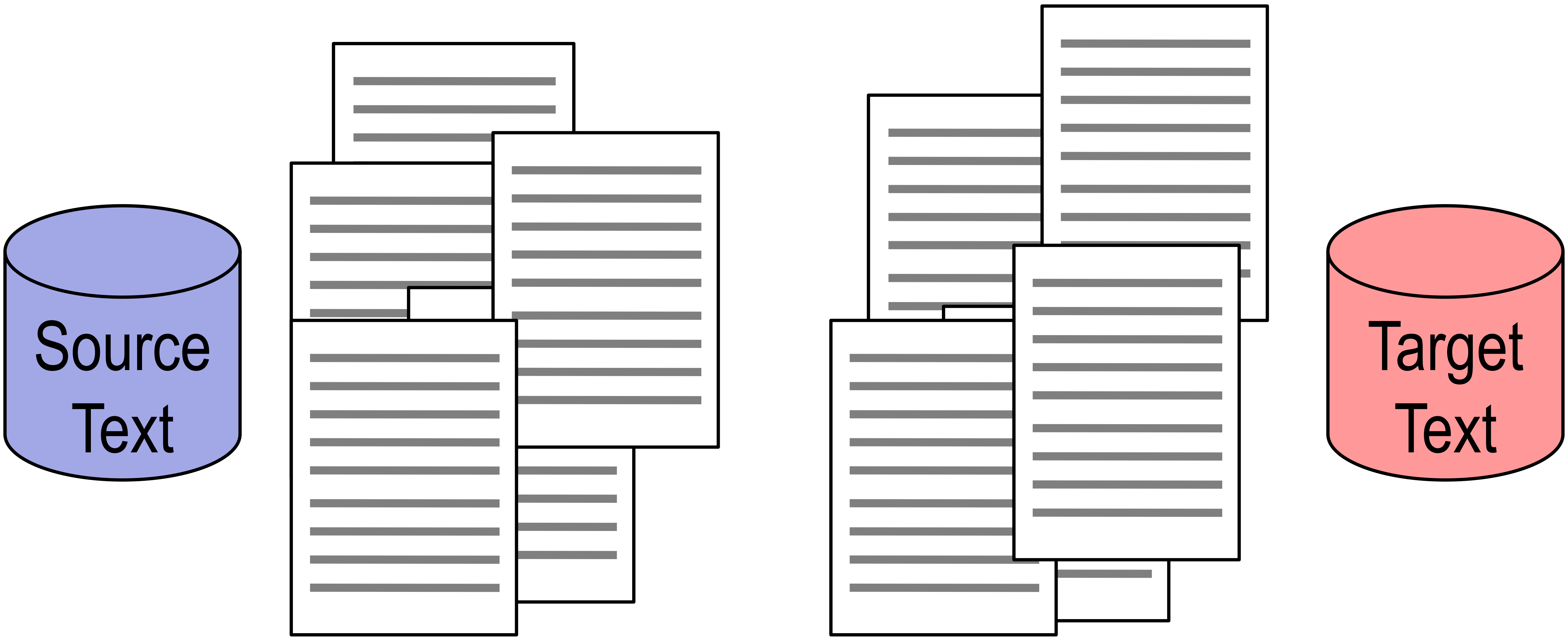
- Unsupervised Grammar Refinement
- Unsupervised Coreference Resolution
- Unsupervised Translation Mining

# Standard MT Approach




- Trained using parallel sentences
- May not always be available

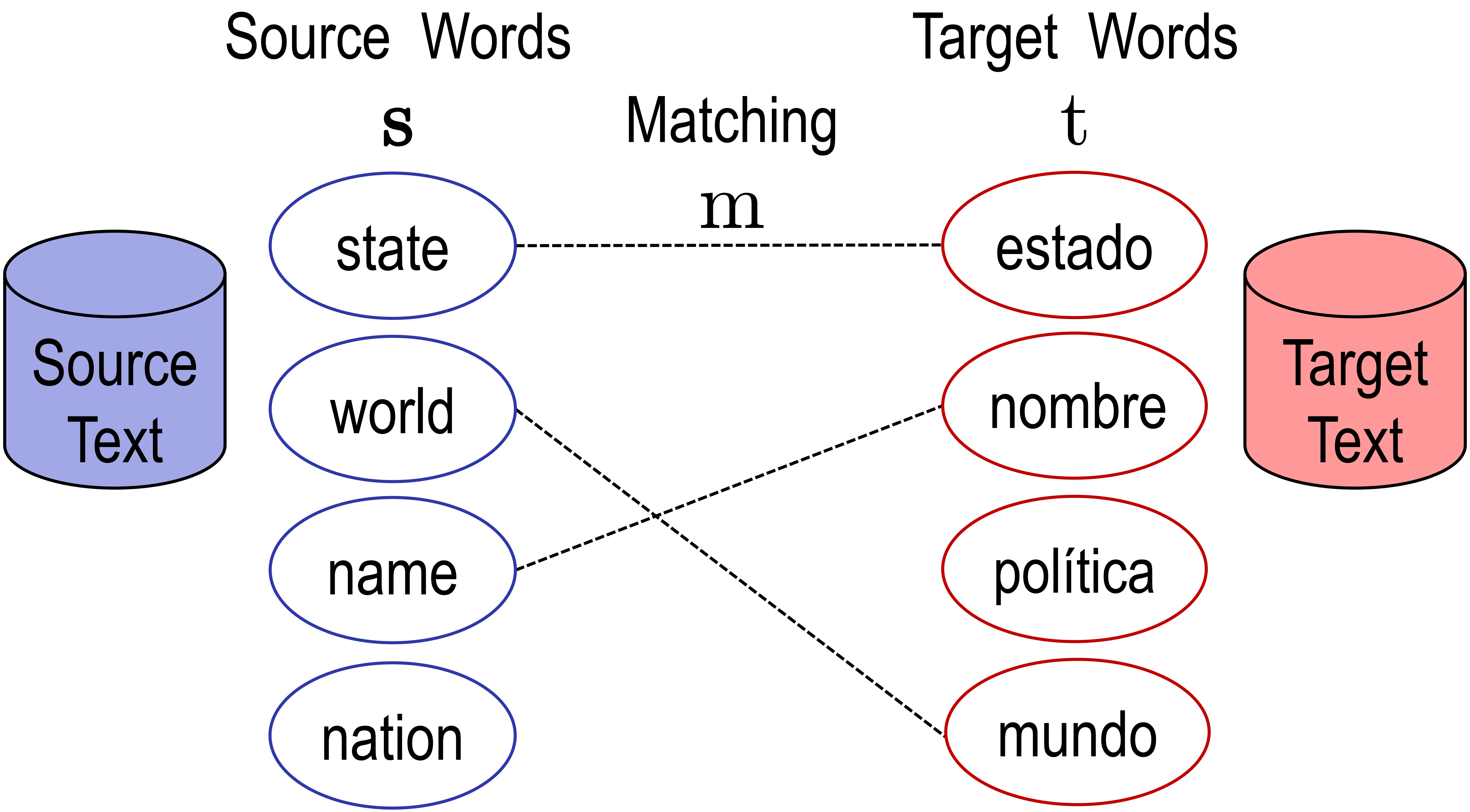
# MT from Monotext



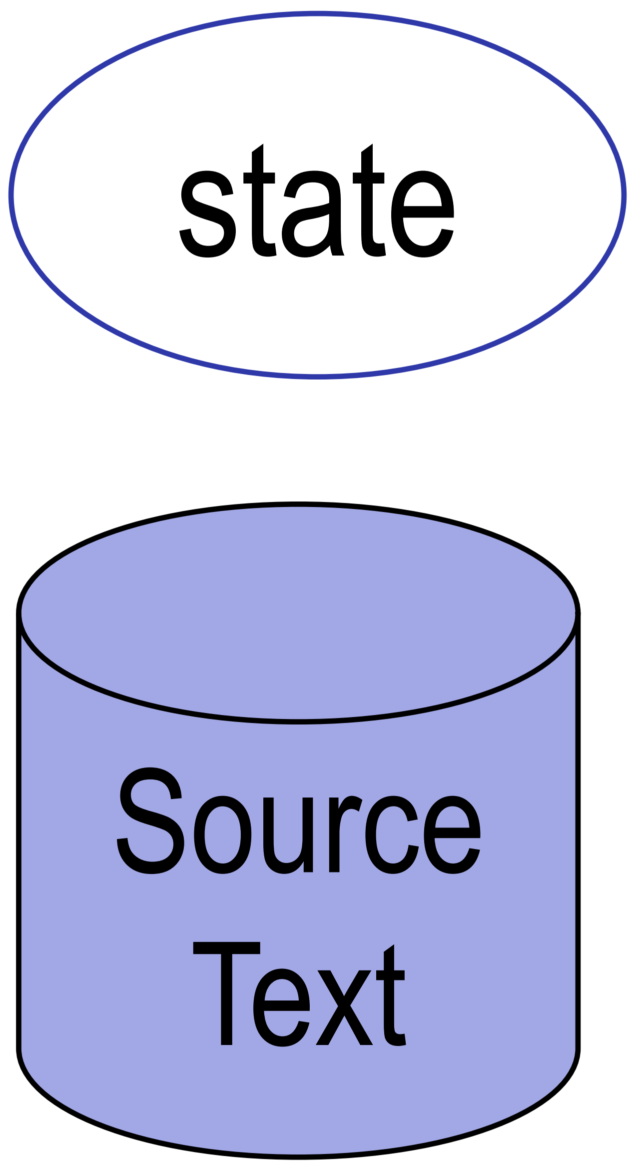

- Translation without parallel text?

[Fung 95, Koehn and Knight 02, Haghighi and Klein 08]

# Task: Lexicon Induction



# Data Representation




Orthographic Features

| | |
|---|---|
| #st | 1.0 |
| tat | 1.0 |
| te# | 1.0 |

Context Features

| | |
|---|---|
| world | 20.0 |
| politics | 5.0 |
| society | 10.0 |

# Data Representation

state

Source Text

**Orthographic Features**

| | |
|---|---|
| #st | 1.0 |
| tat | 1.0 |
| te# | 1.0 |

**Context Features**

| | |
|---|---|
| world | 20.0 |
| politics | 5.0 |
| society | 10.0 |

estado

Target Text

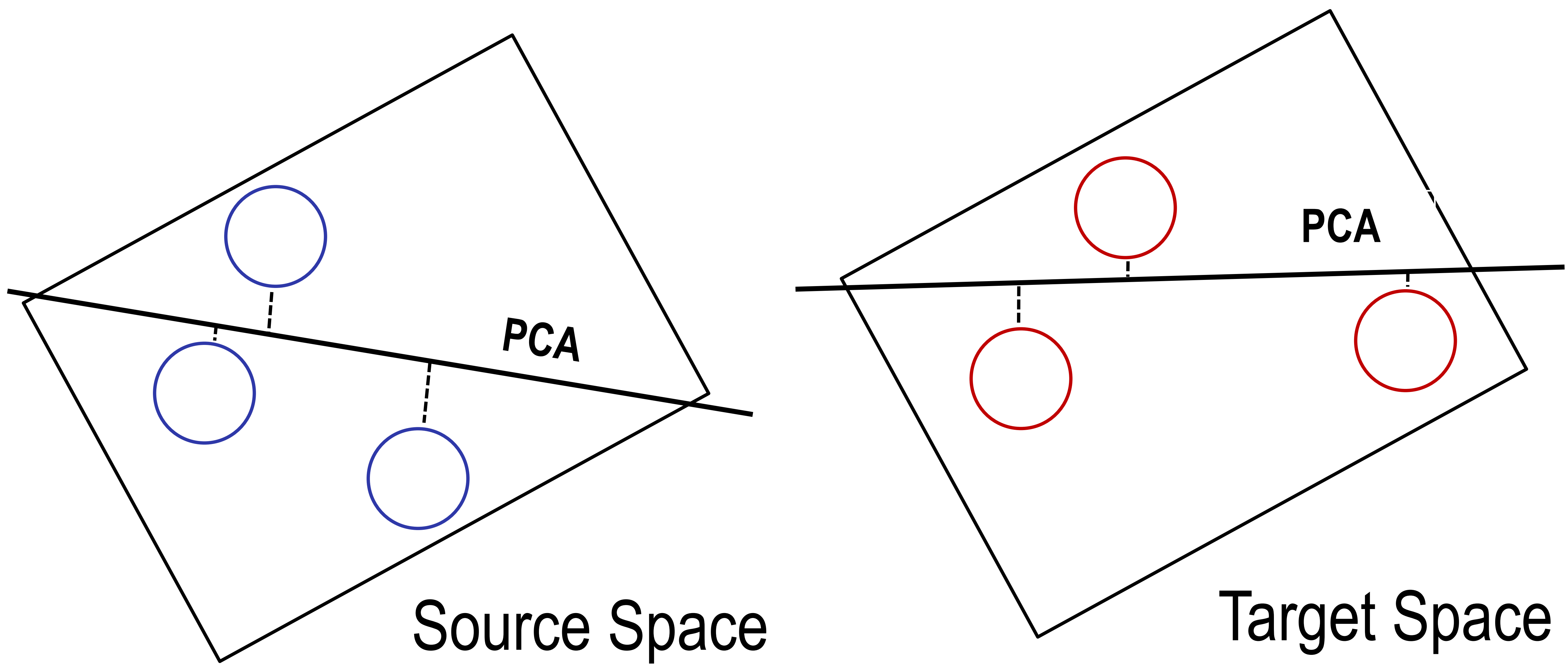
**Orthographic Features**

| | |
|---|---|
| #es | 1.0 |
| sta | 1.0 |
| do# | 1.0 |

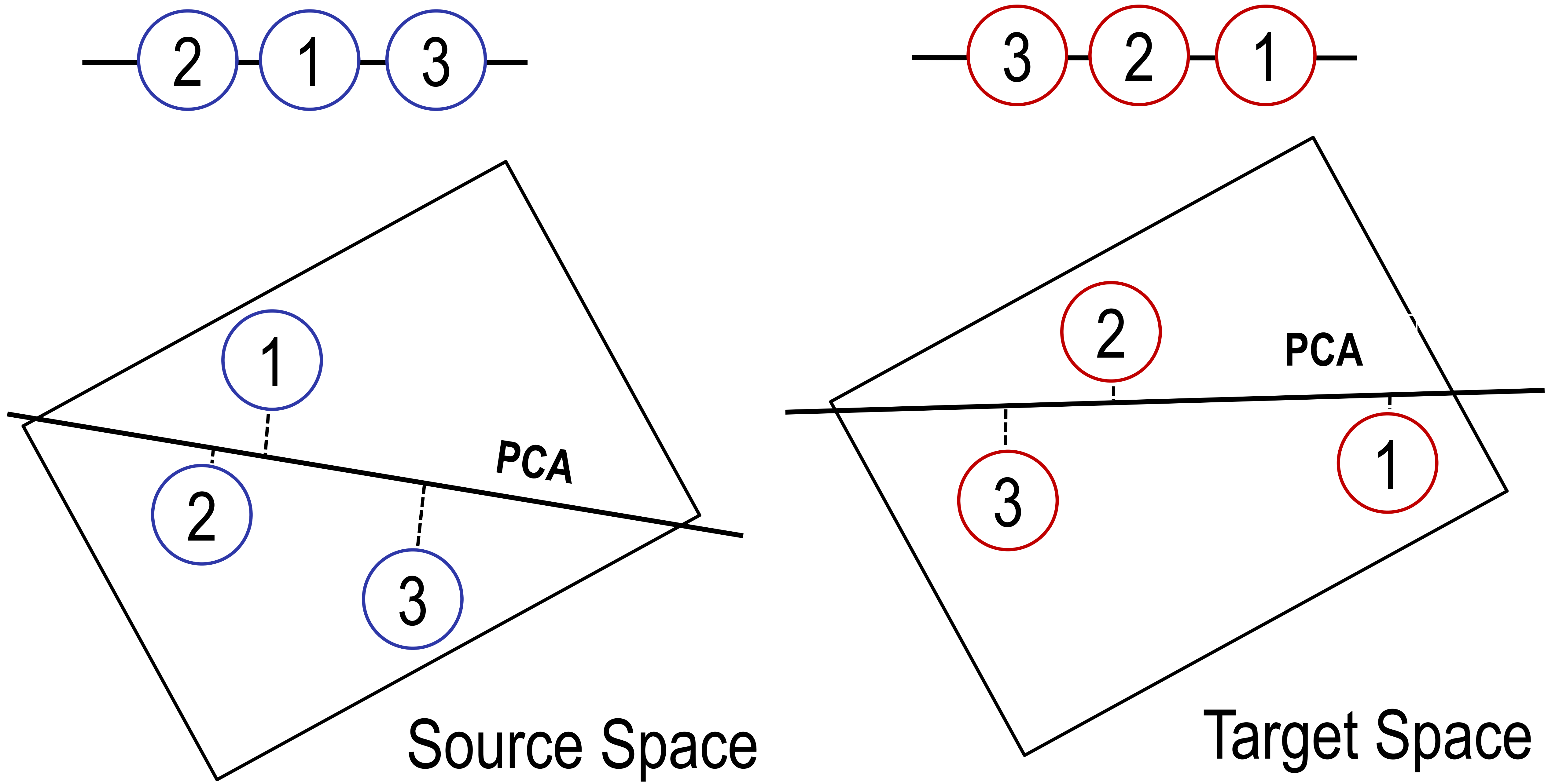
**Context Features**

| | |
|---|---|
| mundo | 17.0 |
| politica | 10.0 |
| sociedad | 6.0 |

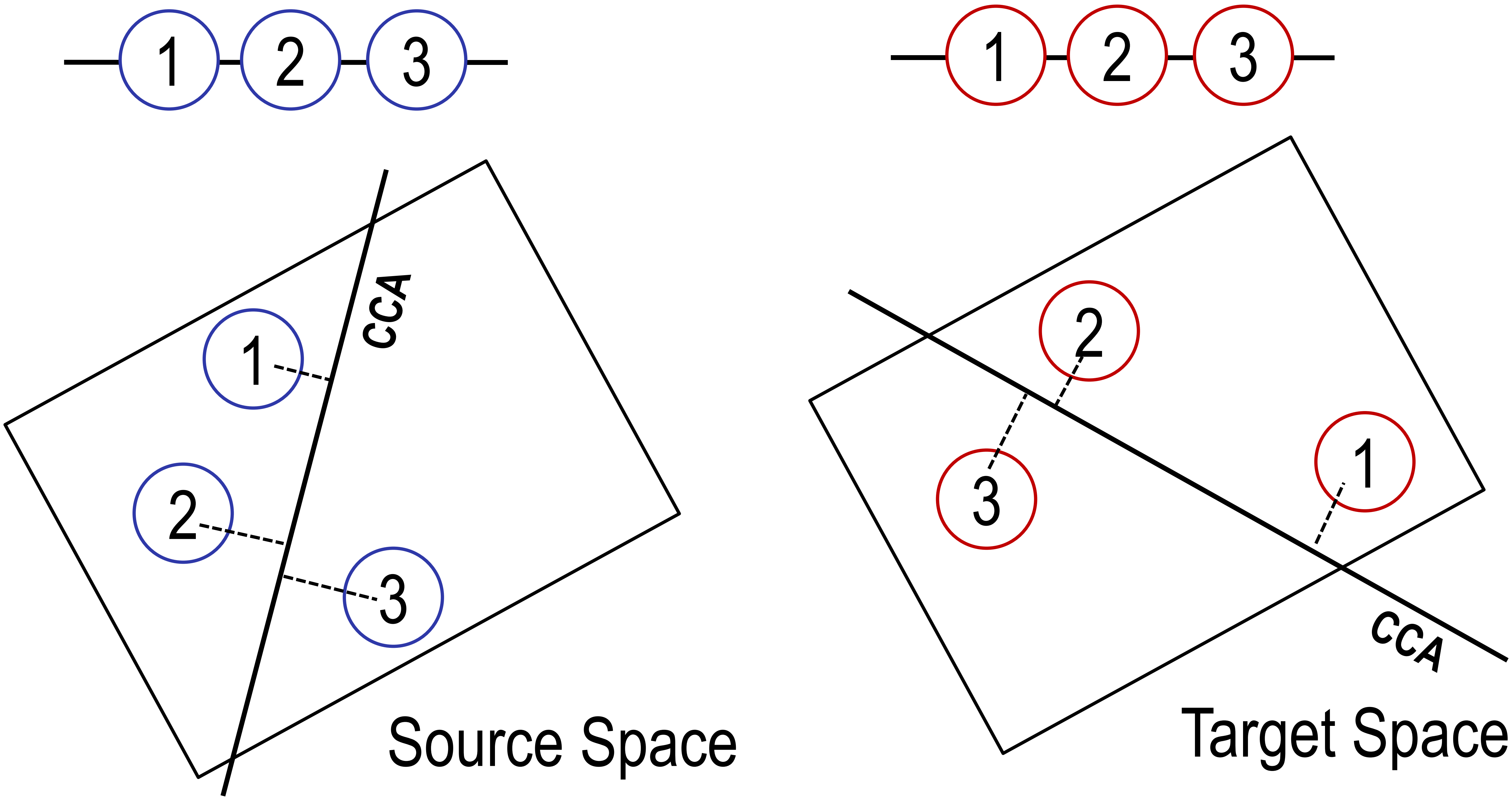
# Canonical Correlation Analysis



# Canonical Correlation Analysis

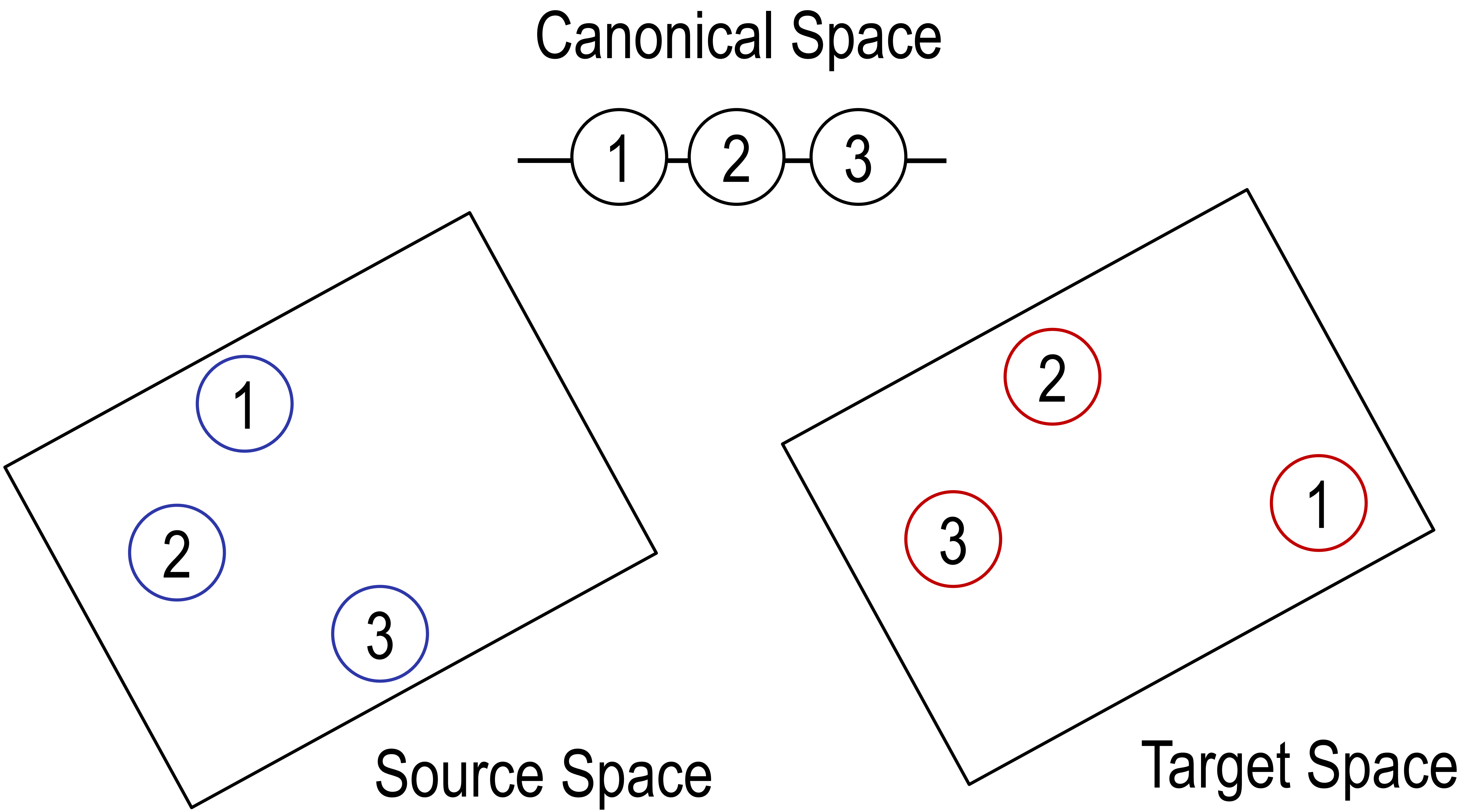


# Canonical Correlation Analysis



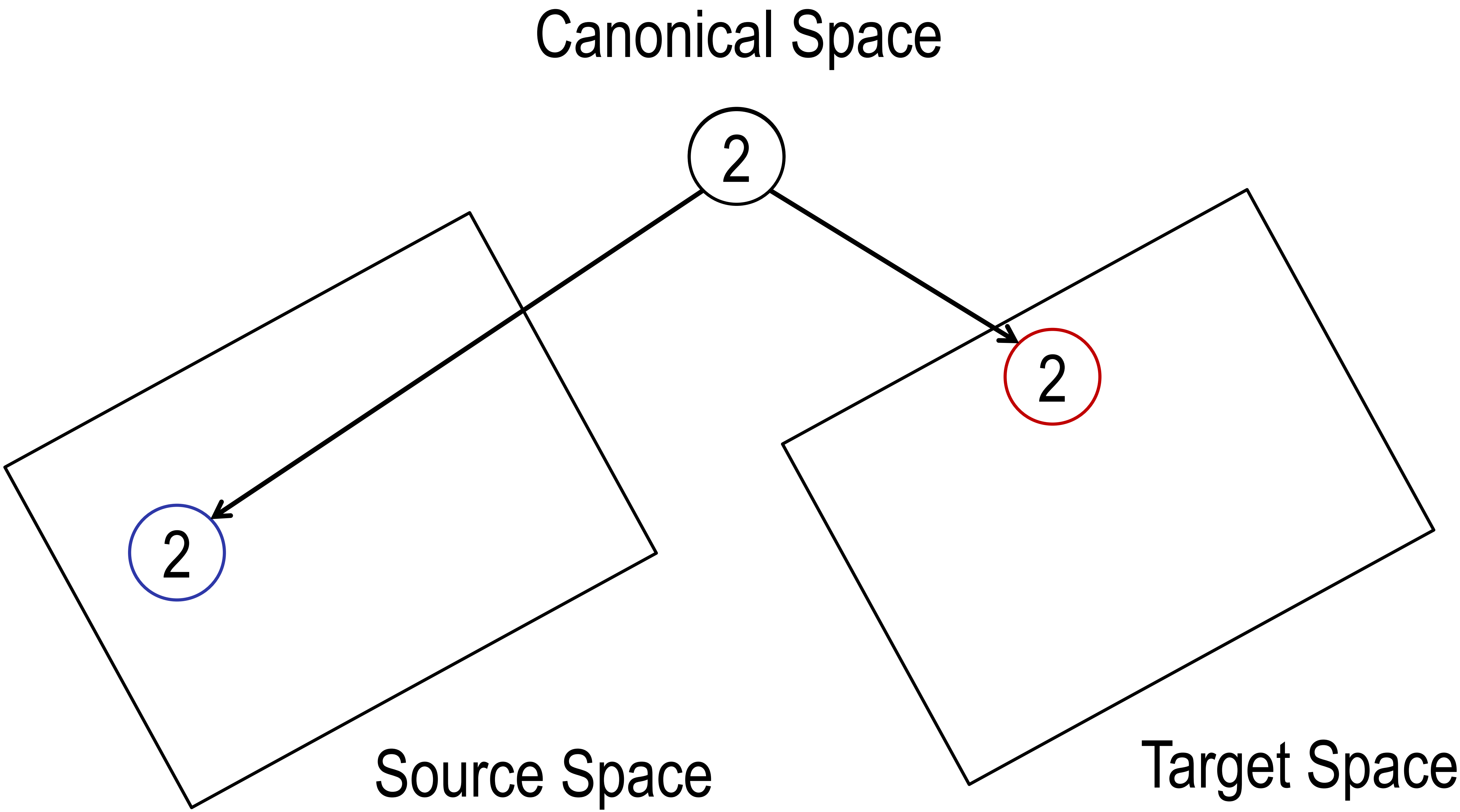
[Bach and Jordan 06]

# Canonical Correlation Analysis

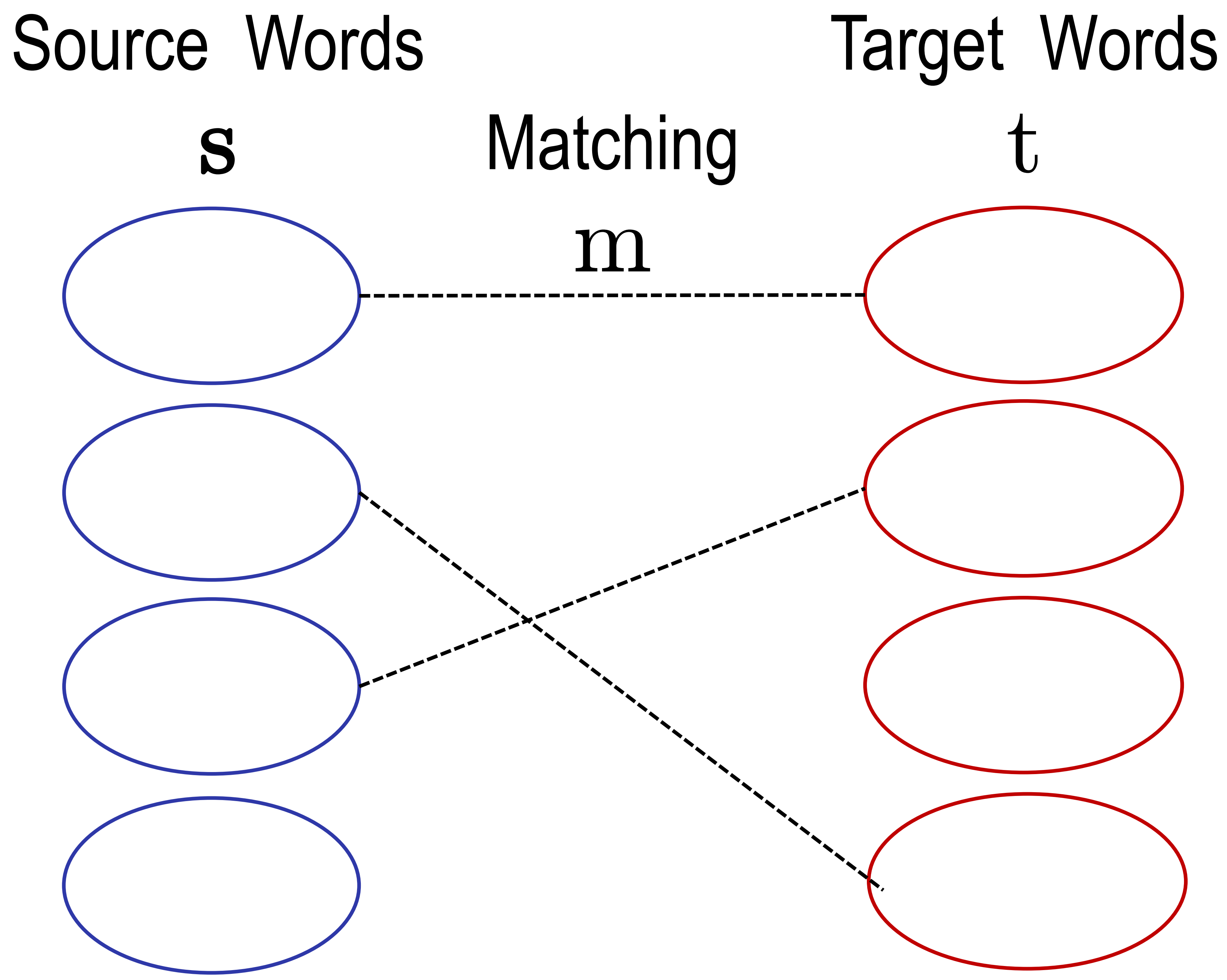


[Bach and Jordan 06]

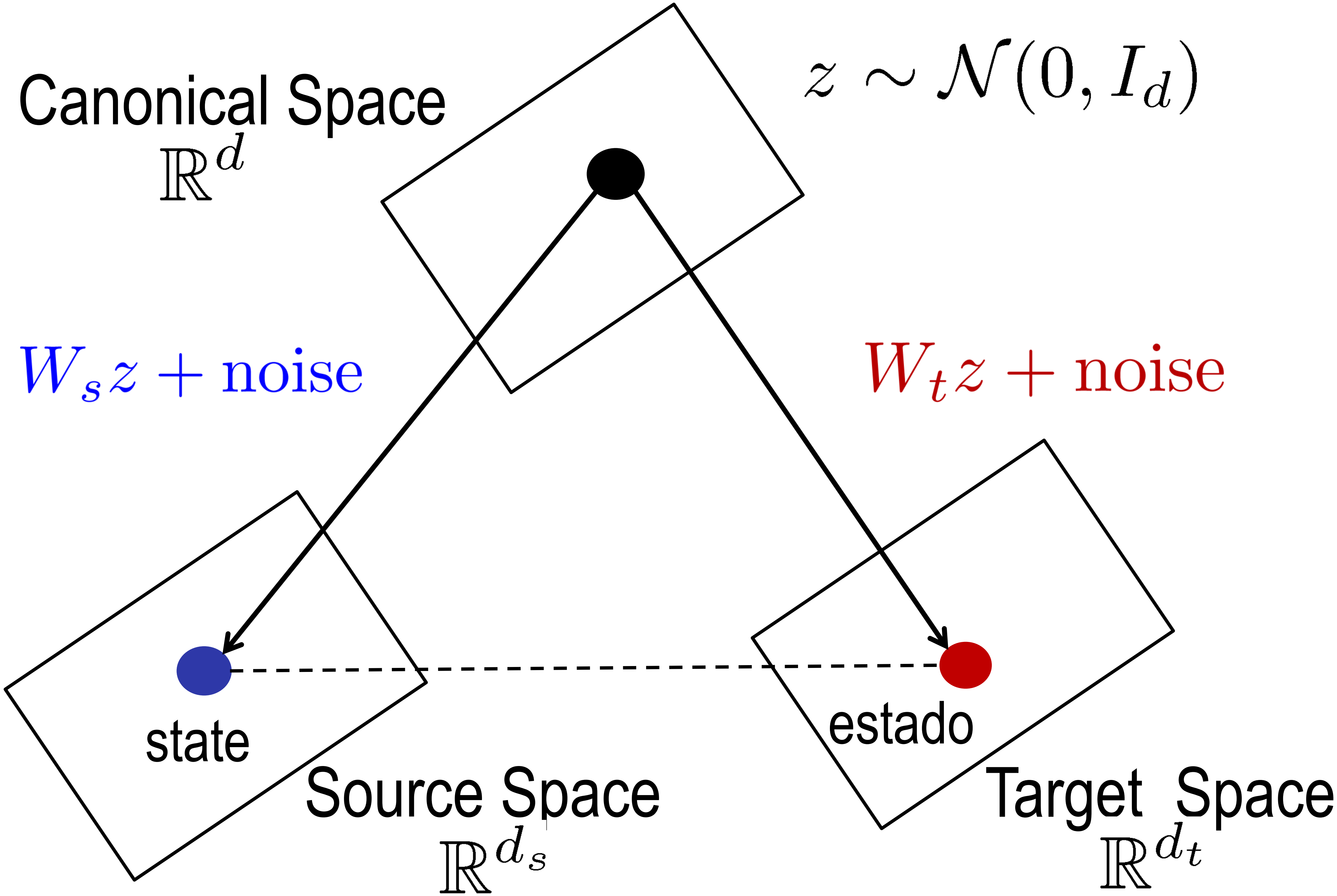
# Canonical Correlation Analysis



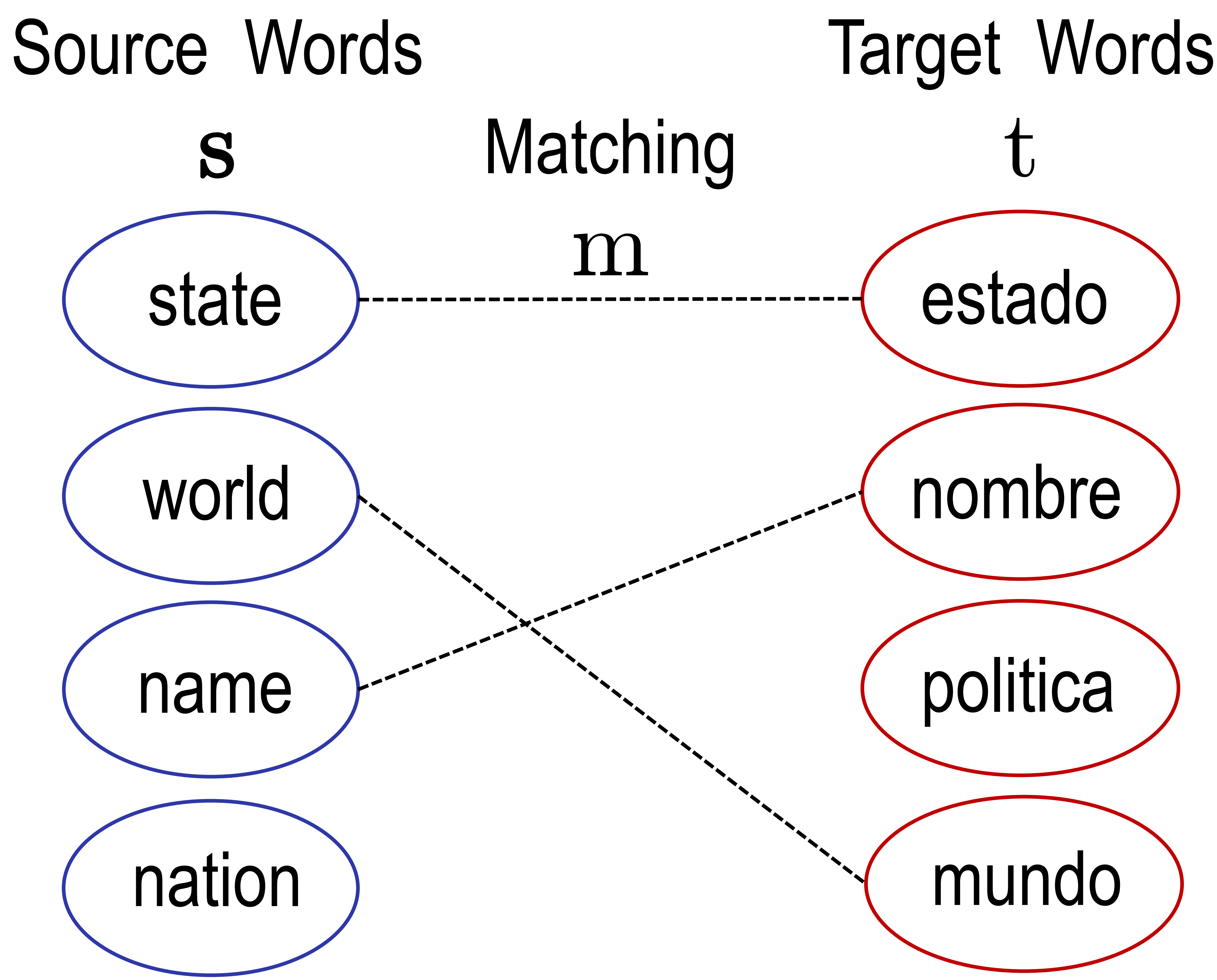
# Generative Model



# Generative Model



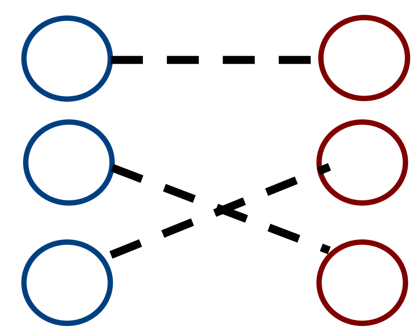
# Generative Model



# Learning: EM?

**E-Step:** Obtain posterior over matching

$$P(\mathbf{m}|\mathbf{s}, \mathbf{t})$$



**M-Step:** Maximize CCA Parameters

$$\max_{(W_s, W_t)} \mathbb{E}_{P(\mathbf{m}|\mathbf{s},\mathbf{t})} \left[ \sum_{(i,j)\in\mathbf{m}} \log p(s_i, t_j | \mathbf{m}; W_s, W_t) \right]$$

# Inference: Hard EM

**Hard E-Step:** Find best matching

$$\begin{aligned} w_{ij} &= \log p(s_i, t_j | \mathbf{m}; W_s, W_t) - \log \mathrm{NULL}_S(s_i) \\ &\quad - \log \mathrm{NULL}_T(t_j) \end{aligned}$$

**M-Step:** Solve CCA

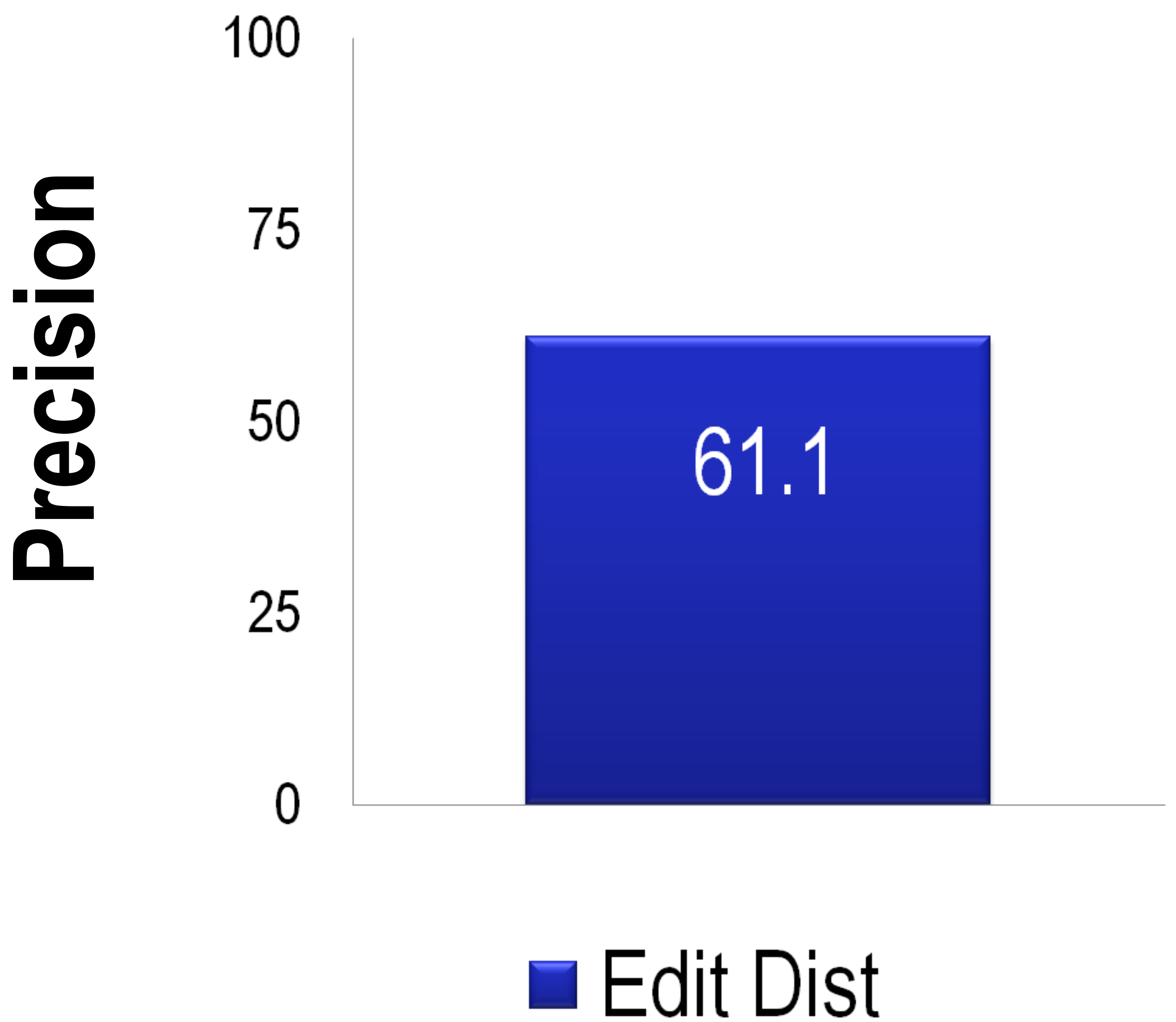
$$\max_{(W_s, W_t)} \left[ \sum_{(i,j) \in \mathbf{m}} \log p(s_i, t_j | \mathbf{m}; W_s, W_t) \right]$$

# Experimental Setup

- Data: 2K most frequent nouns, texts from Wikipedia
- Seed: 100 translation pairs
- Evaluation: Precision and Recall against lexicon obtained from Wiktionary
  - Report $p_{0.33}$, precision at recall 0.33

# Feature Experiments

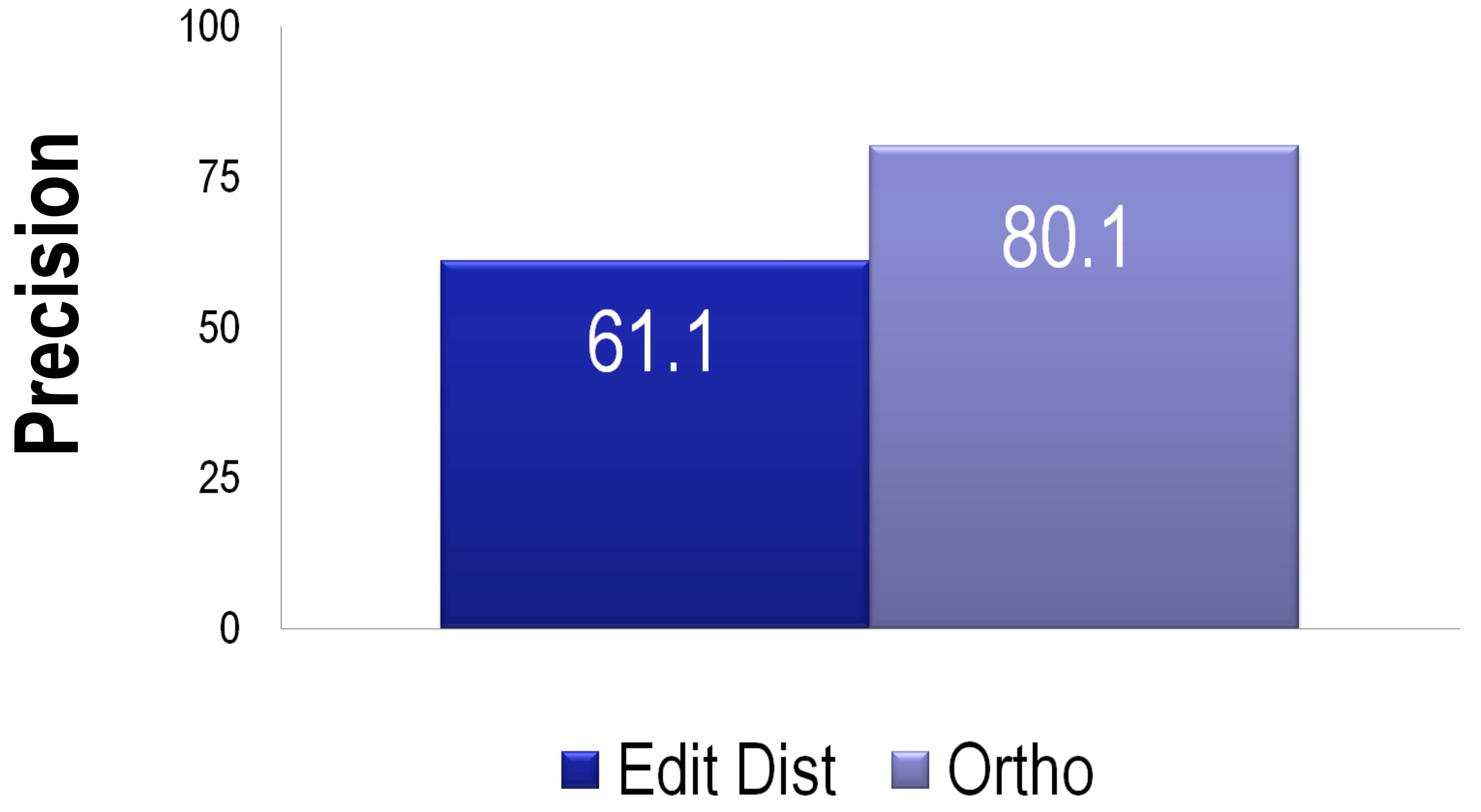
- Baseline: Edit Distance




4k EN-ES Wikipedia Articles

# Feature Experiments

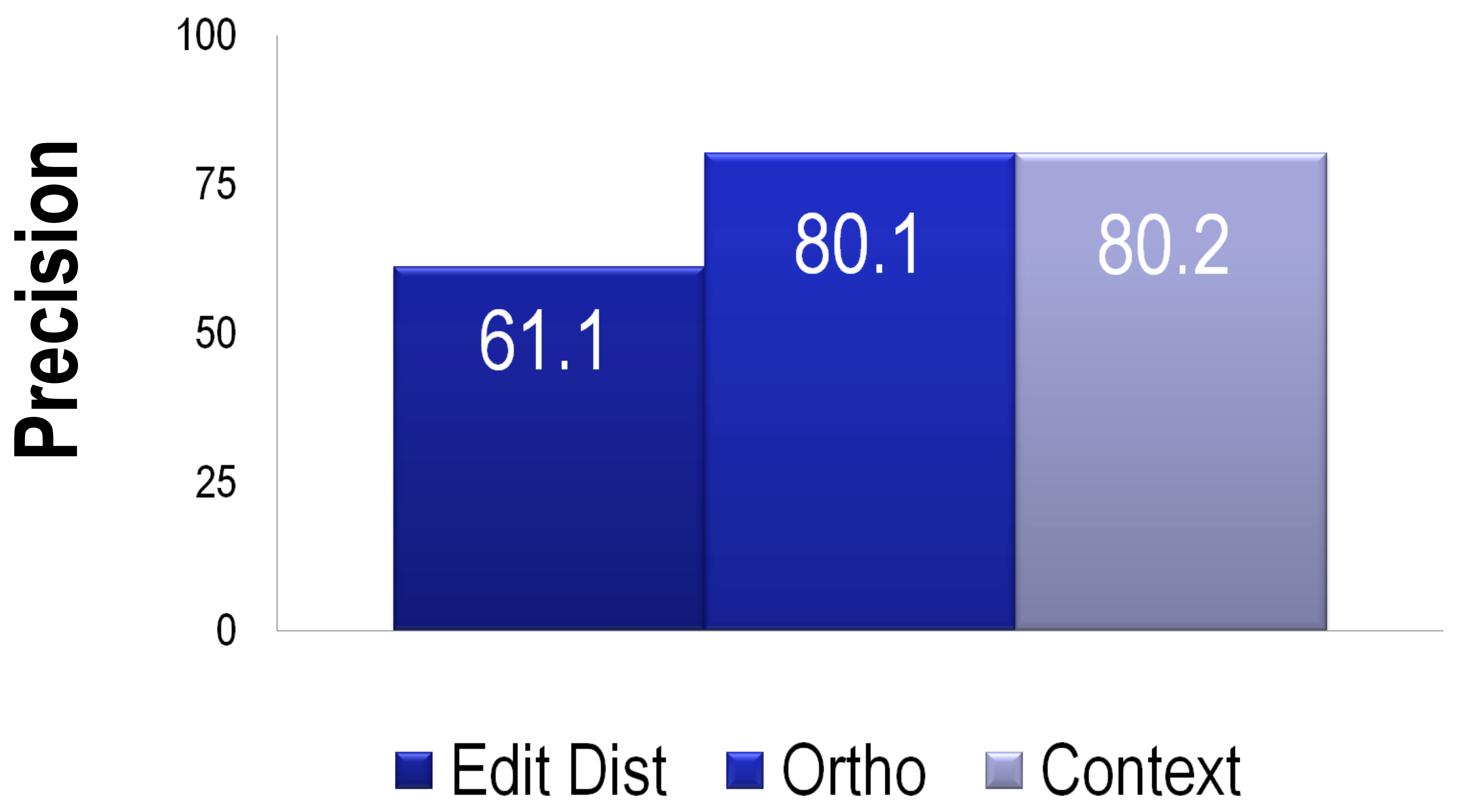
- MCCA: Only orthographic features




4k EN-ES Wikipedia Articles

# Feature Experiments

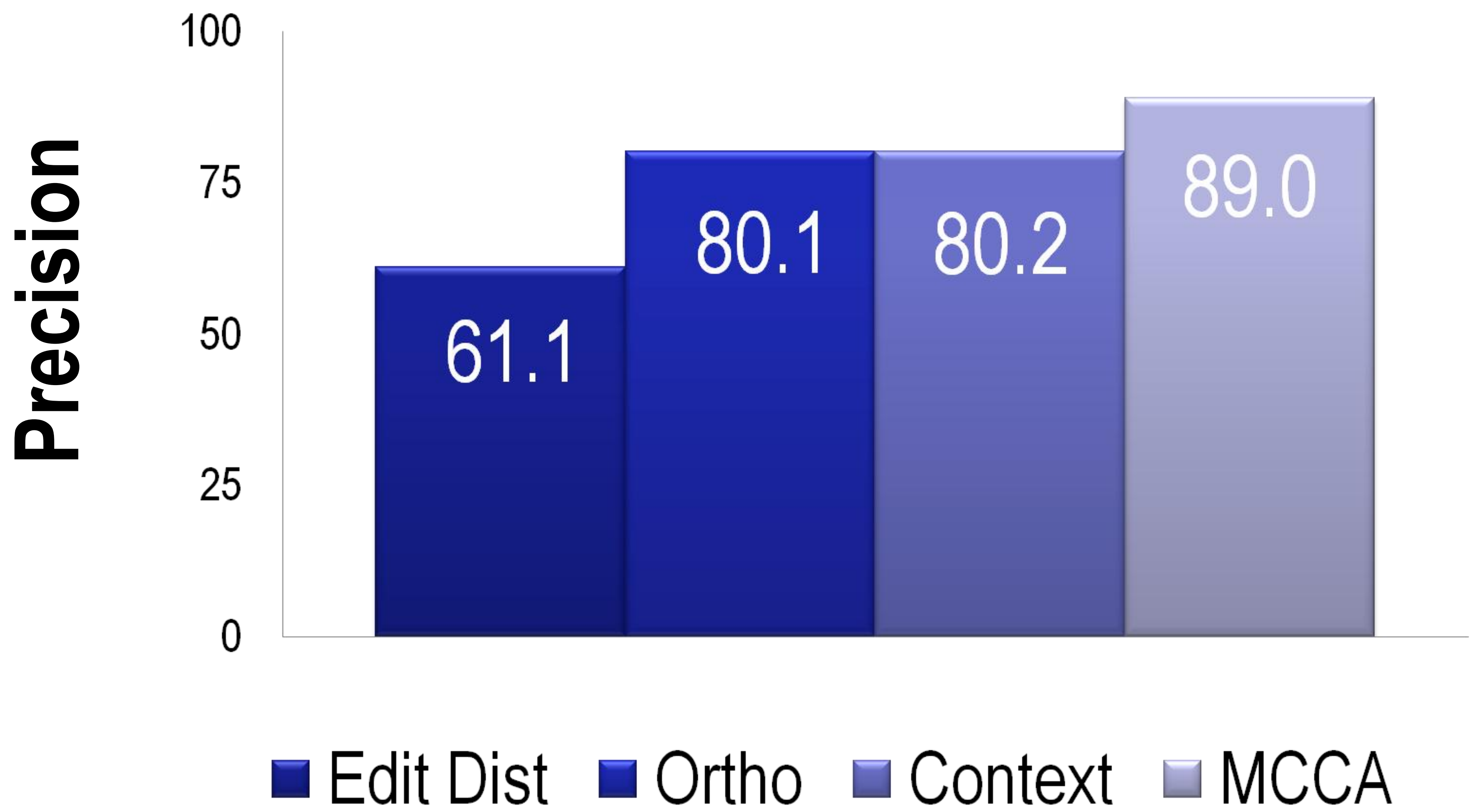
- MCCA: Only context features




4k EN-ES Wikipedia Articles

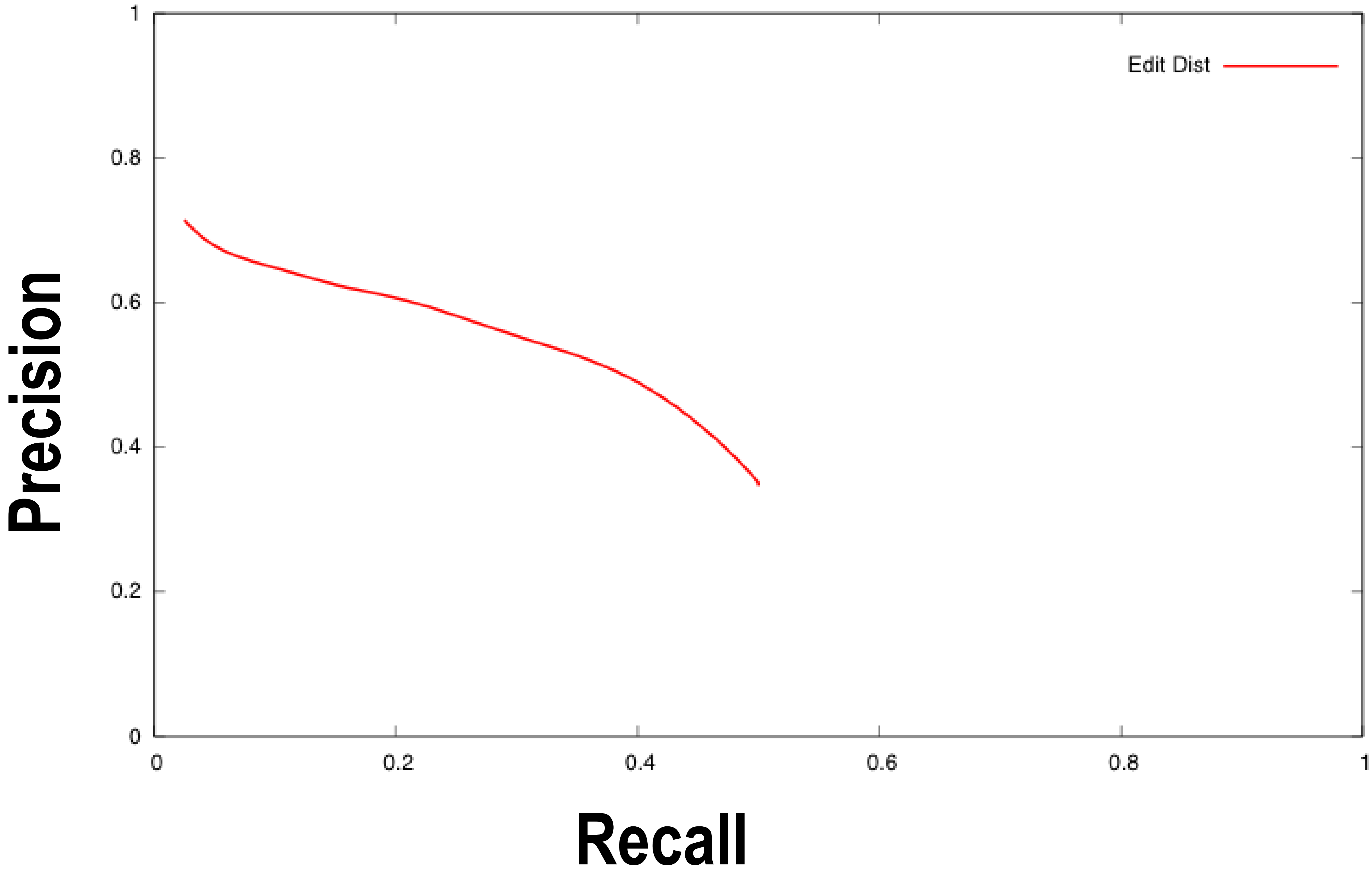
# Feature Experiments

- MCCA: Orthographic and context features

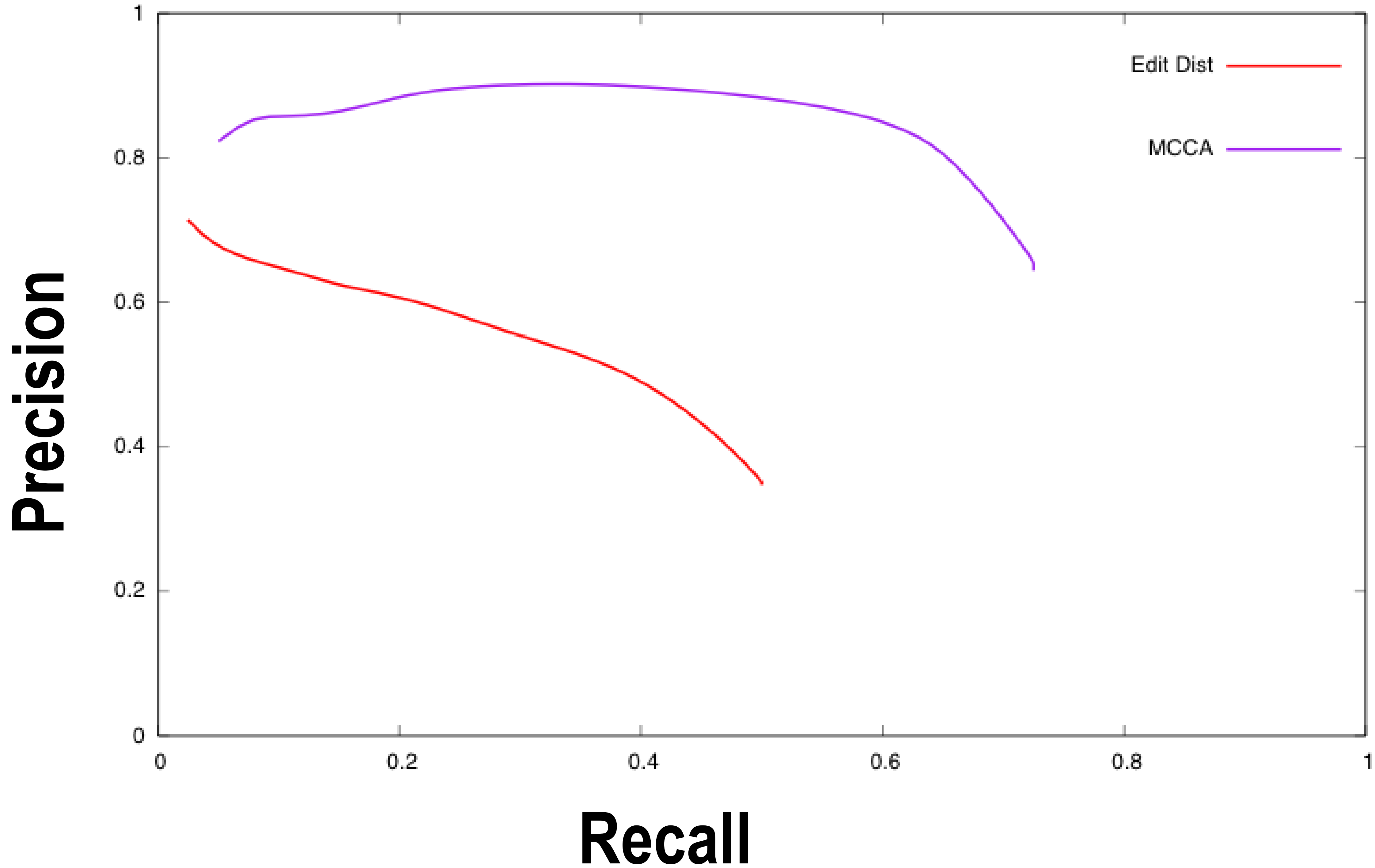



4k EN-ES Wikipedia Articles

# Feature Experiments

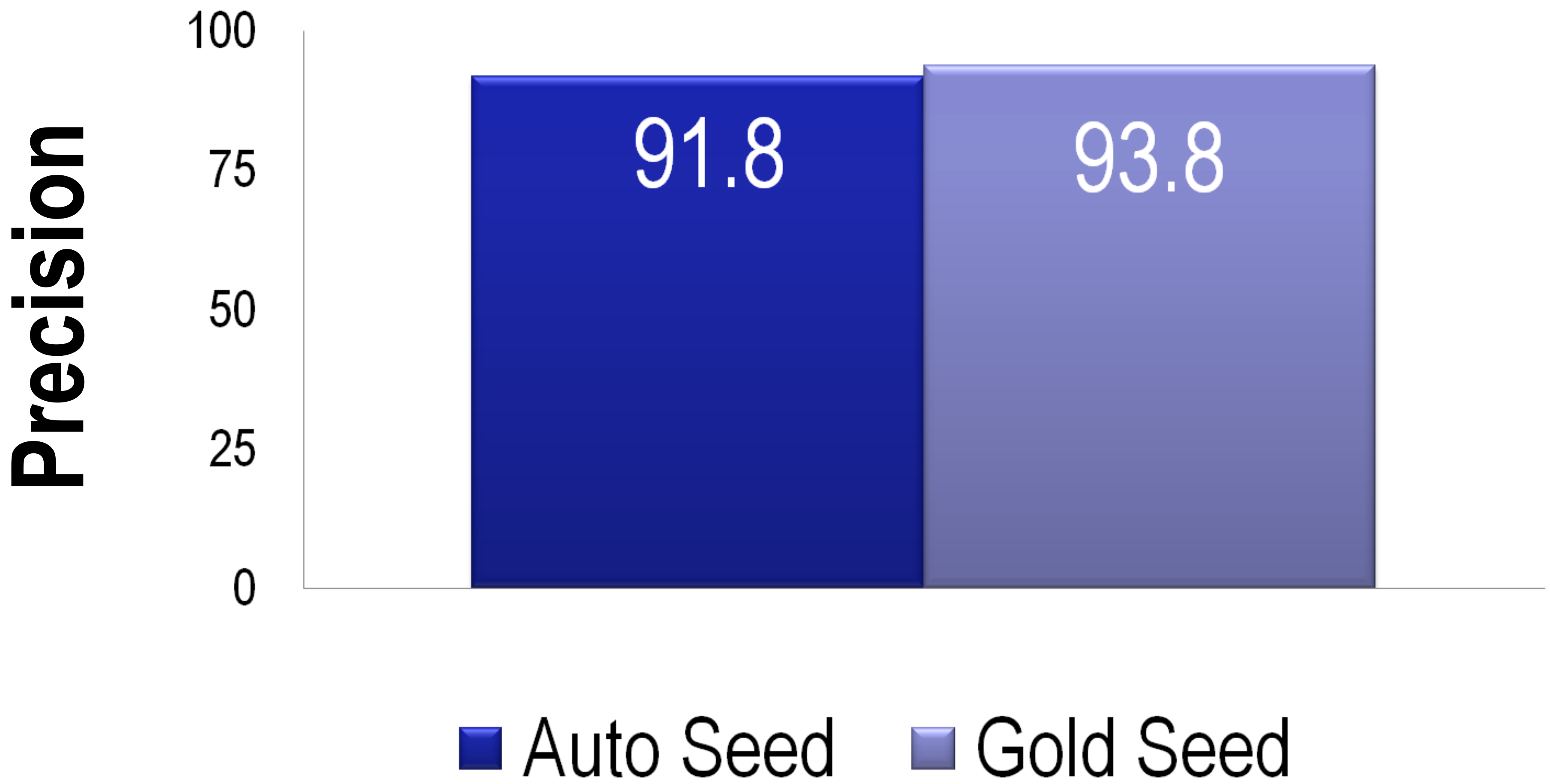


# Feature Experiments



# Seed Lexicon Source

- Automatic Seed
  - Edit distance seed [Koehn & Knight 02]




4k EN-ES Wikipedia Articles

# Analysis

English-Spanish

| Source | Target | Correct |
|---|---|---|
| education | educación | Y |
| pacto | pact | Y |
| stability | estabilidad | Y |
| corruption | corrupción | Y |
| tourism | turismo | Y |
| organisation | organización | Y |
| convenience | conveniencia | Y |
| syria | siria | Y |
| cooperation | cooperación | Y |
| culture | cultura | Y |
| protocol | protocolo | Y |
| north | norte | Y |
| health | salud | Y |
| action | reacción | N |

# Analysis

Top Non-Cognates

| | |
|---|---|
| health | salud |
| traceability | rastreabilidad |
| youth | juventud |
| report | informe |
| advantages | ventajas |

# Analysis

## Interesting Mistakes

| | |
|---|---|
| liberal | partido |
| Kirkhope | Gorsel |
| action | reacción |
| Albanians | Bosnia |
| a.m. | horas |
| Netherlands | Bretaña |

# Language Variation

| English-French | | |
|---|---|---|
| Source | Target | Correct |
| xenophobia | xénophobie | Y |
| corruption | corruption | Y |
| subsidiarity | subsidiarité | Y |
| programme | programme-cadre | N |
| traceability | traçabilité | Y |

# Language Variation

| English-Chinese | | |
|---|---|---|
| Source | Target | Correct |
| prices | 价格 | Y |
| network | 网络 | Y |
| population | 人口 | Y |
| reporter | 孙 | N |
| oil | 石油 | Y |

# Analysis

## Orthography Features

| Source Feature | Closest Target Features | Example Translations |
|---|---|---|
| #st | #es, est | (statue,estatua) |
| ty# | ad#, d# | (felicity,felicidad) |
| ogy | gía, gí | (geology,geología) |

## Context Features

| Source Feature | Closest Context Features |
|---|---|
| party | partido, izquierda |
| democrat | socialistas, demócratas |
| beijing | pekín, kioto |

# Summary

- Learned bilingual lexicon from monotext
  - Matching + CCA model
  - Possible even from unaligned corpora
  - Possible for non-related languages
  - High-precision, but much left to do!

# Conclusion

- Three cases of unsupervised learning of non-trivial linguistic structure for NLP problems
  - Incremental structure learning
  - Careful control of structured training
  - Targeted modeling choices
- In some cases, unsupervised systems are competitive with supervised systems (or better!)
- Much more left to do!

# Thank you!

nlp.cs.berkeley.edu

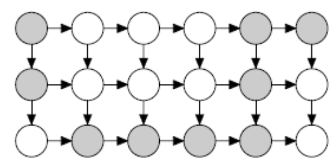
# Outline

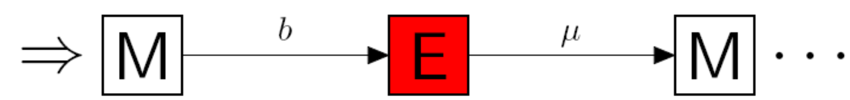
- Latent-Variable Grammar Learning
- Unsupervised Coreference Resolution
- Unsupervised Translation Mining
- Other Unsupervised Work

# Agreement-Based Learning

Problem: learning complex hidden-variable models

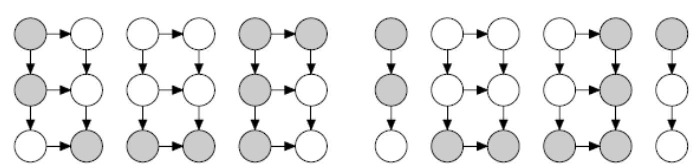
Traditional solution: approximate EM

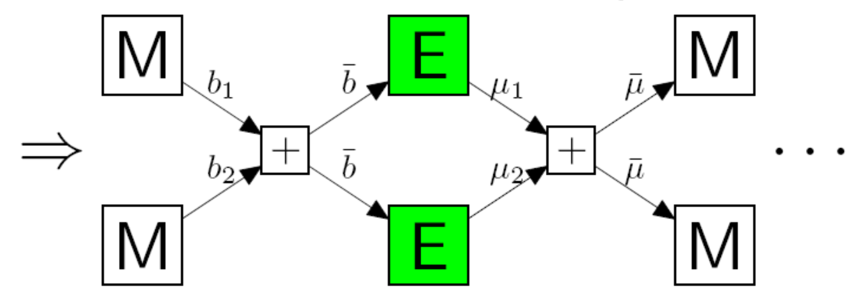





one intractable model

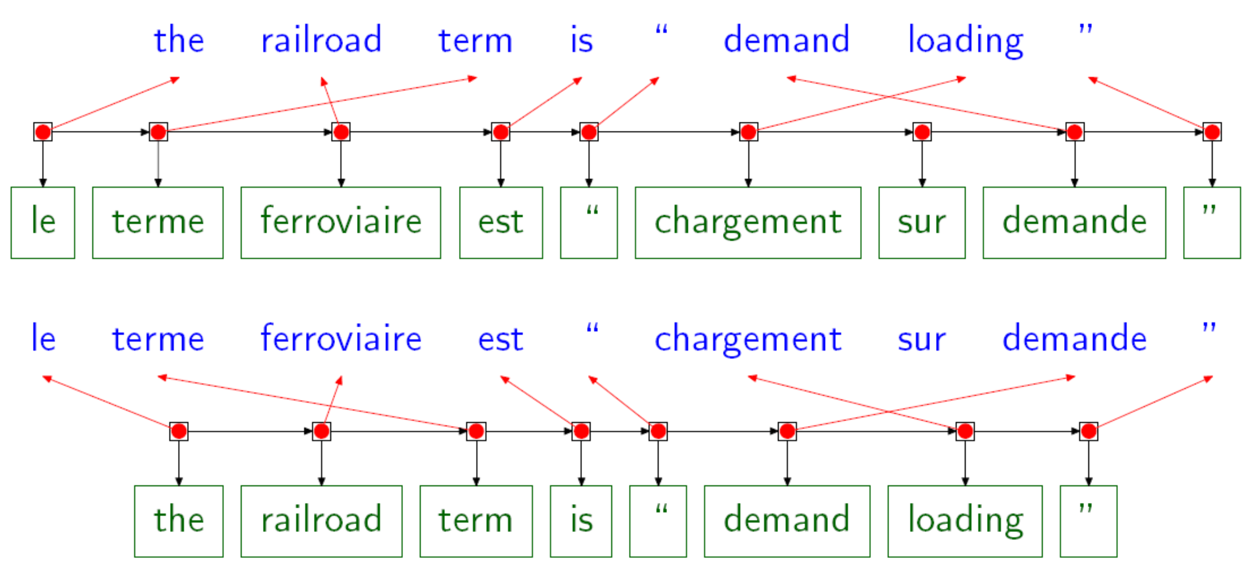
Our solution: product EM (train submodels to agree)

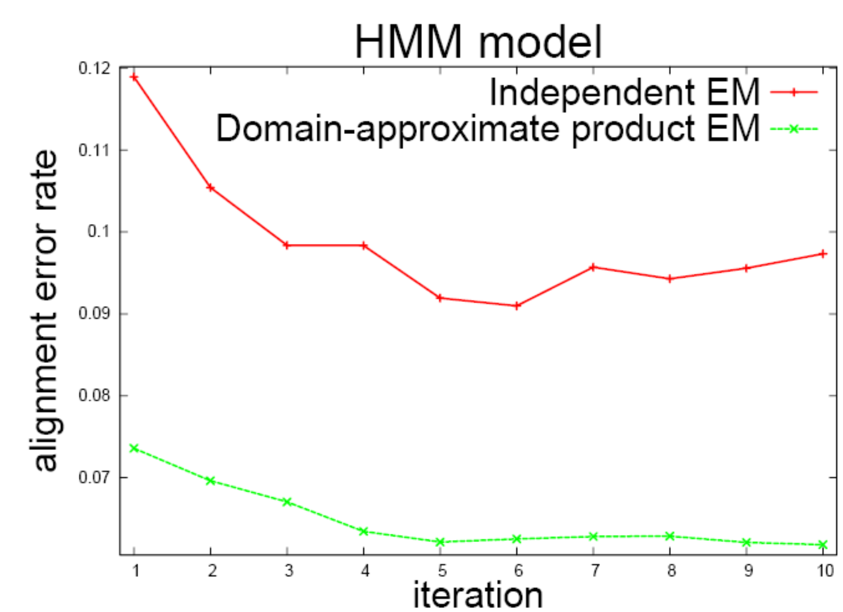





two tractable submodels

Applications: unsupervised NLP, phylogenetic HMMs



# Weakly Supervised Learning

Newly remodeled 2 Bdrms/1 Bath, spacious upper unit, located in Hilltop Mall area. Walking distance to shopping, public transportation, schools and park. Paid water and garbage. No dogs allowed.

Prototype Lists

| | |
|---|---|
| FEATURE | kitchen, laundry |
| LOCATION | near, close |
| TERMS | paid, utilities |
| SIZE | large, feet |
| RESTRICT | cat, smoking |

Information Extraction

| | | | |
|---|---|---|---|
| NN | president | IN | of |
| VBD | said | NNS | shares |
| CC | and | TO | to |
| NNP | Mr. | PUNC | . |
| JJ | new | CD | million |
| DET | the | VBP | are |

English POS

# Language Evolution

| Gloss | Latin | Italian | Spanish | Portuguese |
|---|---|---|---|---|
| Word/verb | verbum | verbo | verbo | verbu |
| Fruit | fructus | frutta | fruta | fruta |
| Laugh | ridere | ridere | reir | rir |
| Center | centrum | centro | centro | centro |
| August | augustus | agosto | agosto | agosto |
| Swim | natare | nuotare | nadar | nadar |