# Discovery of Linguistic Relations Using Lexical Attraction

Deniz Yuret

Overview

- Motivation
- Demonstration
- Theory, Learning, Algorithm
- Evaluation
- Contributions

Syntax and Semantics independently constrain linguistic relations

- I saw the Statue of Liberty flying over New York.

  – Lenat, 1984

- I hit the boy with the girl with long hair with a hammer with vengeance.

  – Schank, 1973

- Colorless green ideas sleep furiously.

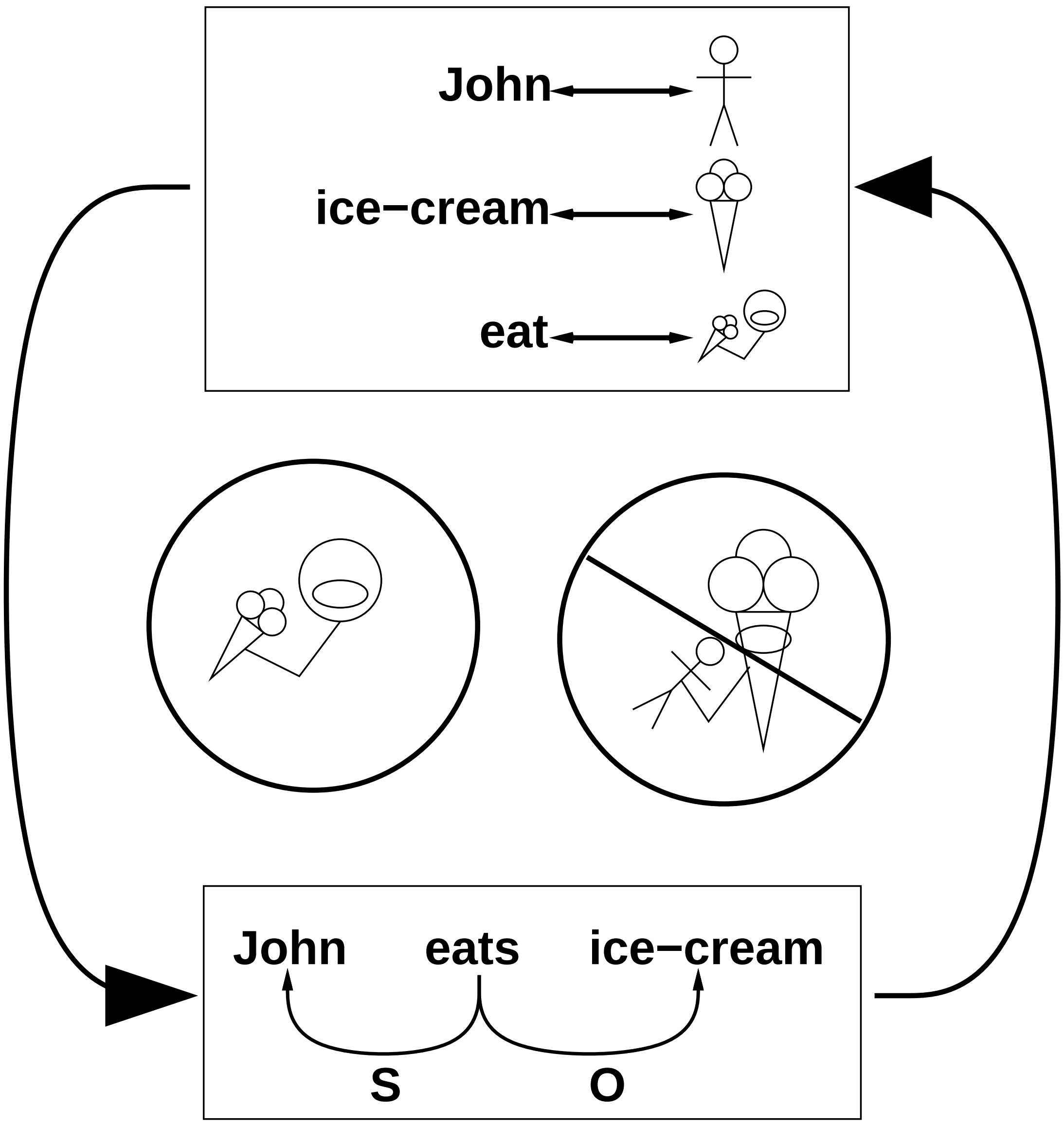
  – Chomsky, 1956

Contributions of this thesis

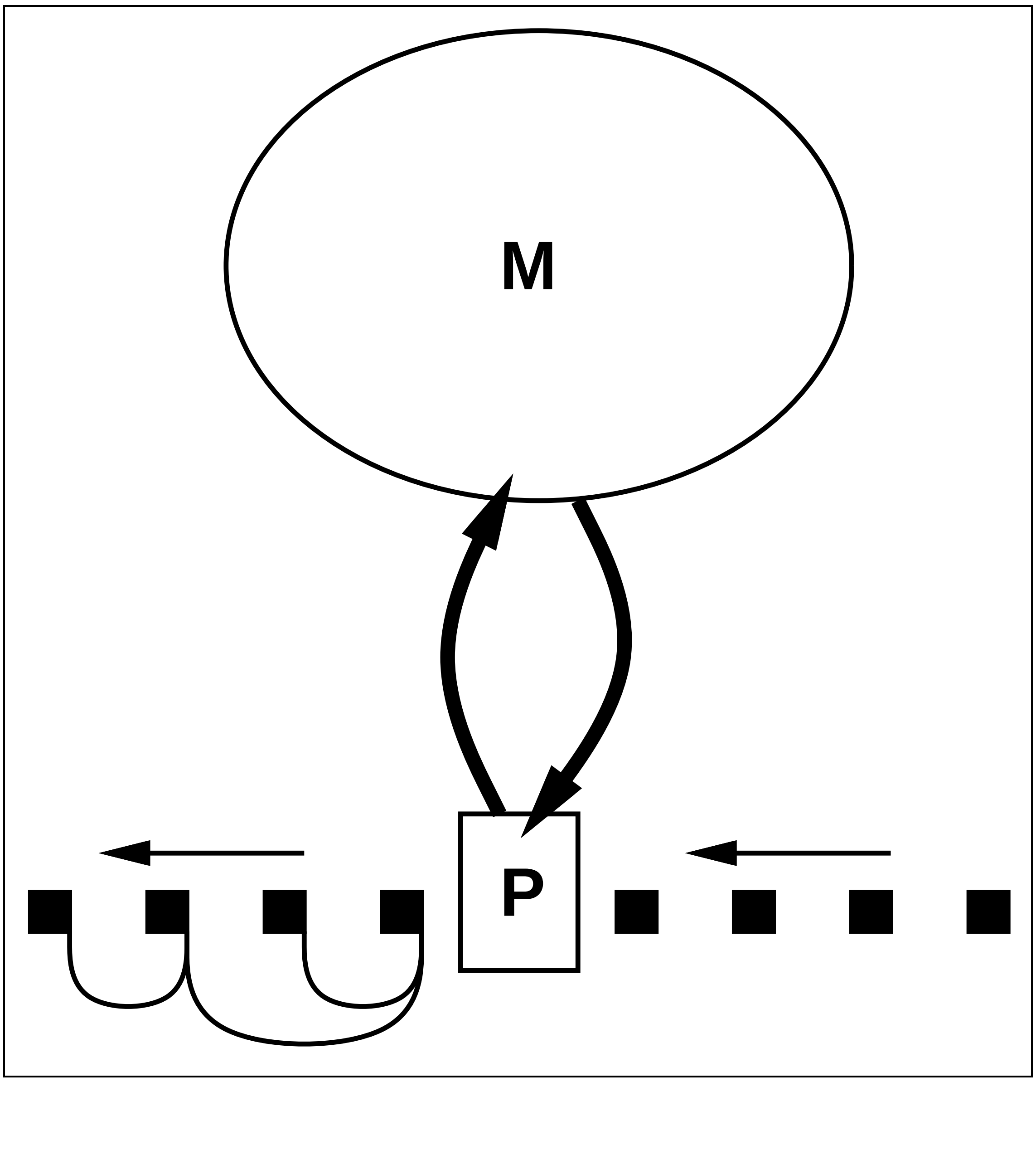
- Opening a door for the use of common sense knowledge in language processing and acquisition.

- A learning paradigm that bootstraps by interdigitating learning with processing.

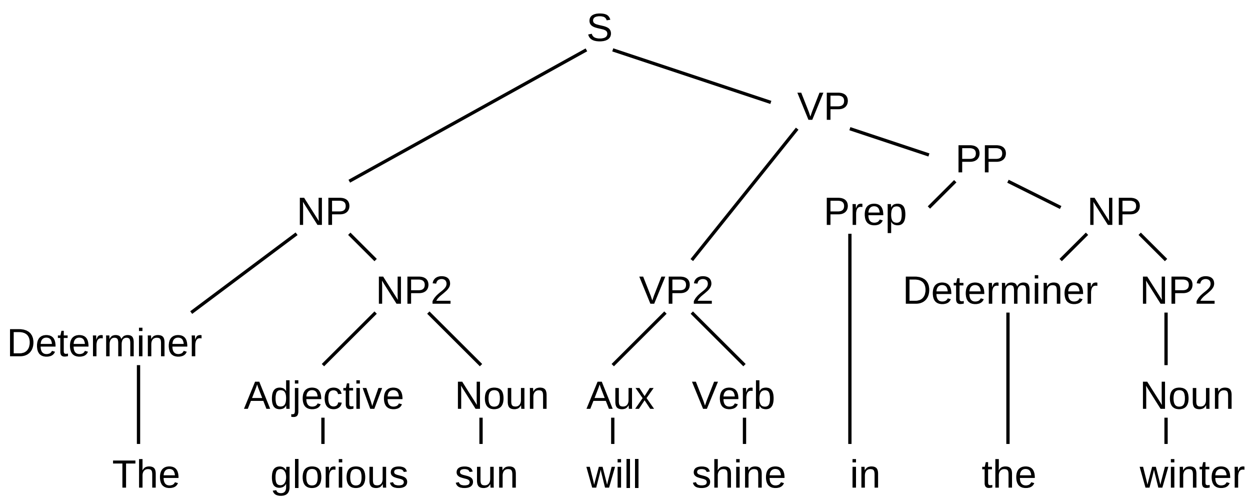
Bringing common sense into language

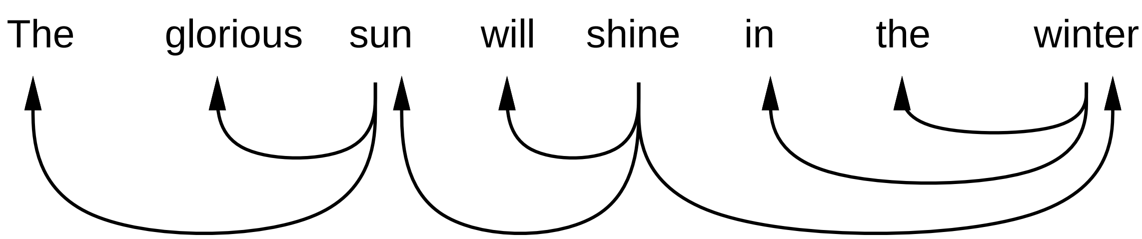


Bootstrapping by interdigitating learning and processing



# Phrase structure versus dependency structure



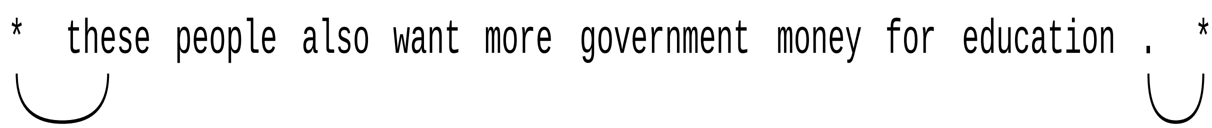



# Discovery of Linguistic Relations
# An Example

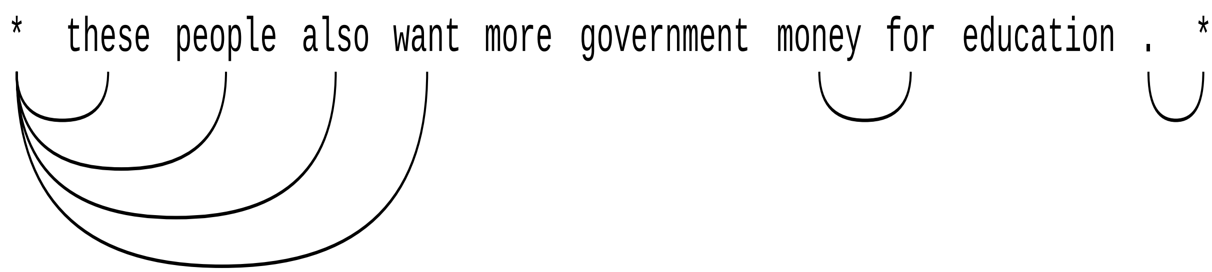
## Simple Sentence 1/5
## (Before training)

```
*  these people also want more government money for education .  *
```

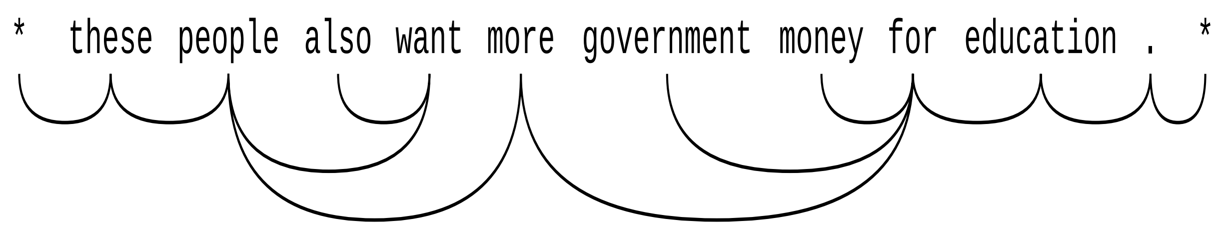
# Simple Sentence 2/5
# (After 1000 words of training)



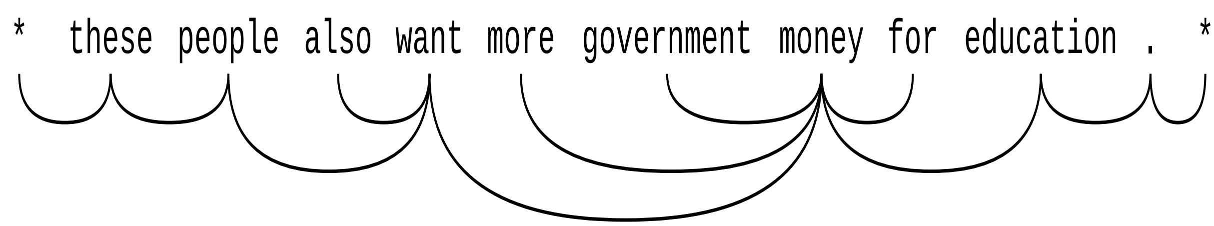
Simple Sentence 3/5
(After 10,000 words of training)



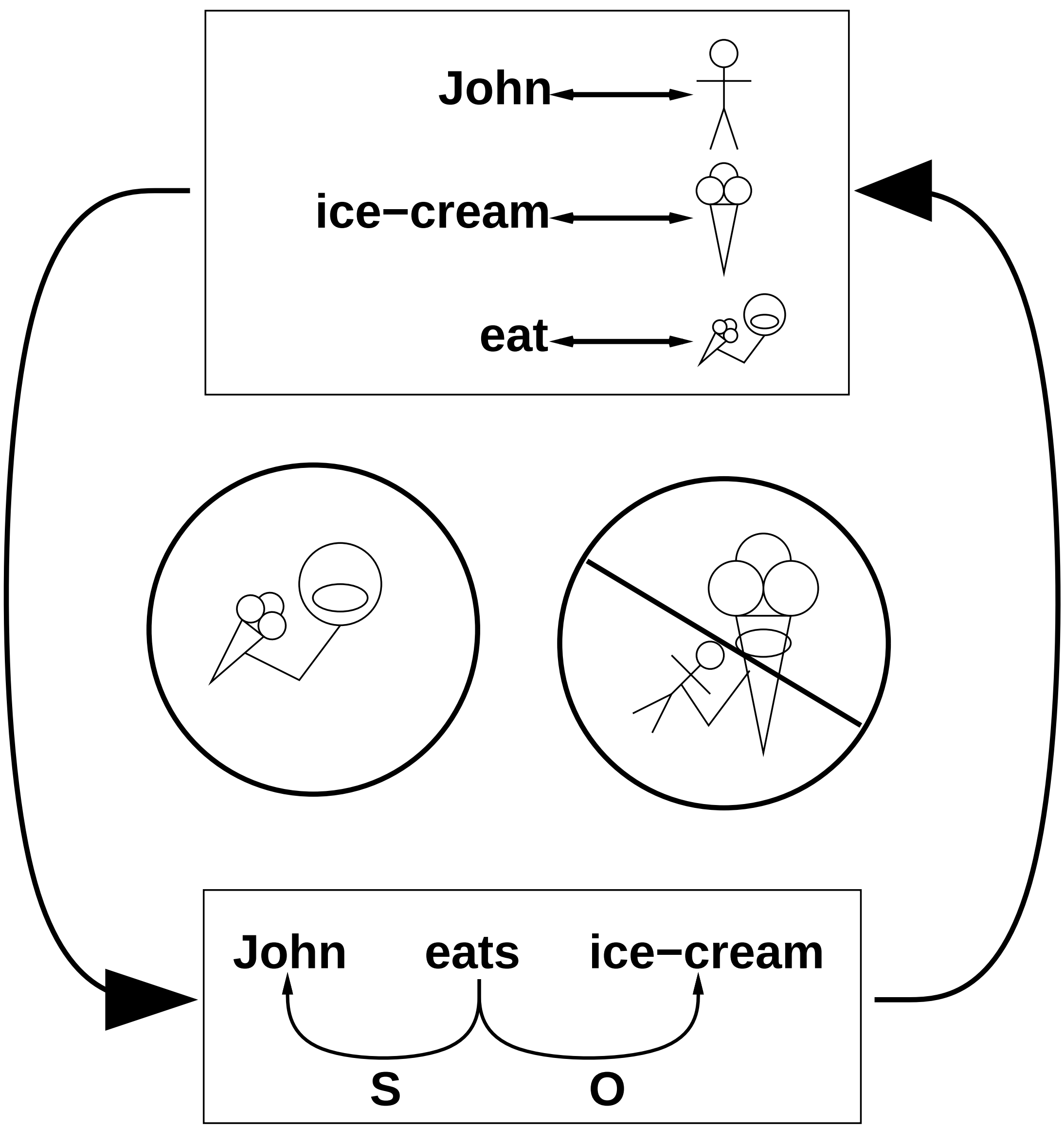
# Simple Sentence 4/5
# (After 100,000 words of training)



# Simple Sentence 5/5
(After 1,000,000 words of training)



Bringing common sense into language
The theory



A Theory of Syntactic Relations

- Lexical attraction is the likelihood of a syntactic relation
- The context of a word is given by its syntactic relations
- Syntactic relations can be formalized as a graph
- Entropy is determined by syntactic relations

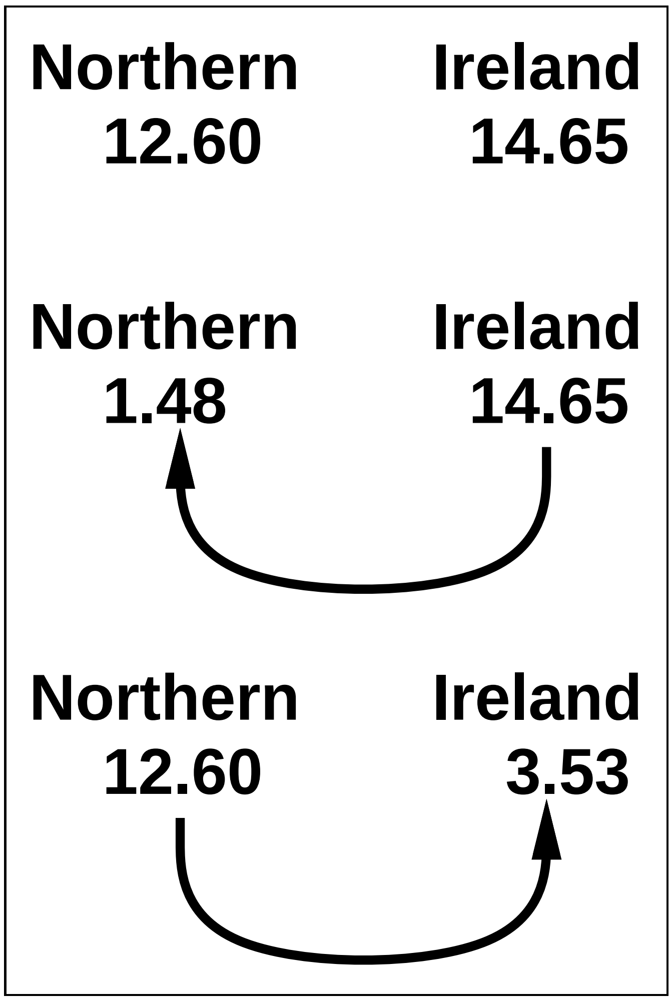
$$H = -\sum p_i \log p_i$$

The information content of a word:

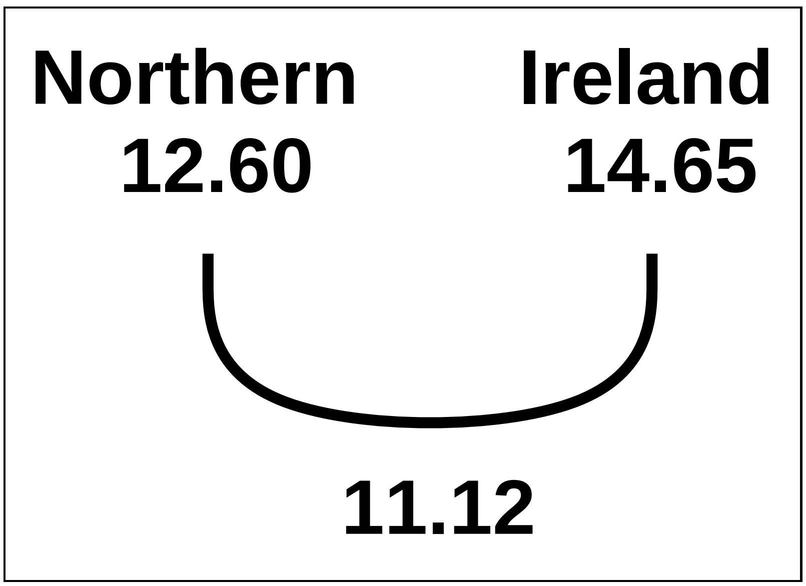
| The | IRA | is | fighting | British | rule | in | Northern | Ireland |
|---|---|---|---|---|---|---|---|---|
| 4.20 | 15.85 | 7.33 | 13.27 | 12.38 | 13.20 | 5.80 | 12.60 | 14.65 |

Total: 99.28 bits

The word pair and relative information:



The lexical attraction link:



## Language Model Determines the Context

| The | IRA | is | fighting | British | rule | in | Northern | Ireland |
|---|---|---|---|---|---|---|---|---|
| 4.20 | 12.90 | 3.73 | 10.54 | 8.66 | 5.96 | 3.57 | 9.25 | 3.53 |

Total: 99.28 $\rightarrow$ 62.34 bits

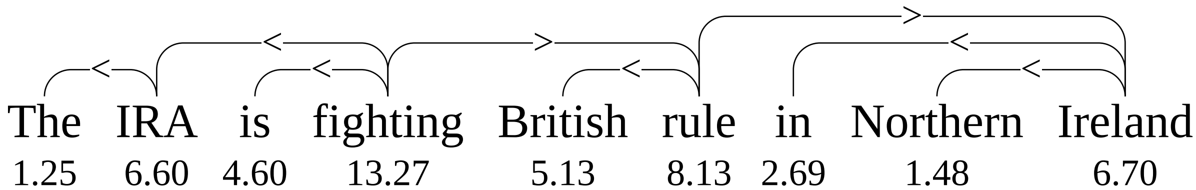
Context should be determined by syntactic relations:

?

The man with the dog spoke

The man with the dog spoke

Context should be determined by syntactic relations:

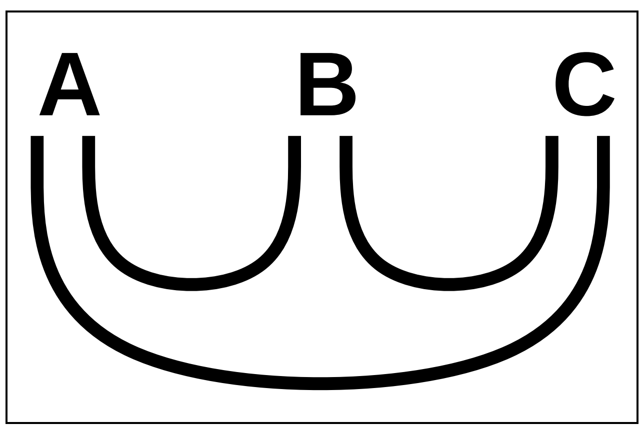



Total: 62.34 → 49.85 bits

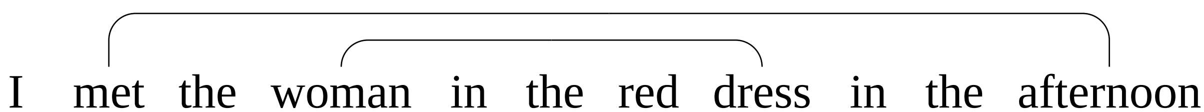
Dependency structure is acyclic:

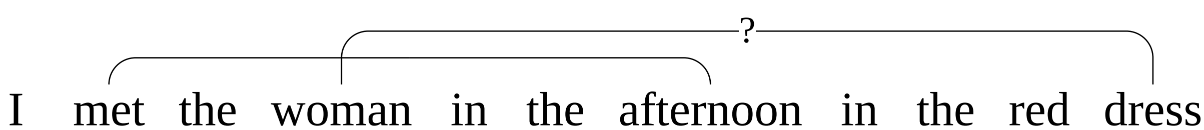
- Mathematically: cannot use all the lexical attraction links in a cycle.

- Linguistically: cannot construct a consistent head-modifier structure.



Syntactic relations form a planar tree:
(Links do not cross)



Syntactic relations form a planar tree:
(Links do not cross)

- Hays and Lecerf (1960) discovered that (almost) all sentences in a language are planar.

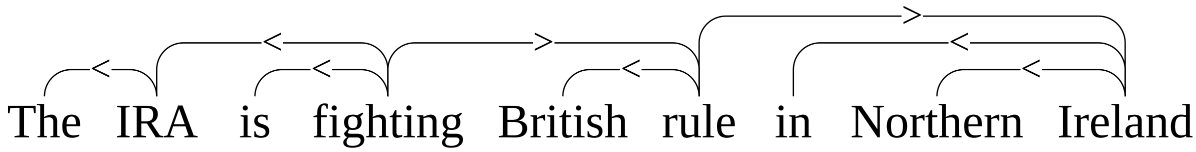
- Gaifman (1965) proved that a planar dependency grammar can generate the same set of languages as a context free grammar.

- Planar trees can be encoded with constant number of bits per word.

Cayley’s formula for counting trees:

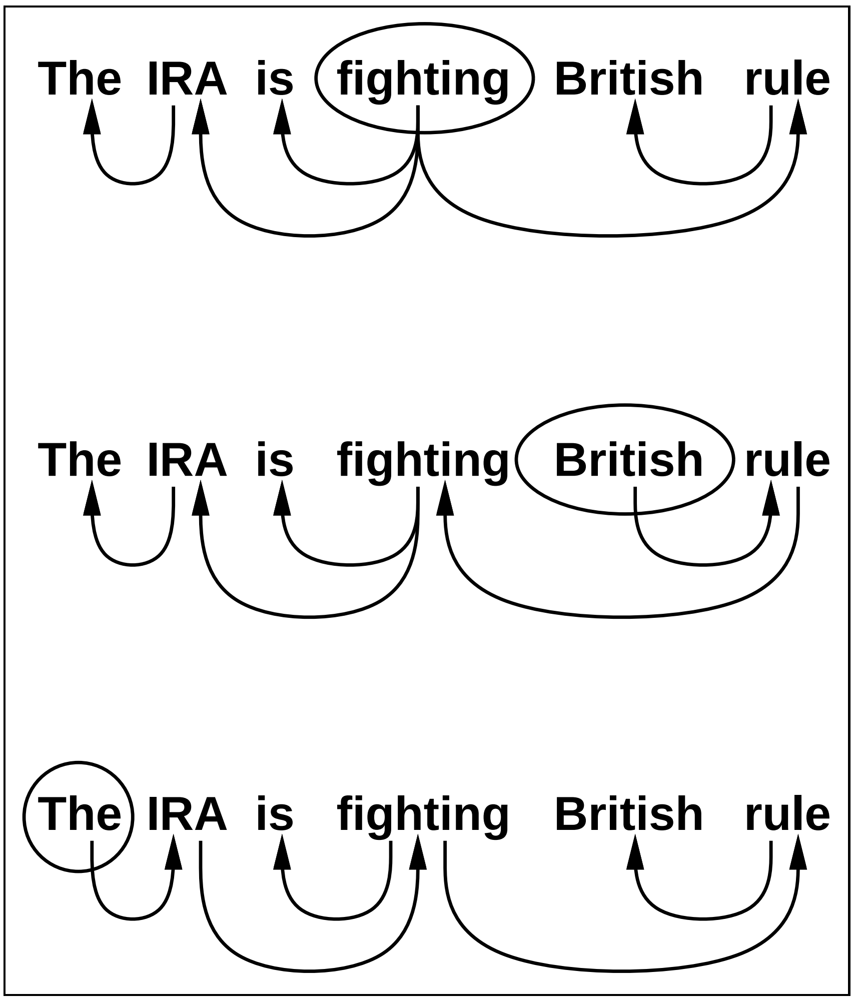
$$T(n) = n^{n-2}$$

Planar trees are polynomial in n:




Encoding: LPLLPPRLPRLPLPPP
L:10 R:11 P:0
Upper bound: 3 bits per word

# Lexical attraction is symmetric

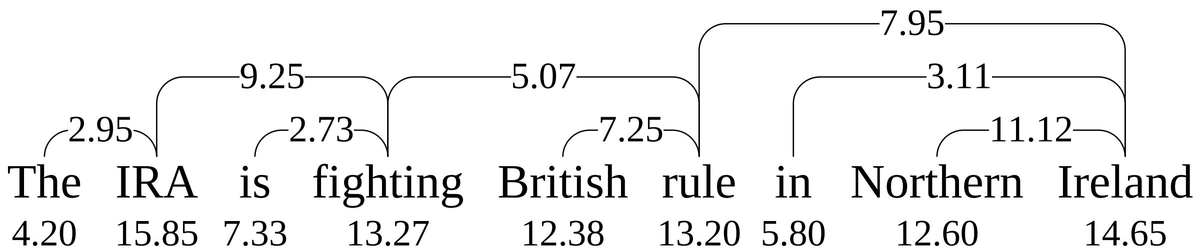


Lexical attraction is symmetric

$$
\begin{aligned}
S &= (W, L, w_0) \\
W &= \{\ w_i\ \} \\
L &= \{\ (w_i, w_j)\ \}
\end{aligned}
$$

$$
\begin{aligned}
P(S) &= P(L)P(w_0) \prod_{(w_i,w_j)\in L} P(w_j \mid w_i) \\
&= P(L)P(w_0) \prod_{(w_i,w_j)\in L} \frac{P(w_i, w_j)}{P(w_i)} \\
&= P(L) \prod_{w_i \in W} P(w_i) \prod_{(w_i,w_j)\in L} \frac{P(w_i, w_j)}{P(w_i)P(w_j)}
\end{aligned}
$$

Dependency structure is an undirected, acyclic, planar graph:



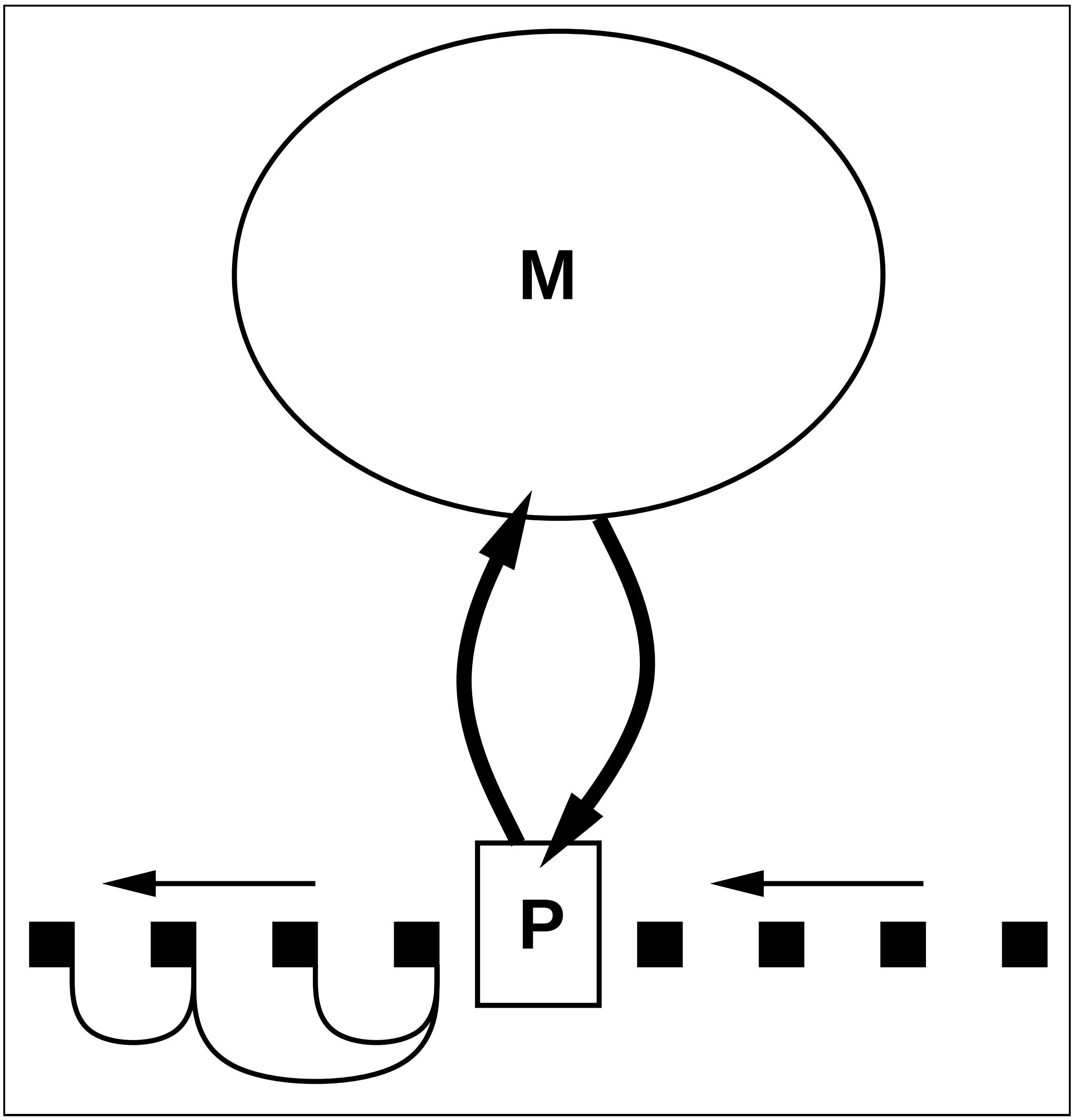
Information in a Sentence =

Information in Words

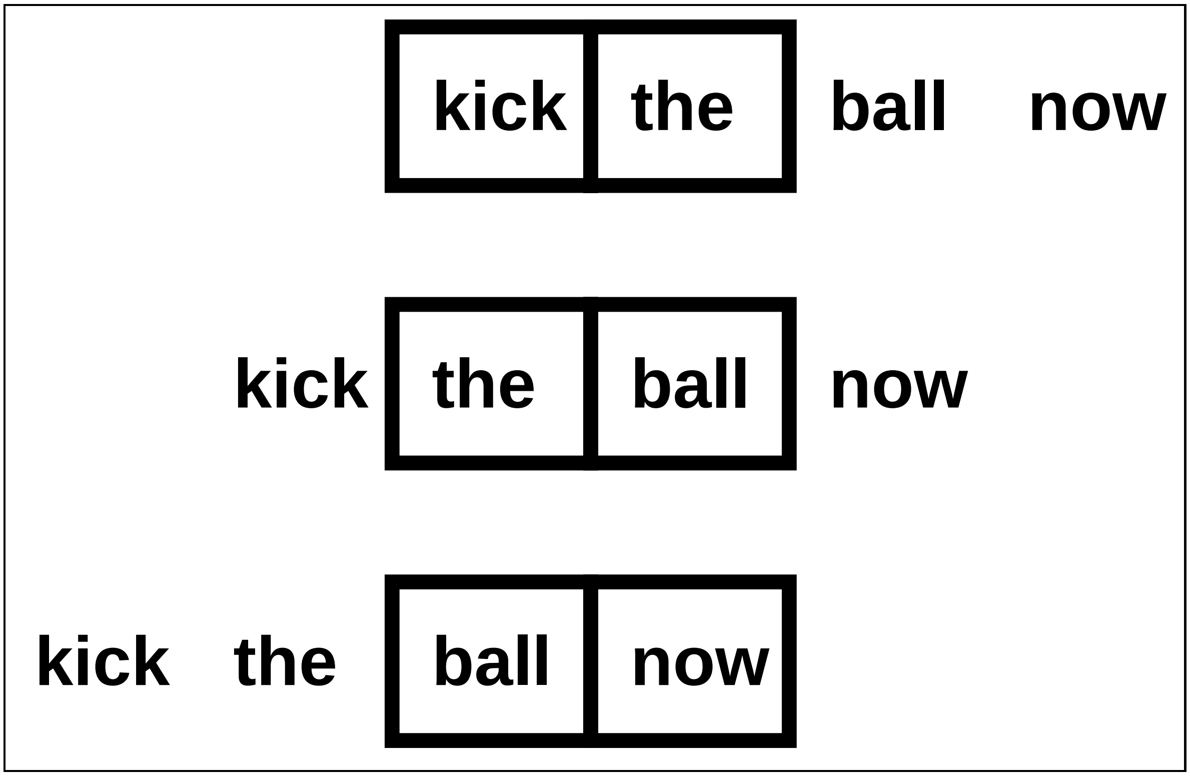
+ Information in the Tree

- Mutual Information in Syntactic Relations

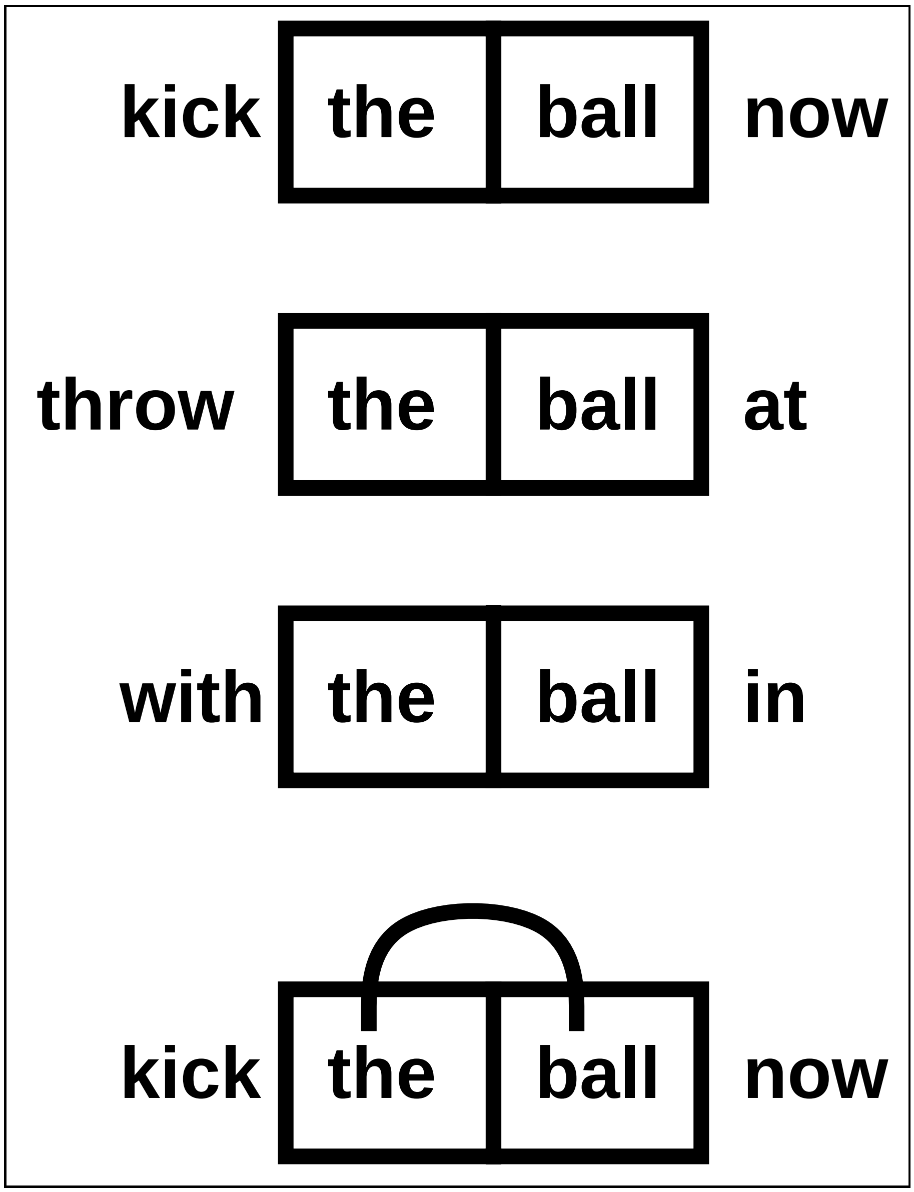
The Memory



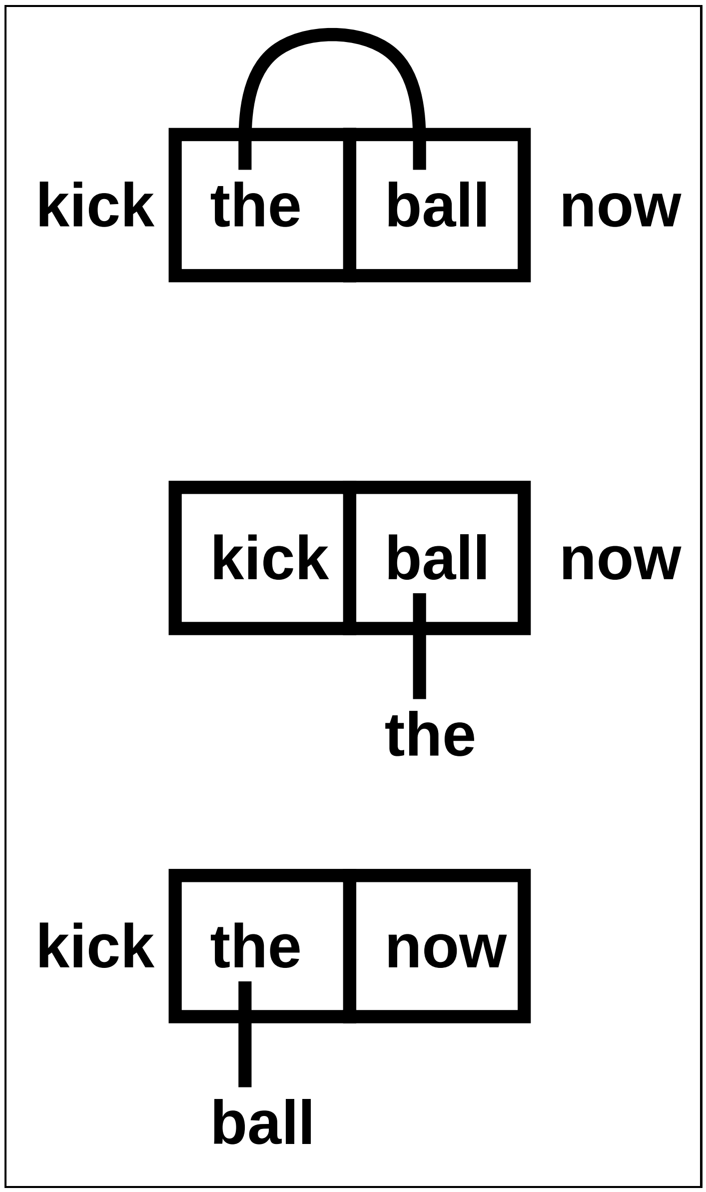
The memory observes the processor



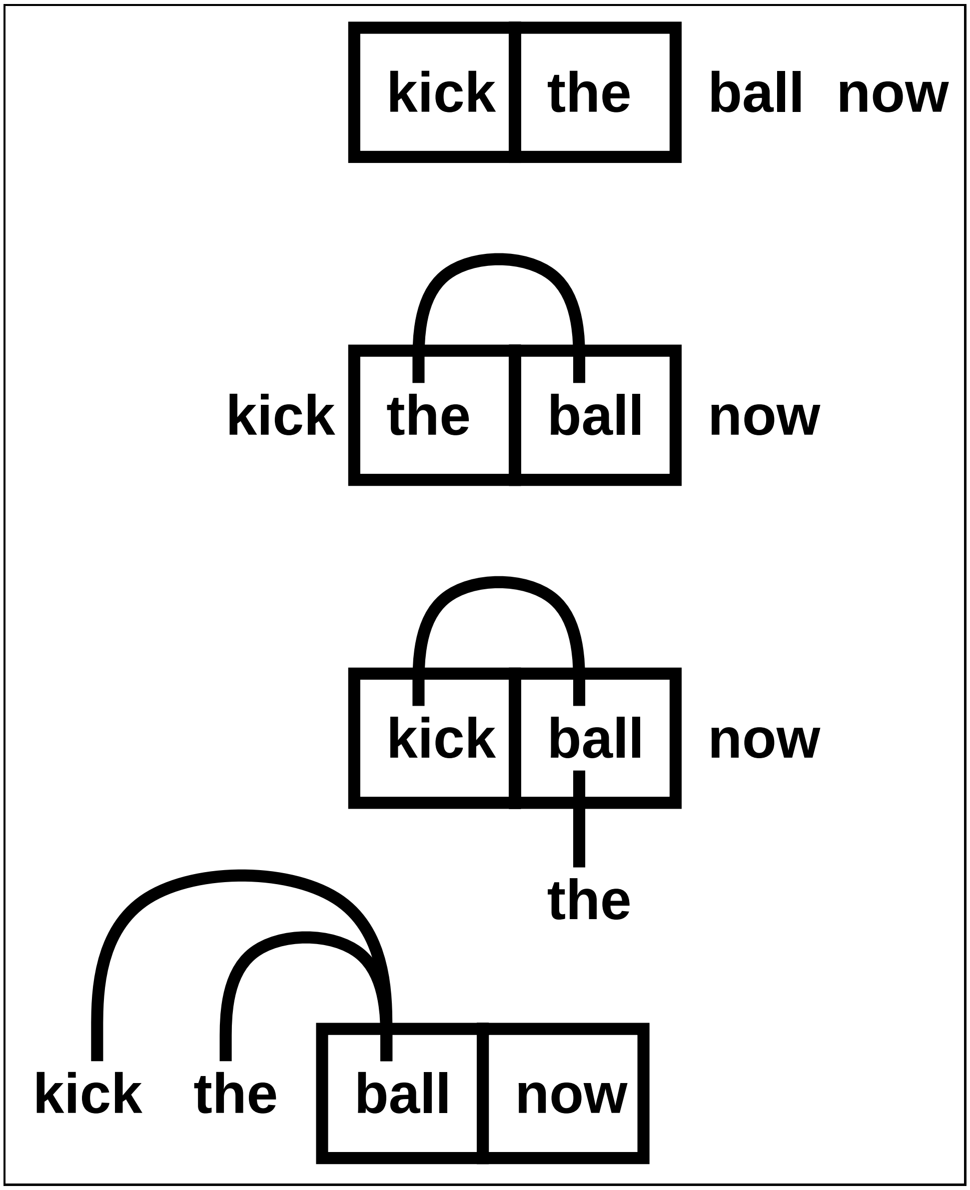
Learning simple structures



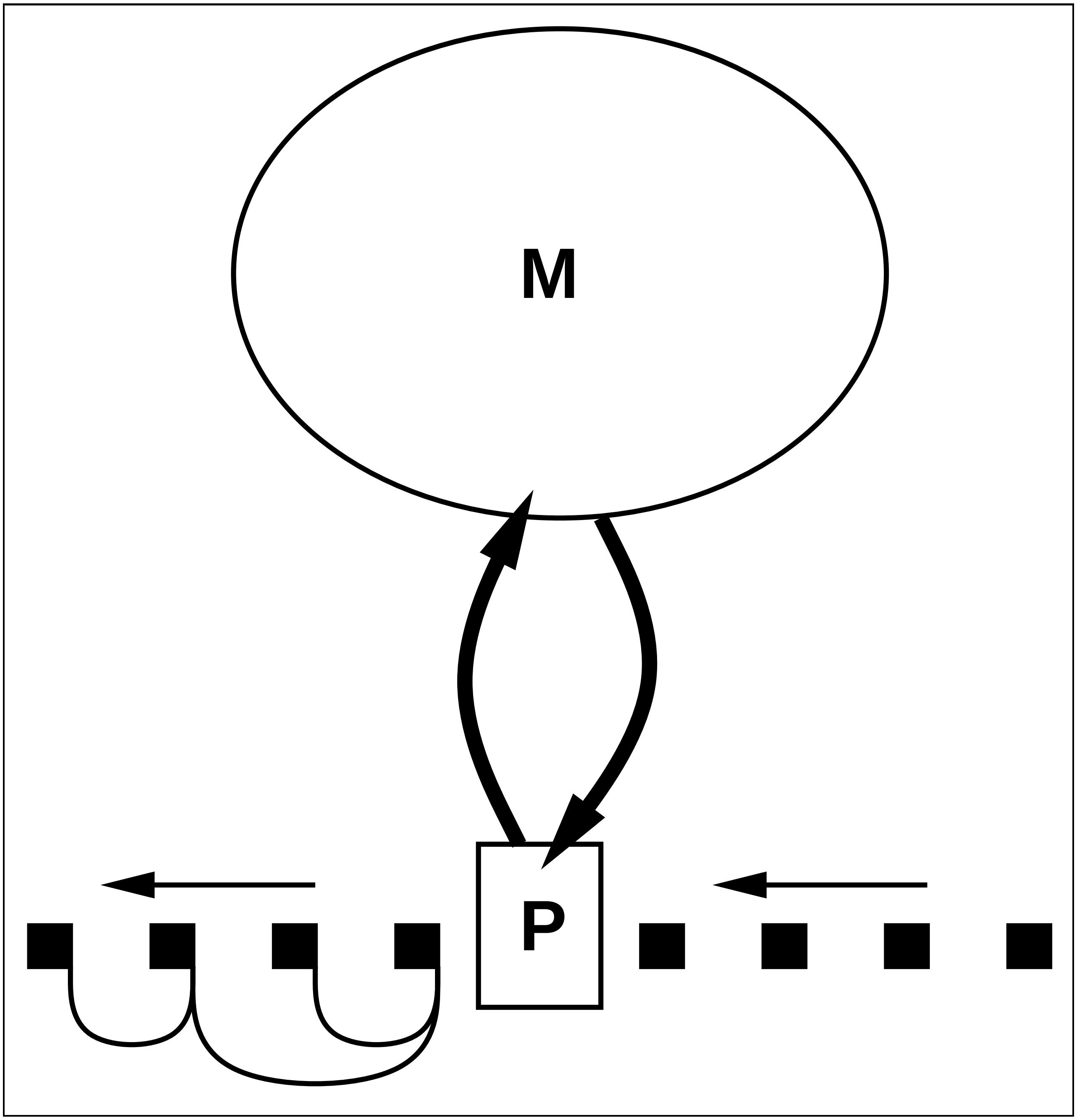
# Simple structures help see complex structures



# Learning complex structures



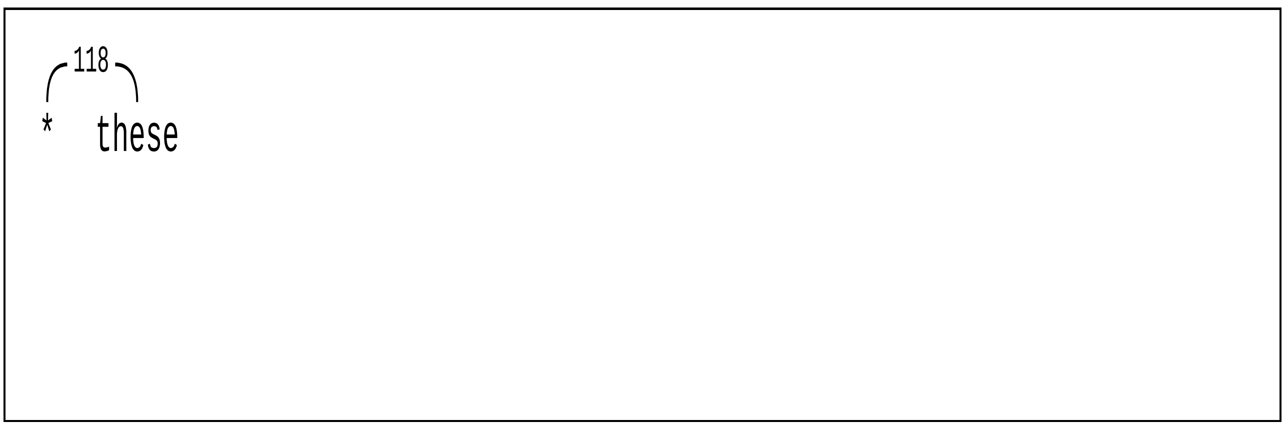
The Processor



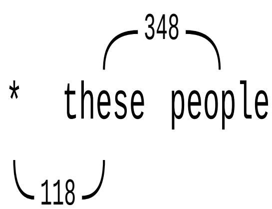
- We need to discover the best linkage.

```
*  these people also want more government money for education .  *
```

- Words are read in left to right order.

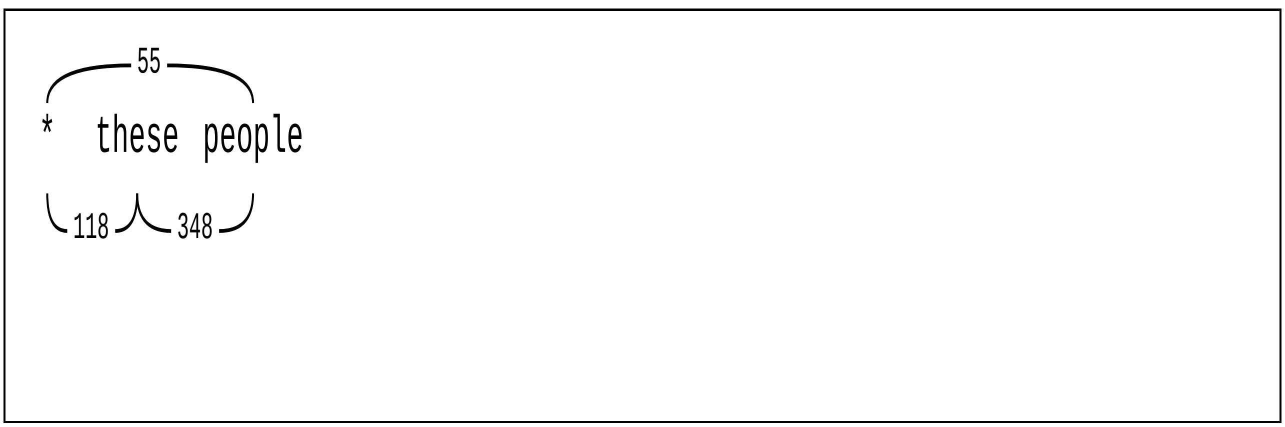


- New word considers links with previous words.

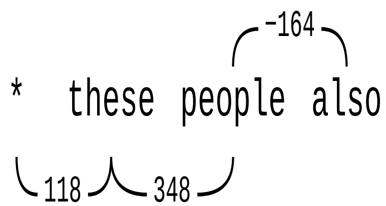


- Cycles are not allowed.

- Link with minimum score gets rejected.

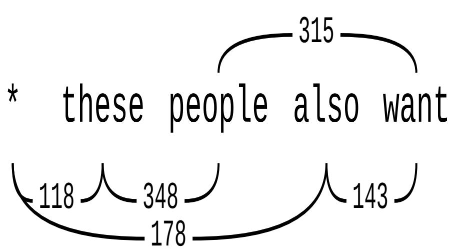


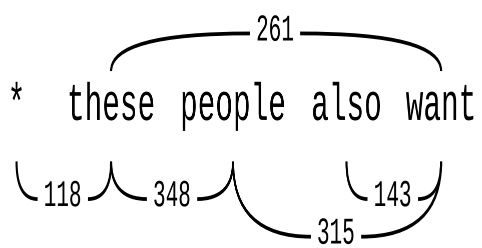
- Link with negative value not accepted.



- Link crossing not allowed.

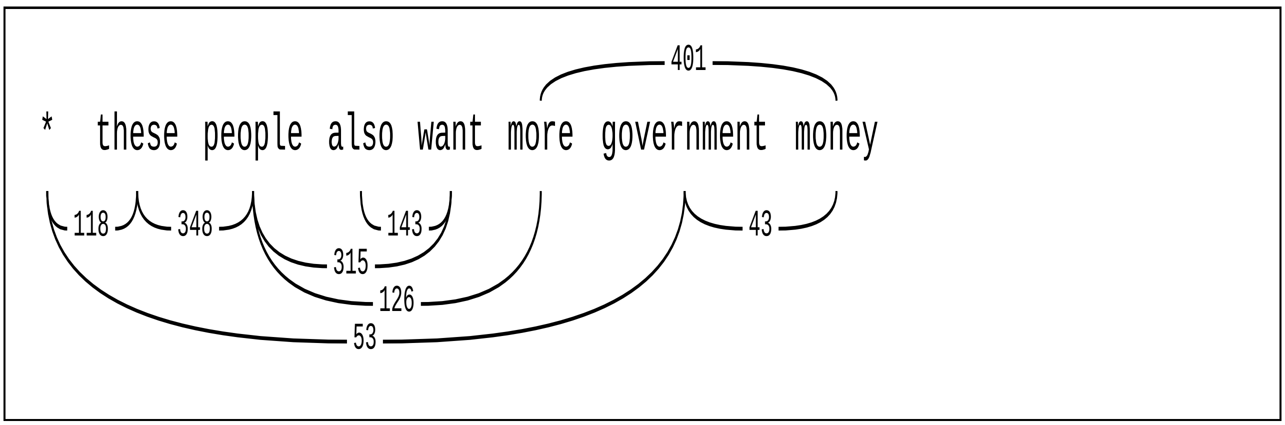
- Link with minimum score gets eliminated.



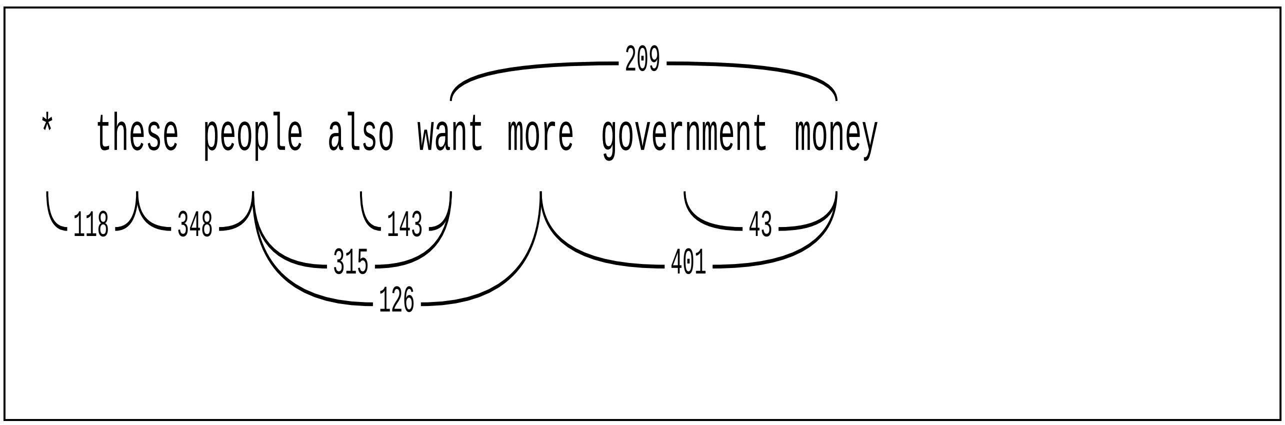


261
* these people also want
118
348
143
315

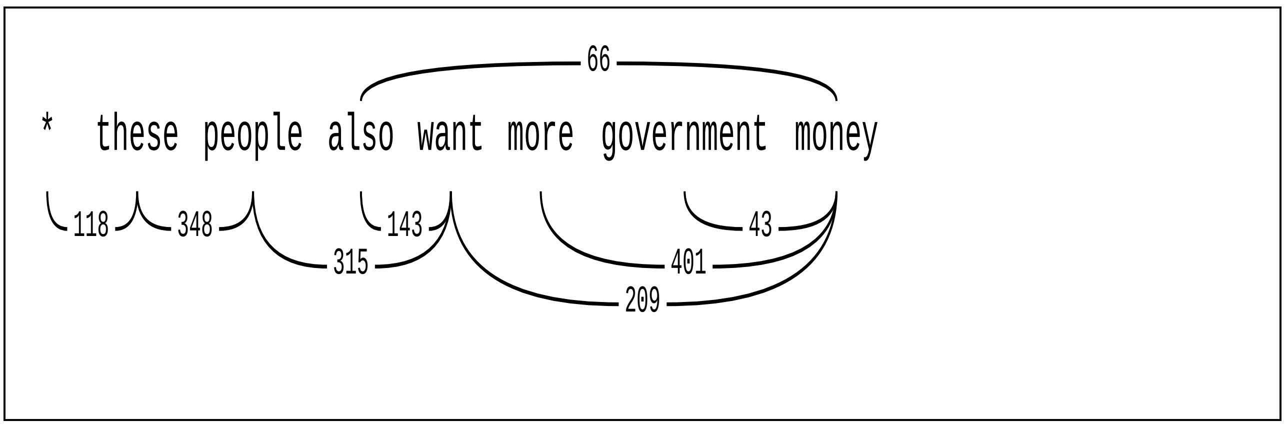
- The two constraints straighten out previous mistakes by eliminating bad links.



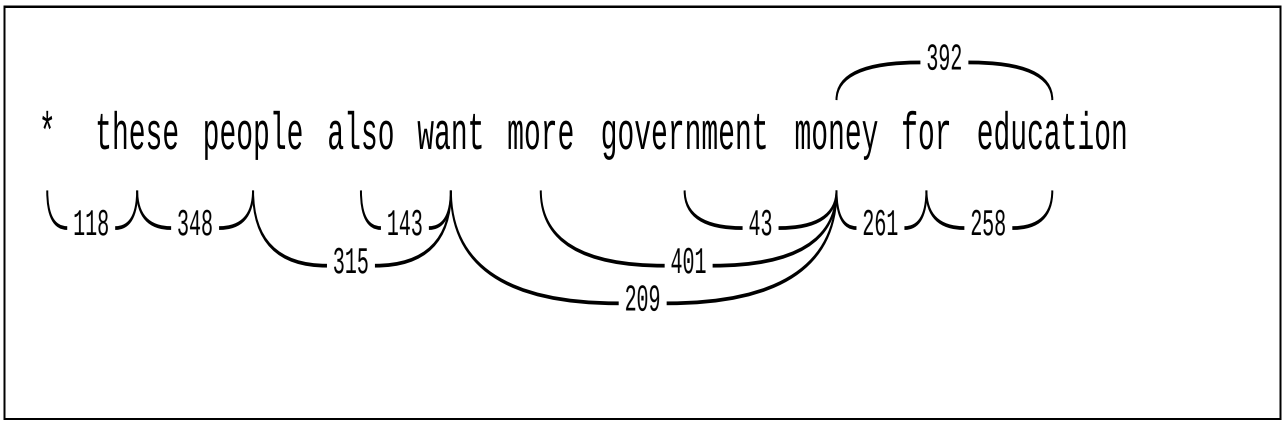
- Eliminating bad links 2/3



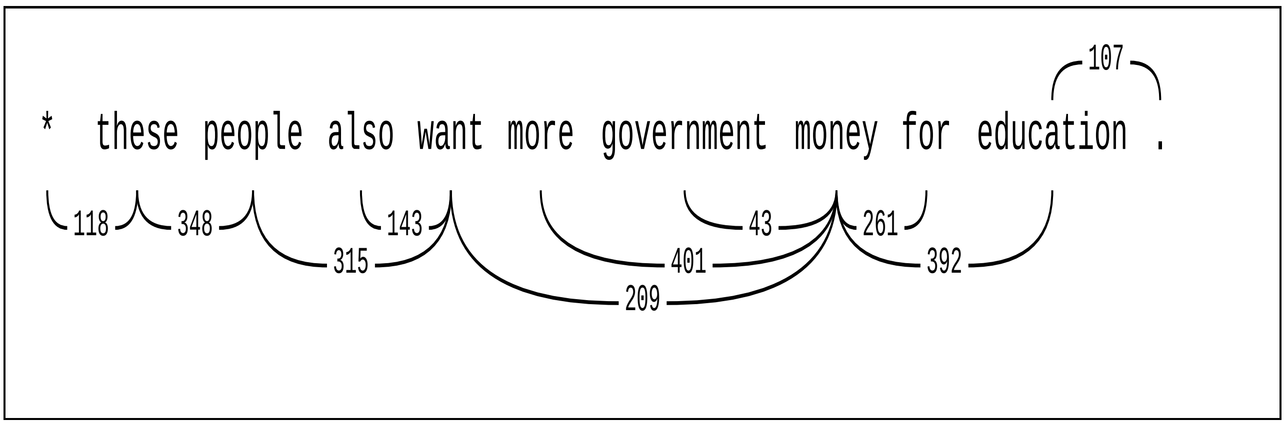
- Eliminating bad links 3/3



• New link can knock off old link in cycle.



- The final result.

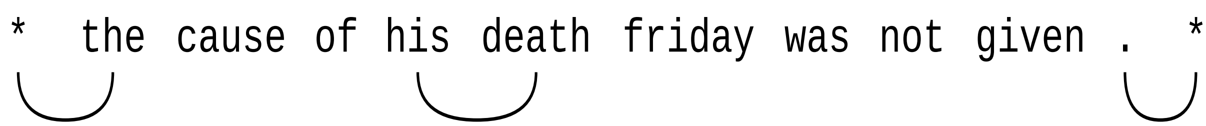


Discovery of Linguistic Relations
Using Lexical Attraction

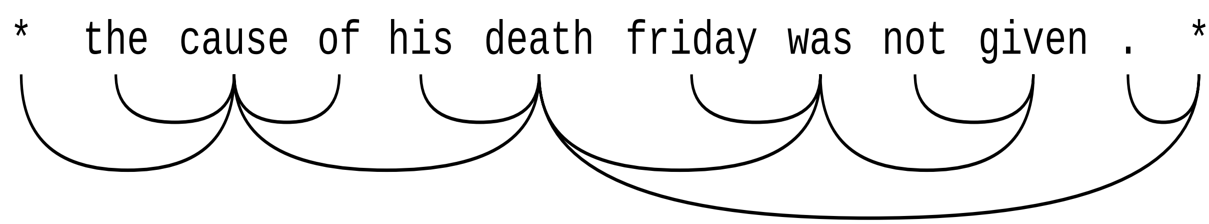
A demonstration

- Long distance link
- Complex noun phrase
- Syntactic ambiguity

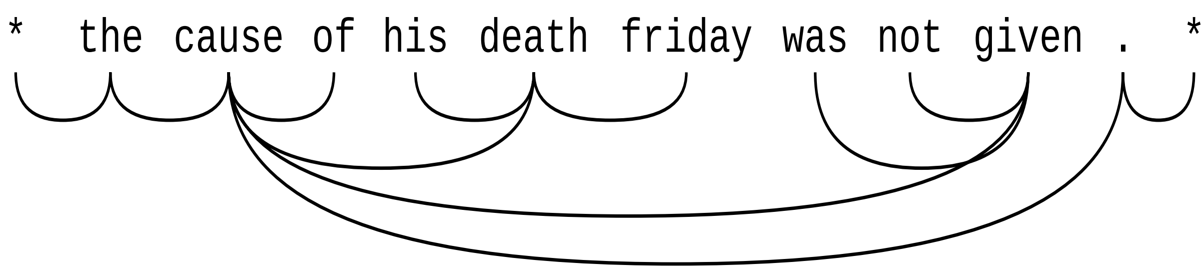
# Long Distance Link 1/3
(After 1,000 words of training)



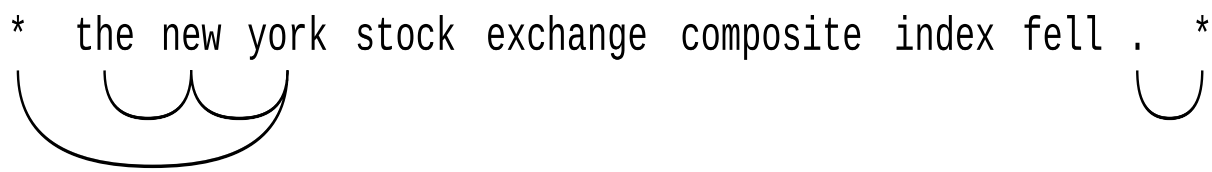
# Long Distance Link 2/3
## (After 100,000 words of training)



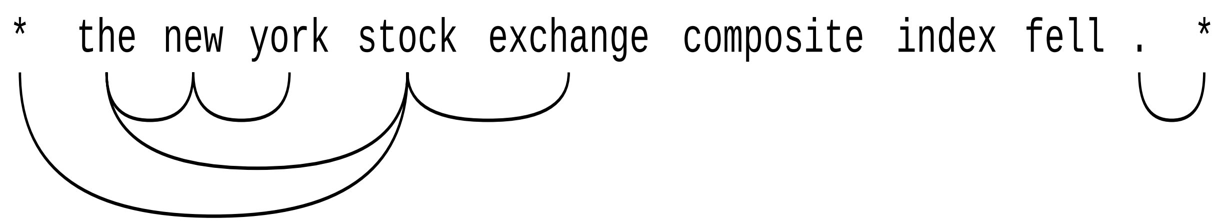
Long Distance Link 3/3
(After 10,000,000 words of training)



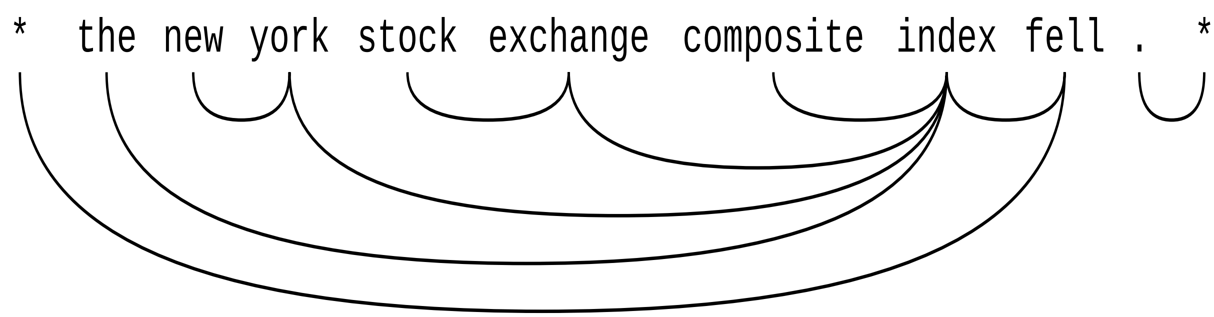
# Complex Noun Phrase 1/4
## (After 10,000 words of training)



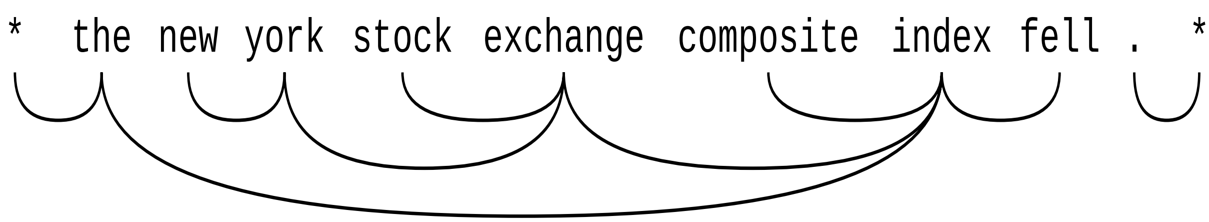
# Complex Noun Phrase 2/4
(After 100,000 words of training)



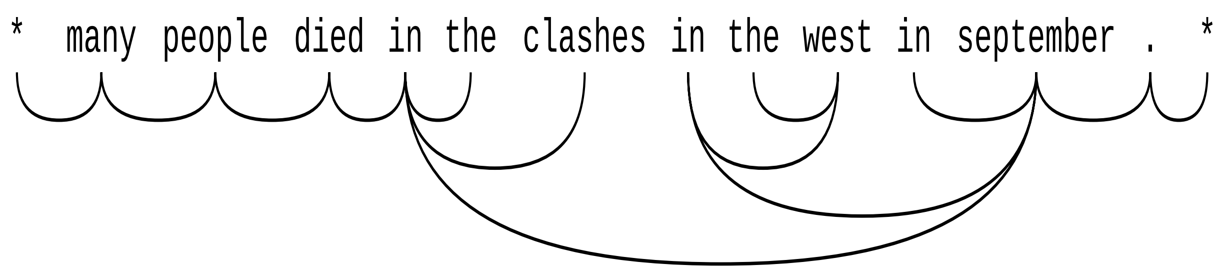
Complex Noun Phrase 3/4
(After 1,000,000 words of training)



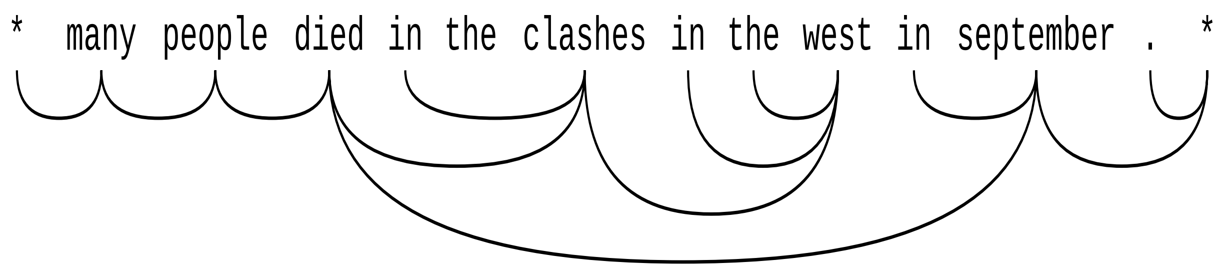
Complex Noun Phrase 4/4
(After 10,000,000 words of training)



Syntactic Ambiguity 1/3
(After 1,000,000 words of training)

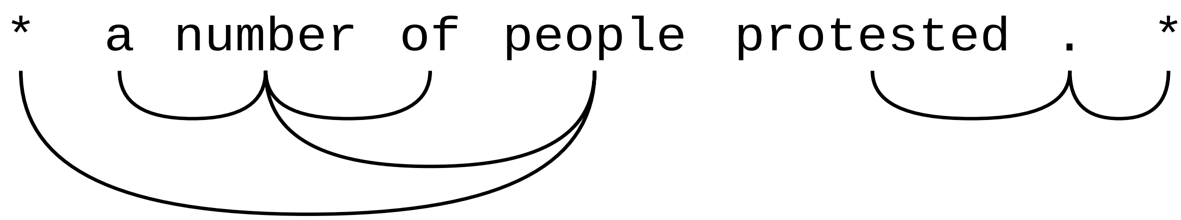


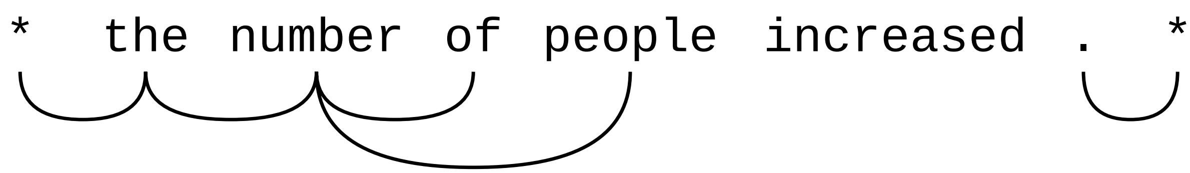
# Syntactic Ambiguity 1/3
(After 10,000,000 words of training)



Syntactic Ambiguity 2/3

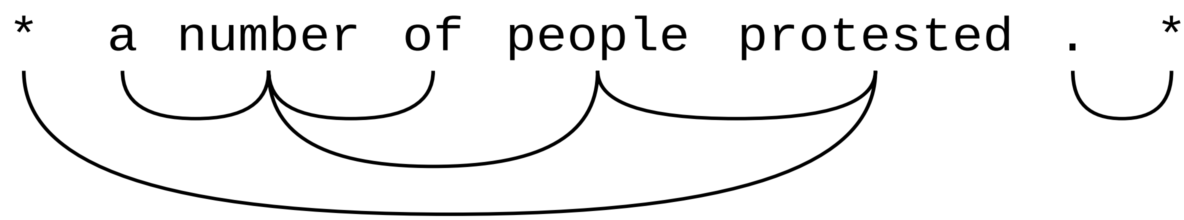
(After 500,000 words of training)

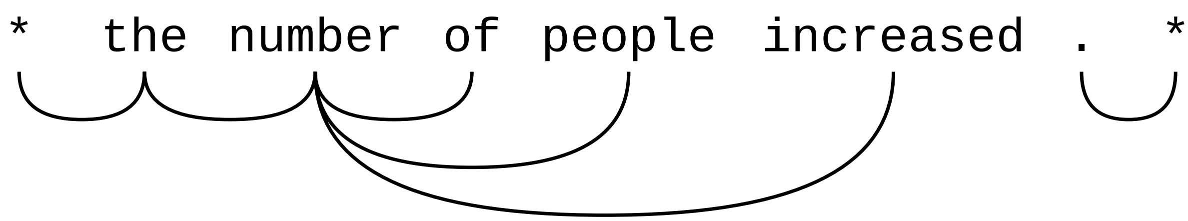


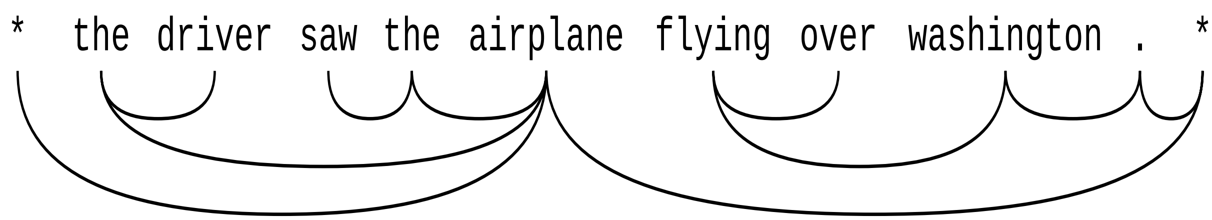
Syntactic Ambiguity 2/3

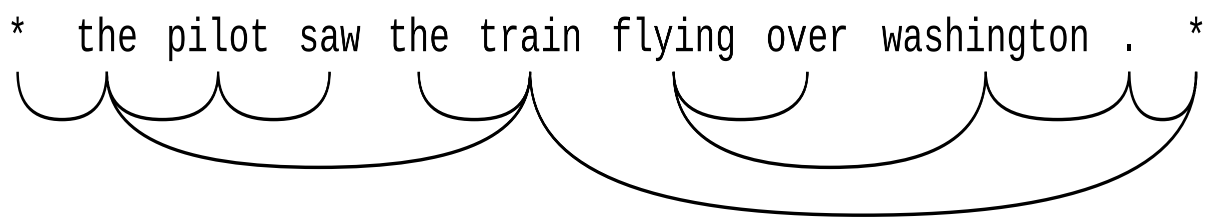
(After 5,000,000 words of training)



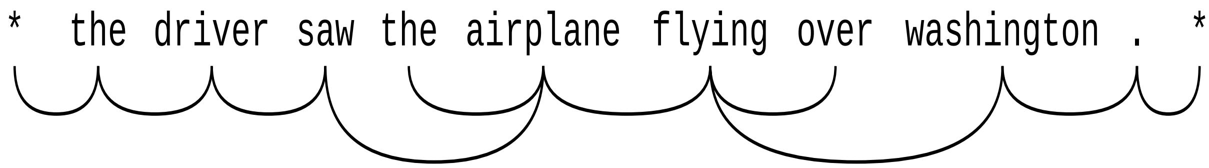



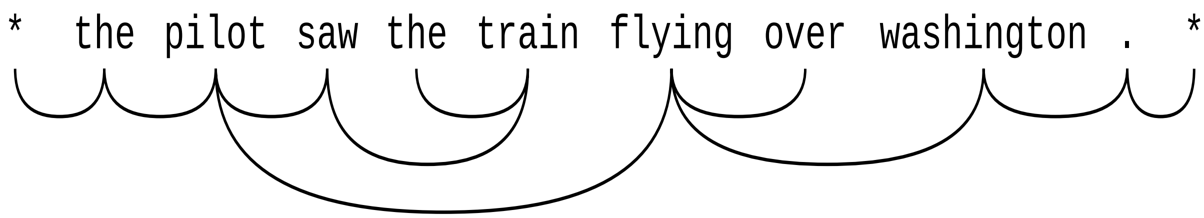
# Syntactic Ambiguity 3/3
## (After 1,000,000 words of training)



# Syntactic Ambiguity 3/3
(After 10,000,000 words of training)

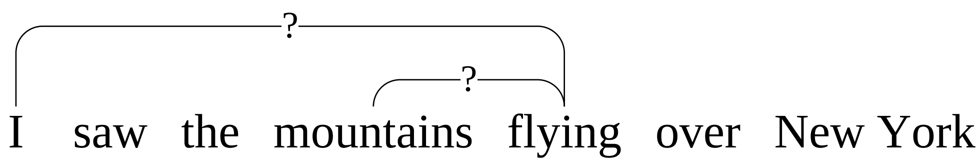


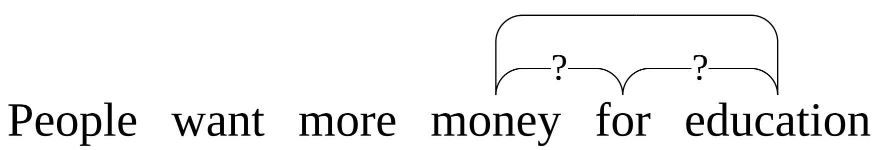



Results

- Evaluation criteria
- Upper and lower bounds
- Link accuracy
- Related work

## Evaluation criteria: Content-word links



Training

- Up to 100 million words of Associated Press material.

Testing

- 200 out-of-sample sentences.

- Selected from 5000 word vocabulary (90% of all the words seen in the corpus).

- 3152 words (15.76 words per sentence).

- Hand parsed with 1287 content-word links.

Accuracy:

n1 = human links
n2 = program links
n12 = common links

- Precision = n12 / n2
- Recall = n12 / n1

Lower bound:

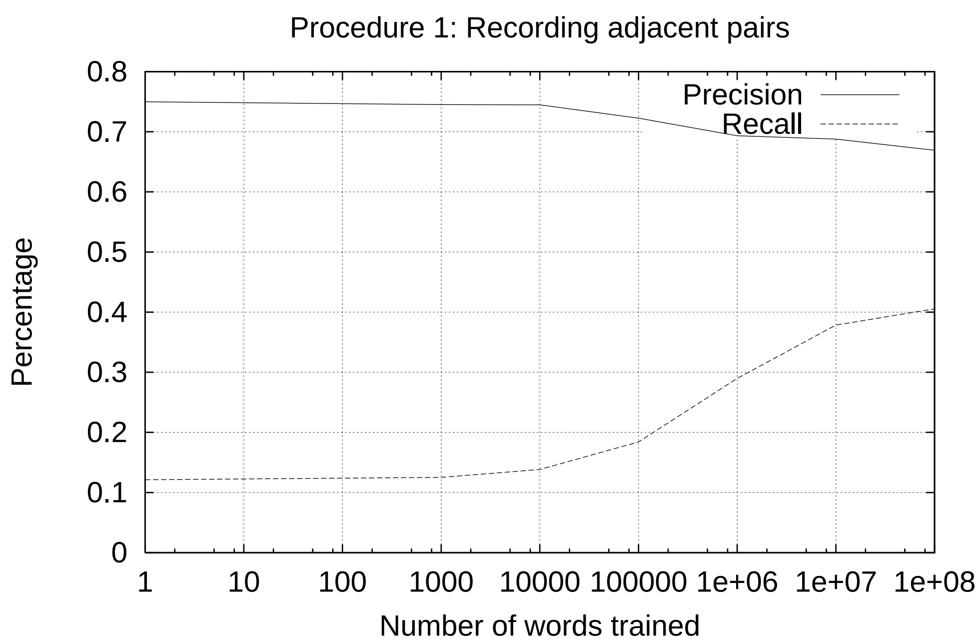
Random lexical attraction $\rightarrow$ 8.9% precision, 5.4% recall

Linking every adjacent word $\rightarrow$ 41% recall

Upper bound:

85% of syntactically related pairs have positive lexical attraction

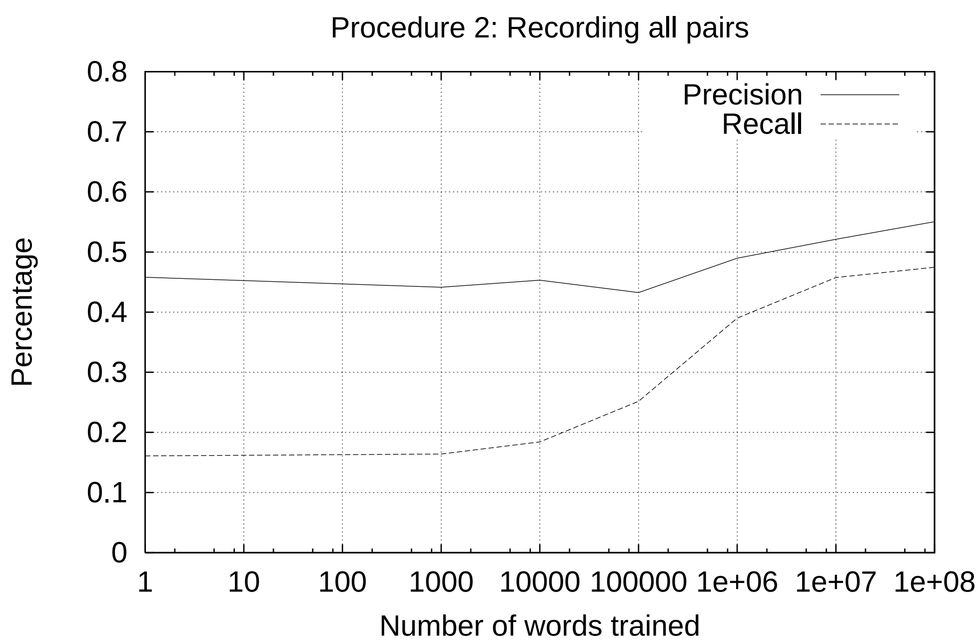
# Recording adjacent pairs




Precision = 67%

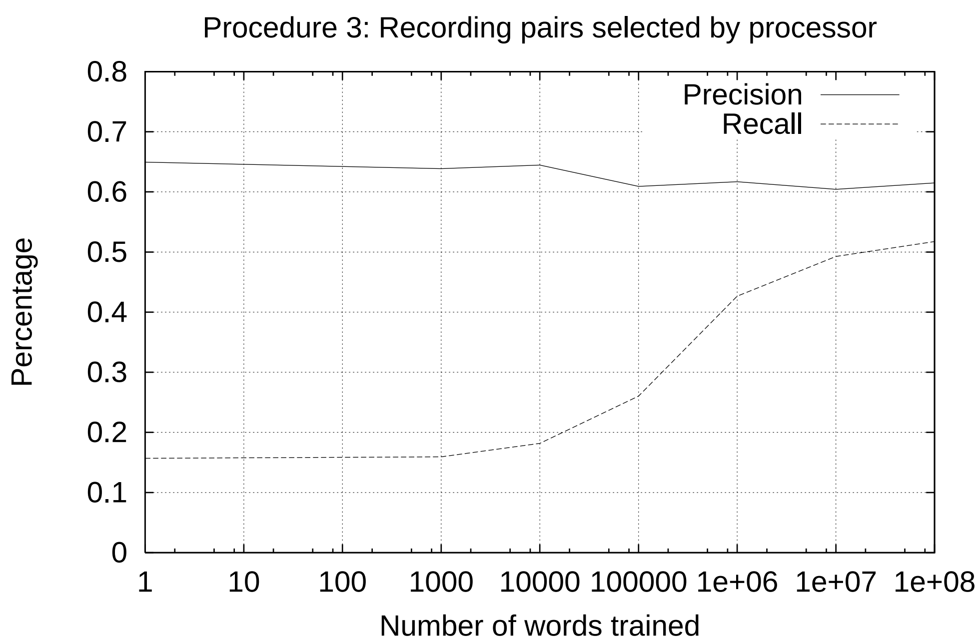
Recall = 41%

# Recording all pairs




Precision = 55%
Recall = 48%

Using feedback from processor

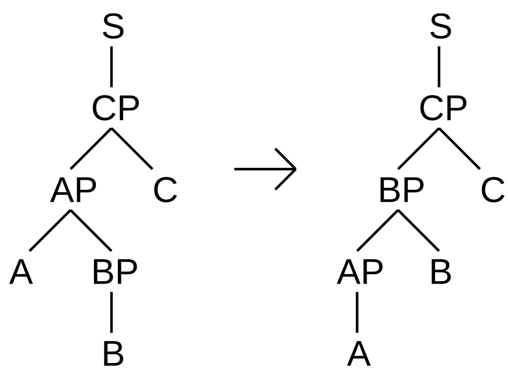



Precision = 62%
Recall = 52%

Related work

- Magerman and Marcus, 1990
- Lari and Young, 1990
- Pereira and Schabes, 1992
- Briscoe and Waegner, 1992
- Carroll and Charniak, 1992
- Stolcke, 1994
- Chen, 1996
- de Marcken, 1996

# de Marcken, 1995




| | | |
|---|---|---|
| AP => A BP | | AP => A |
| BP => B | → | BP => AP B |
| CP => AP C | | CP => BP C |

Lessons learned

- Training with words instead of parts of speech enable the program to learn common but idiosyncratic usages of words.

- Not committing to early generalizations prevent the program from making irrecoverable mistakes early.

- Using a representation that makes the relevant features (such as syntactic relations) explicit simplifies learning.

Contributions

- Opening a door for common sense in language
- Bootstrapping from zero by interdigitating learning and processing

## Future Work

- Second degree models
- History mechanism
- Categorization and generalization