

# 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

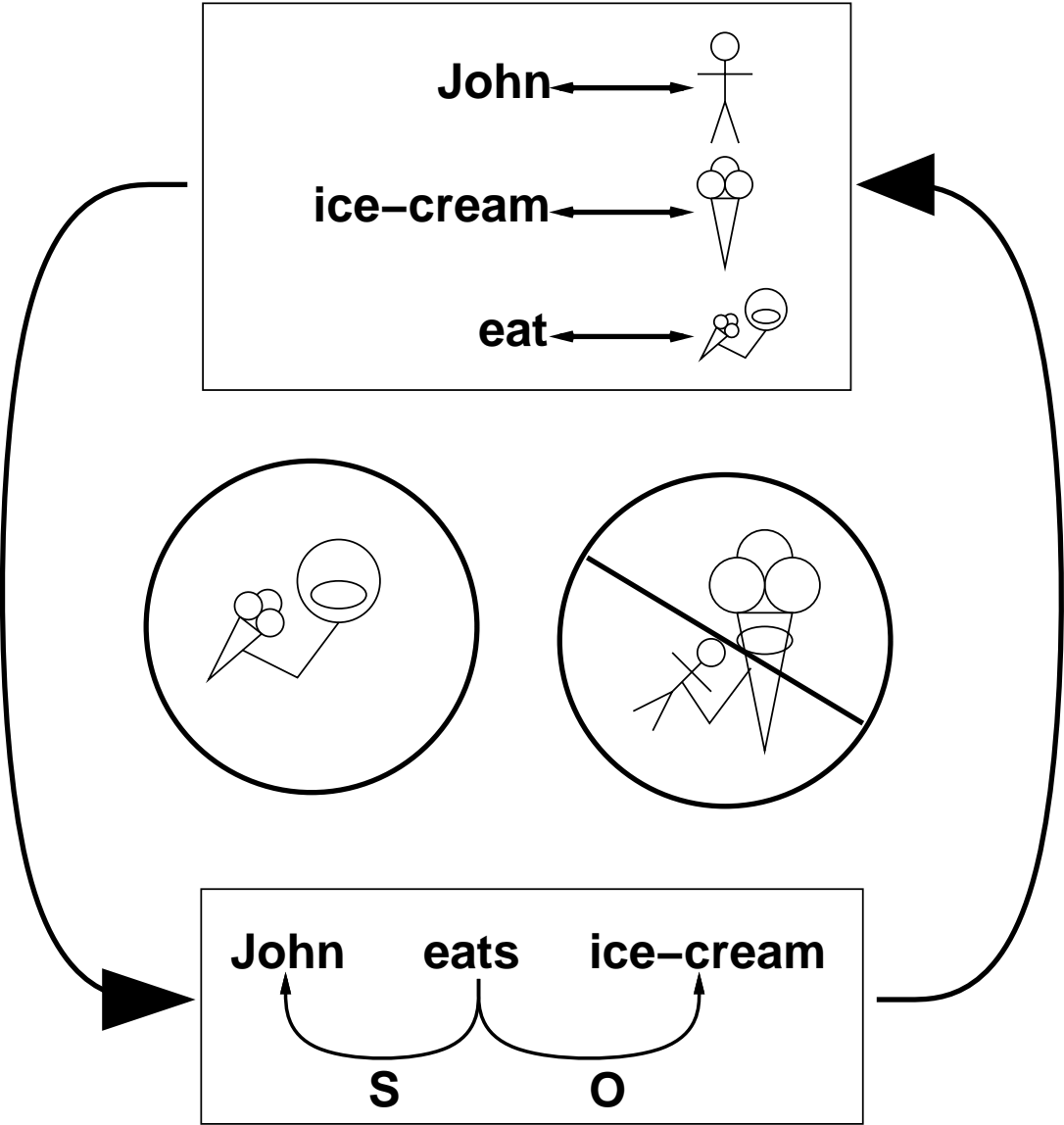
- Colorless green ideas sleep furiously.

– Chomsky, 1956

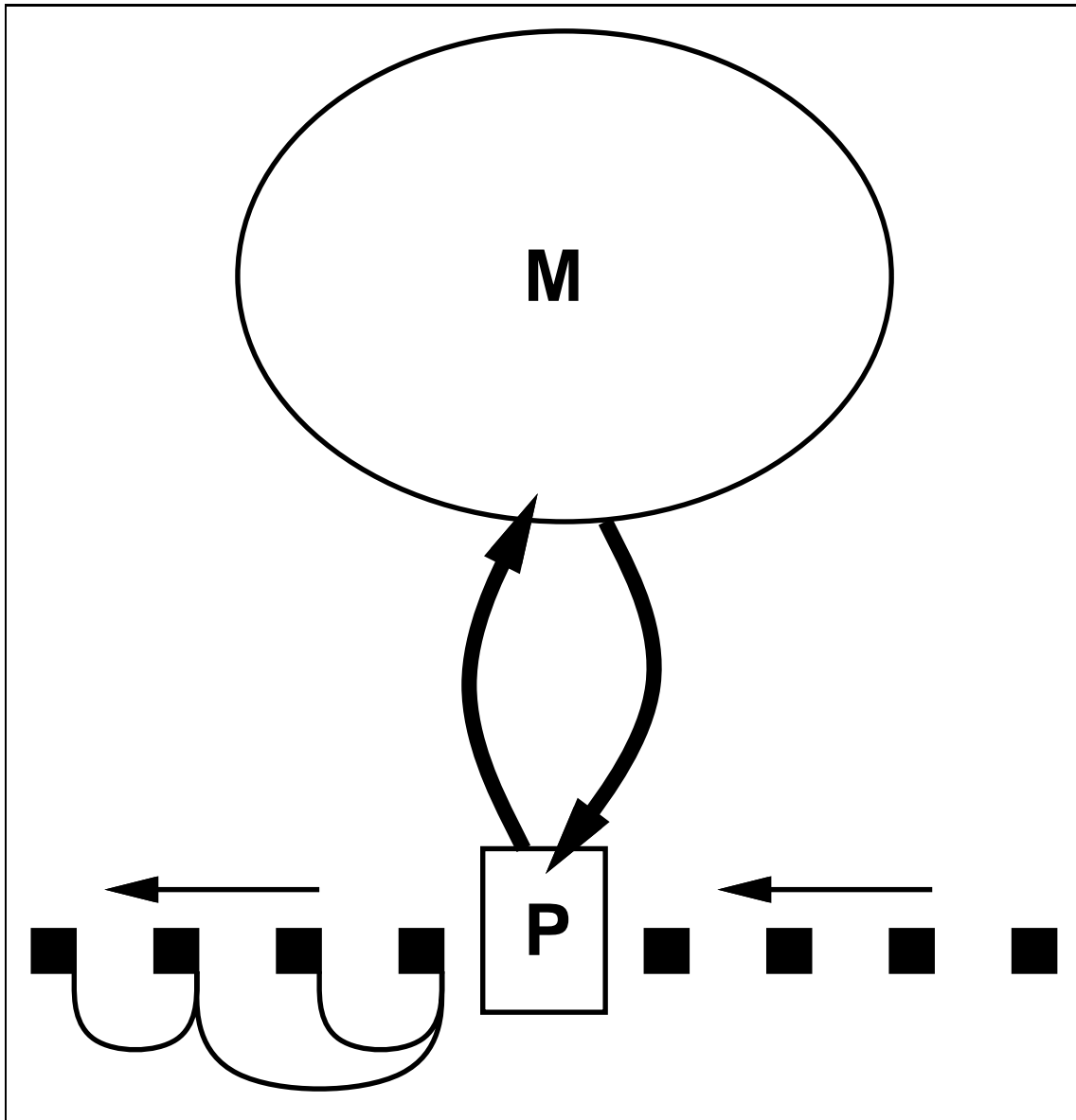
## Contributions of this thesis

- 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.

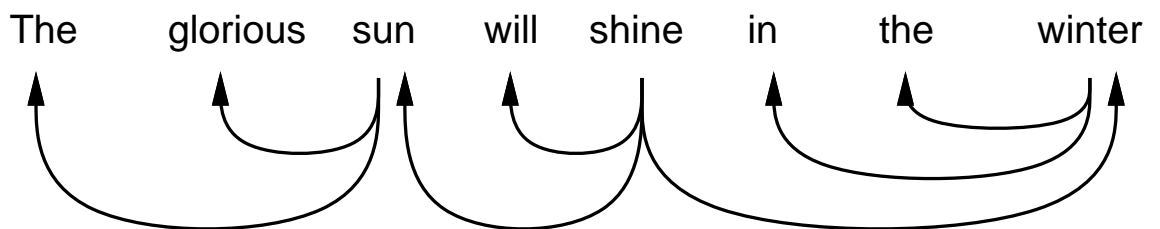
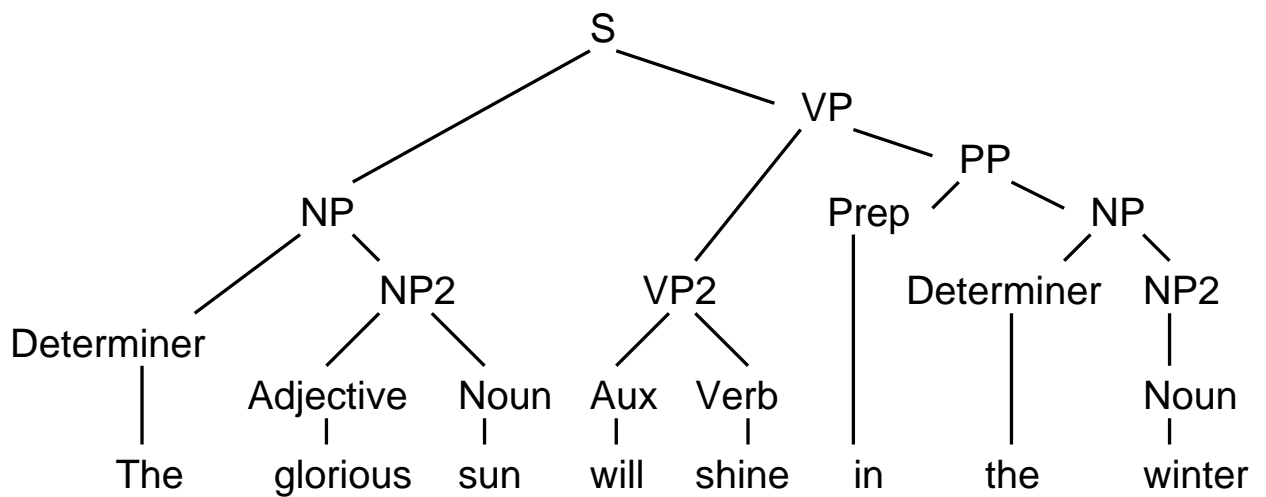
Bringing common sense into language



# 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)

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

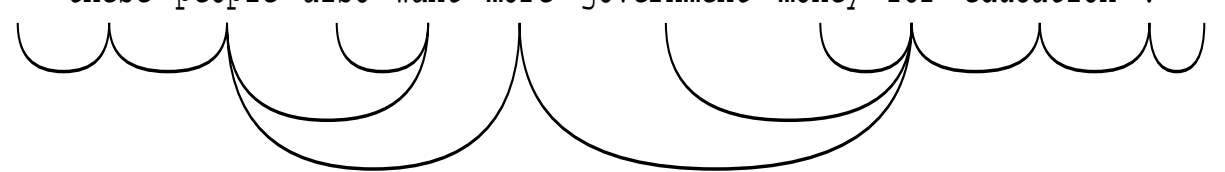
Simple Sentence 3/5  
(After 10,000 words of training)

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



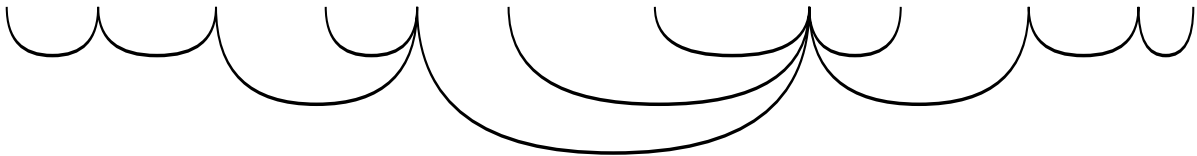
Simple Sentence 4/5  
(After 100,000 words of training)

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

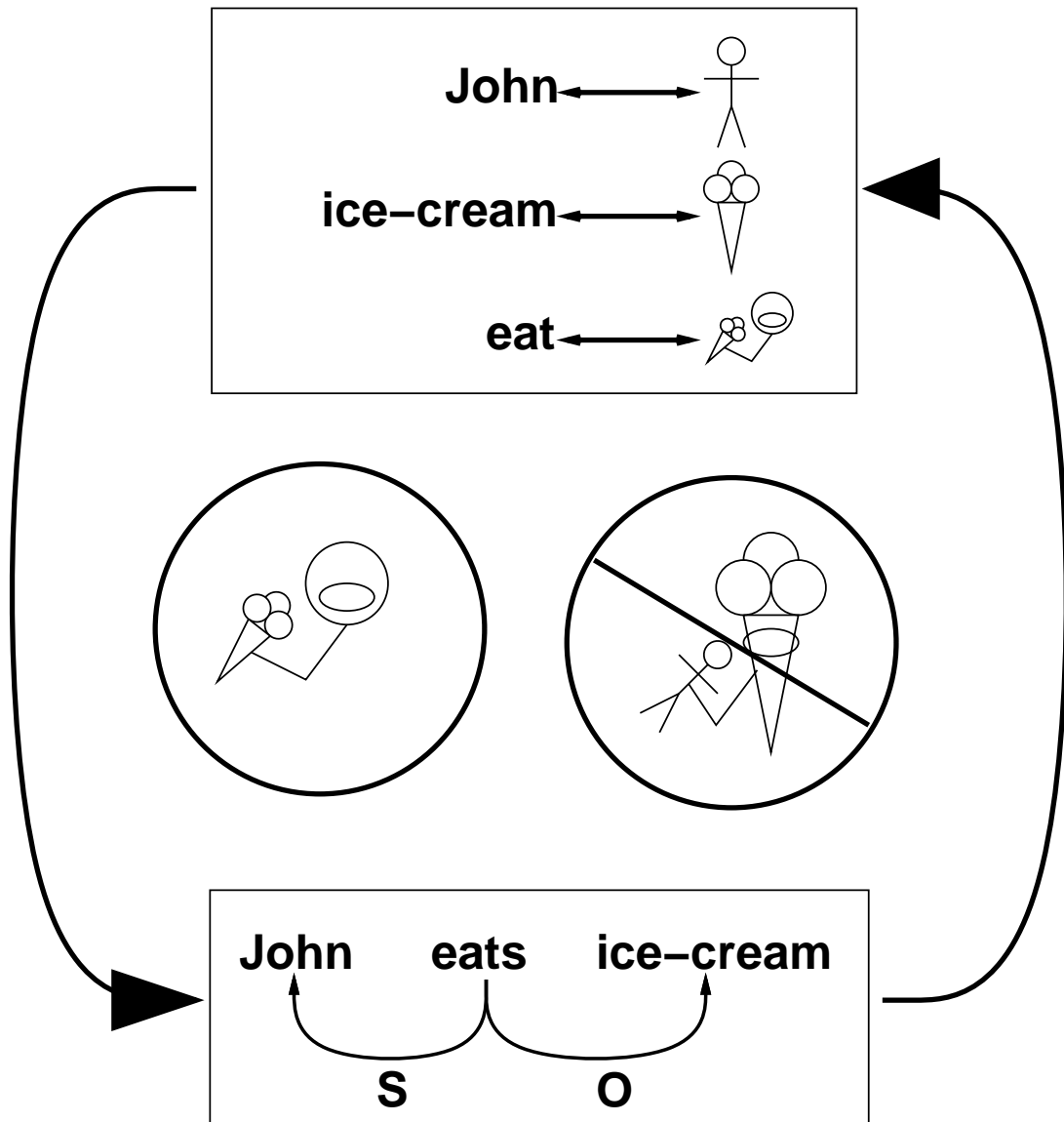


Simple Sentence 5/5  
(After 1,000,000 words of training)

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



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

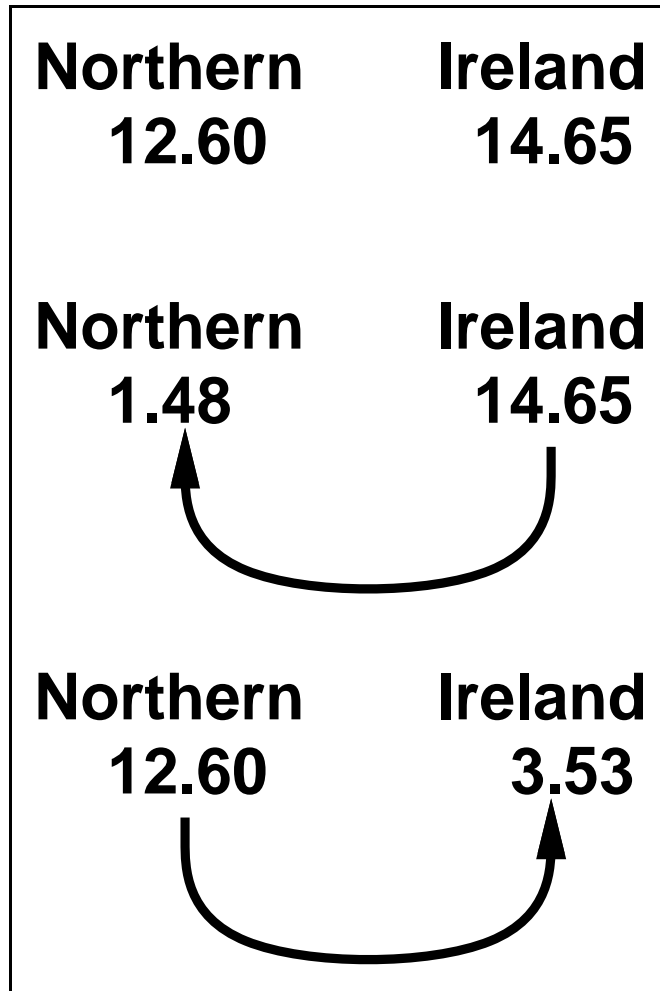
$$H = - \sum p_i \log p_i$$

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

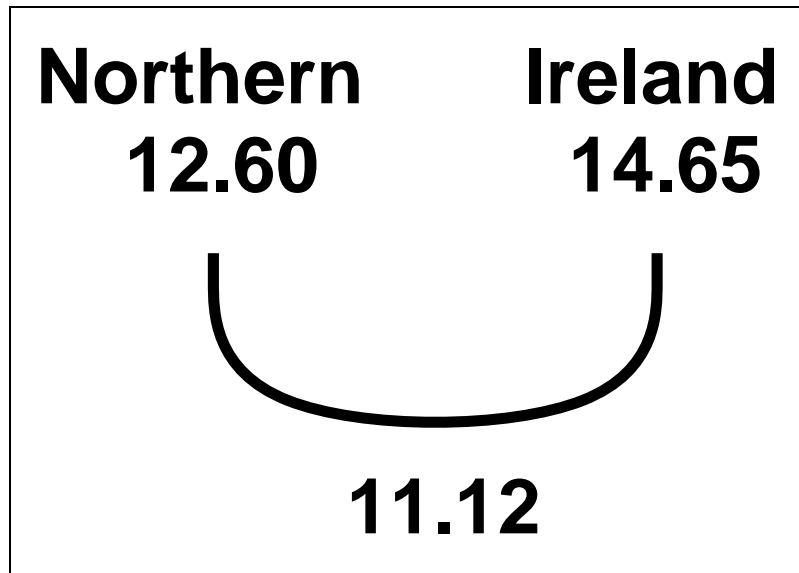
Total: 99.28 bits

The word pair and relative information:





The lexical attraction link:



## Language Model Determines the Context

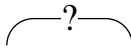
The IRA is fighting British rule in Northern Ireland

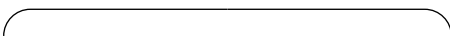
4.20 12.90 3.73 10.54 8.66 5.96 3.57 9.25 3.53

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

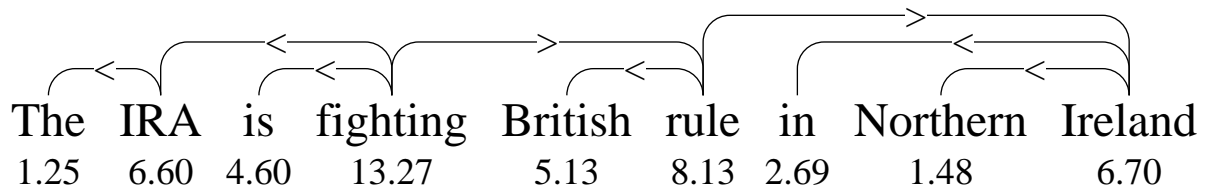
Total: 99.28  $\rightarrow$  62.34 bits

Context should be determined by syntactic relations:

The man with the  dog spoke

 The man with the dog spoke

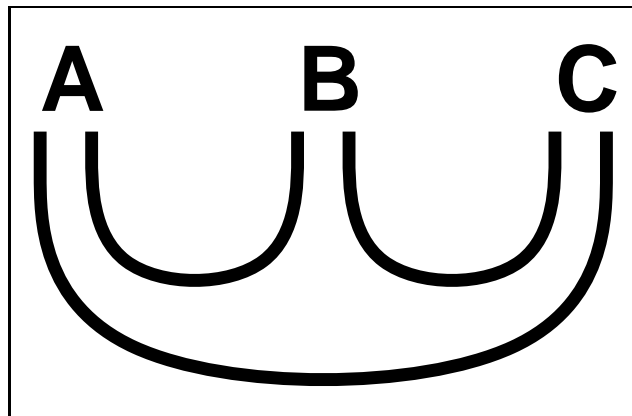
Context should be determined by syntactic relations:



Total: 62.34 → 49.85 bits

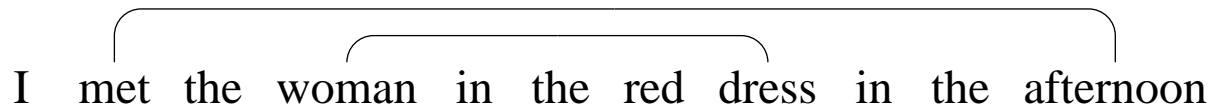
Dependency structure is acyclic:

- Mathematically: cannot use all the lexical attraction links in a cycle.
- Linguistically: cannot construct a consistent head-modifier structure.

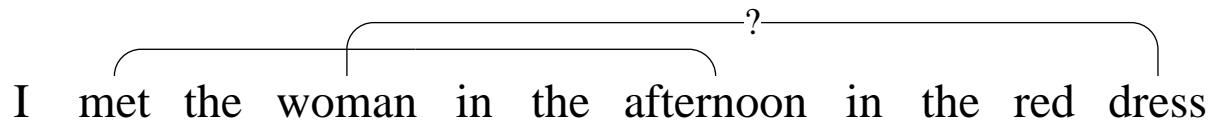


Syntactic relations form a planar tree:  
(Links do not cross)

I met the woman in the red dress in the afternoon



I met the woman in the afternoon in the red dress



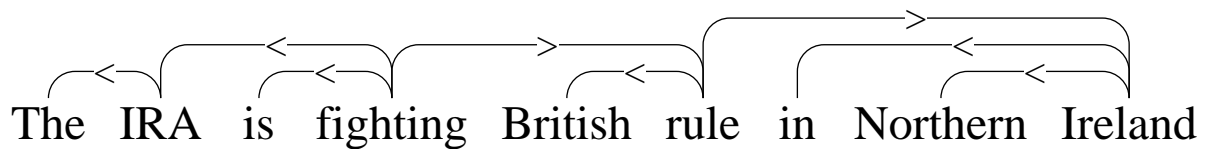
Syntactic relations form a planar tree:  
(Links do not cross)

- 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:



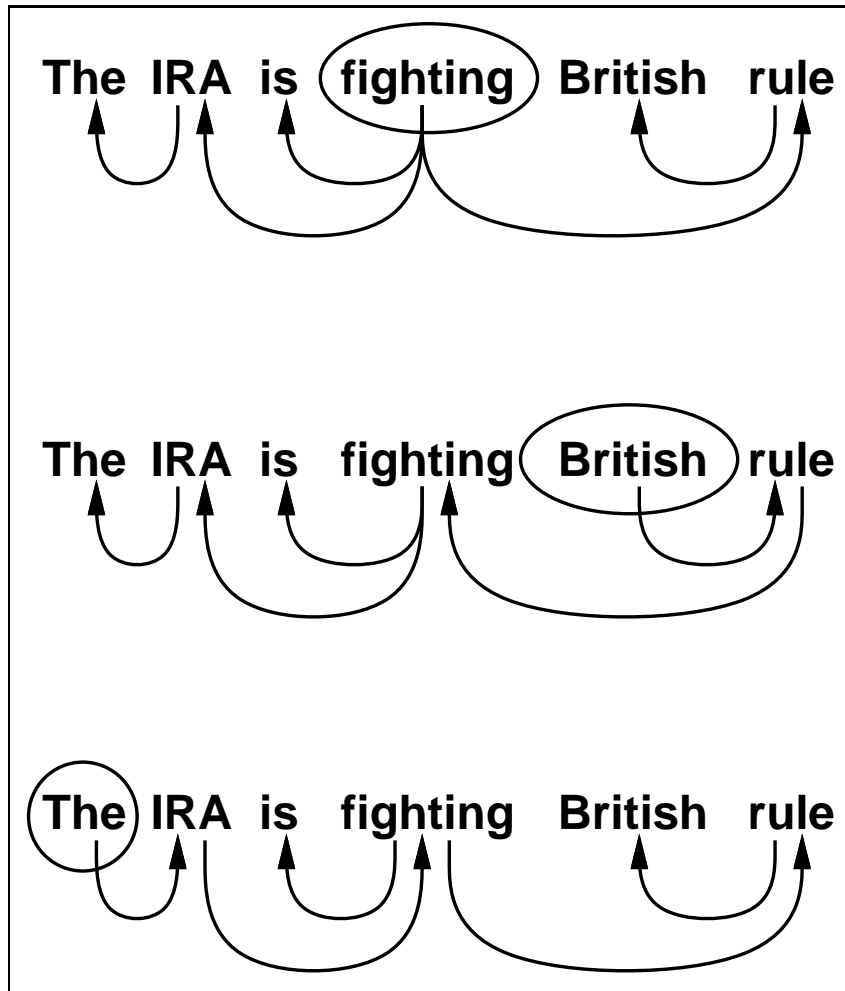
Encoding: LPLLPPRLPRLPLPPP

L:10 R:11 P:0

Upper bound: 3 bits per word



Lexical attraction is symmetric



Lexical attraction is symmetric

$$S = (W, L, w_0)$$

$$W = \{ w_i \}$$

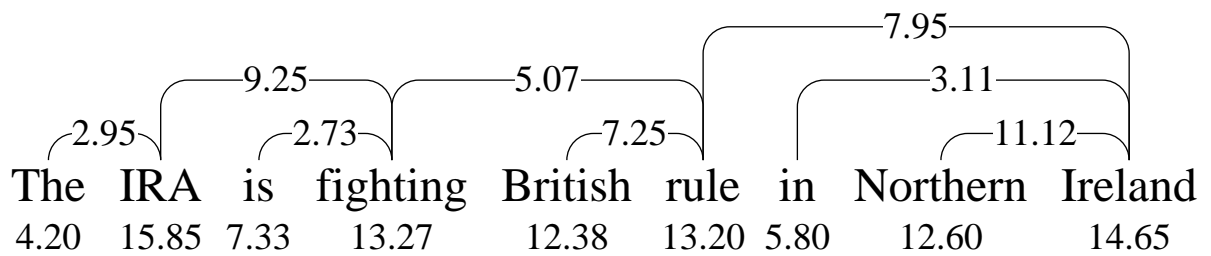
$$L = \{ (w_i, w_j) \}$$

$$P(S) = P(L)P(w_0) \prod_{(w_i, w_j) \in L} P(w_j | w_i)$$

$$= P(L)P(w_0) \prod_{(w_i, w_j) \in L} \frac{P(w_i, w_j)}{P(w_i)}$$

$$= P(L) \prod_{w_i \in W} P(w_i) \prod_{(w_i, w_j) \in L} \frac{P(w_i, w_j)}{P(w_i)P(w_j)}$$

Dependency structure is an undirected, acyclic, planar graph:



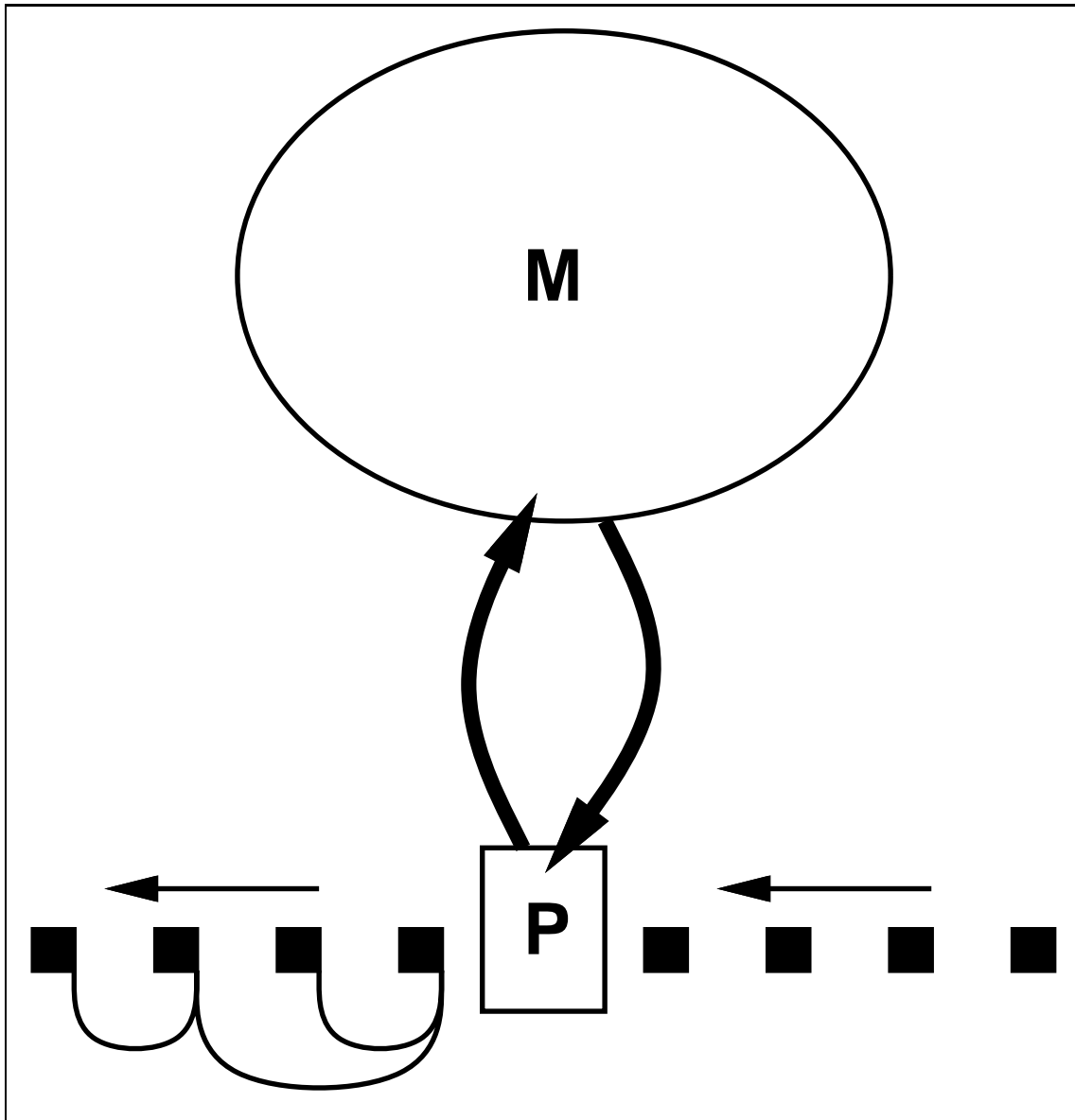
Information in a Sentence =

Information in Words

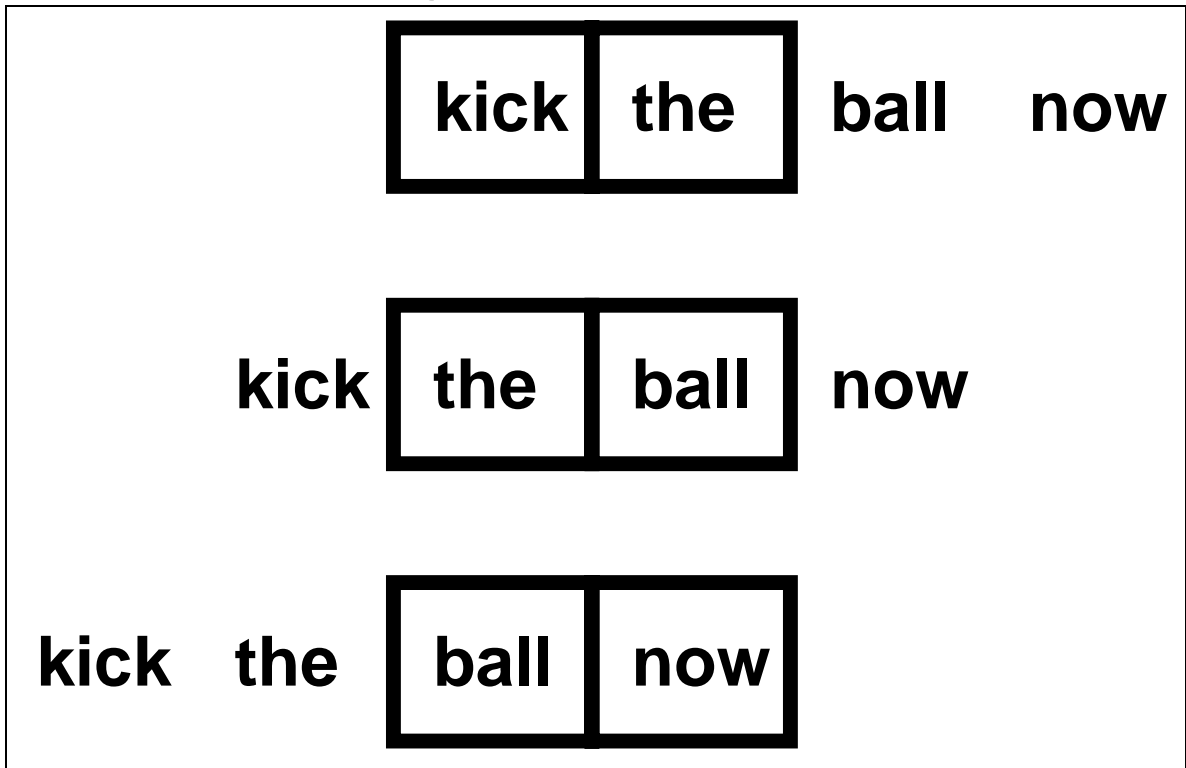
+ Information in the Tree

- Mutual Information in Syntactic Relations

# The Memory



The memory observes the processor



Learning simple structures

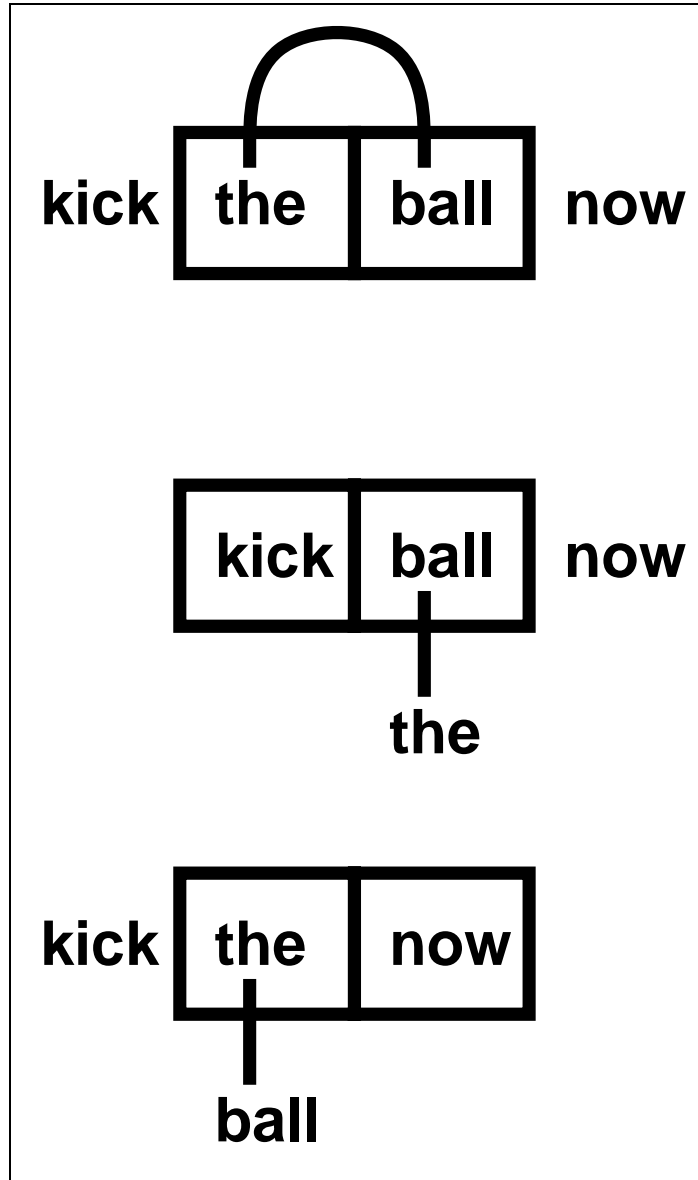
**kick** **the** **ball** **now**

**throw** **the** **ball** **at**

**with** **the** **ball** **in**

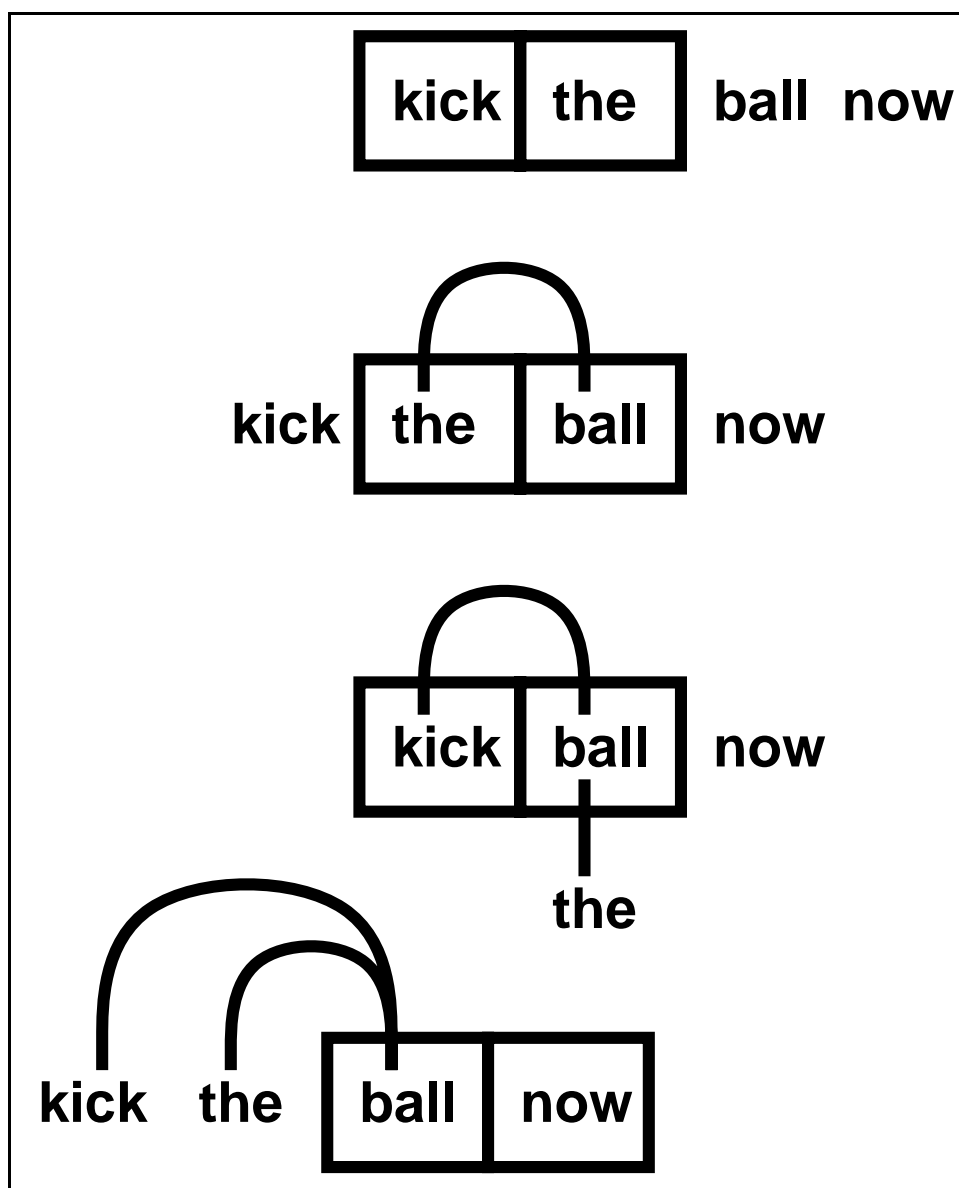
**kick** **the** **ball** **now**

Simple structures help see complex structures

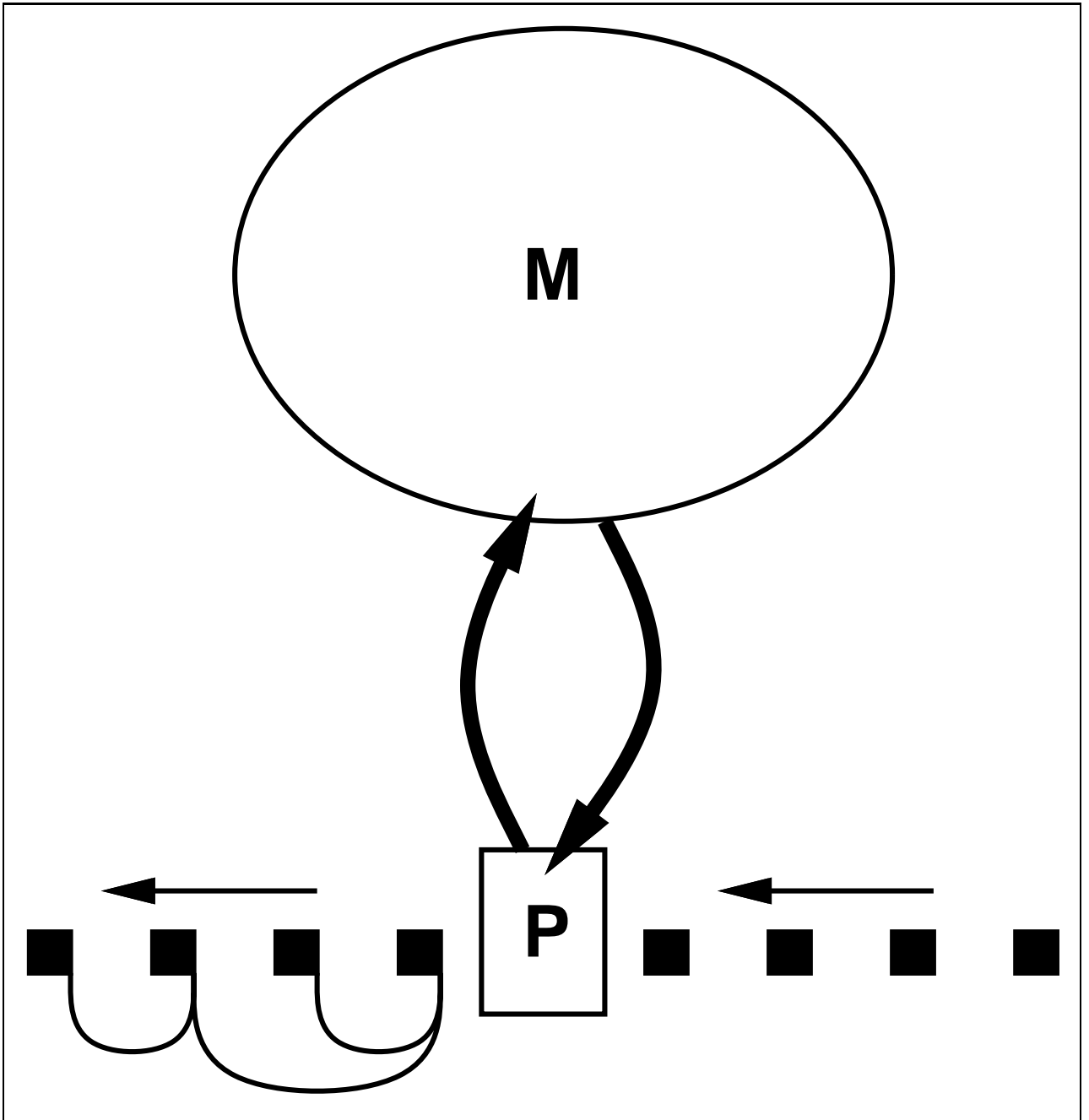




# Learning complex structures



# 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.

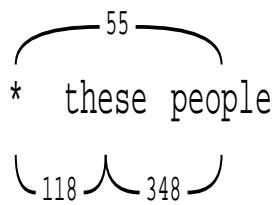


- New word considers links with previous words.

\* these people

The diagram shows the phrase '\* these people' with two curved brackets. The top bracket is positioned above the words 'these' and 'people' and contains the number '348'. The bottom bracket is positioned below the words 'these' and 'people' and contains the number '118'.

- Cycles are not allowed.
- Link with minimum score gets rejected.

  
\* these people

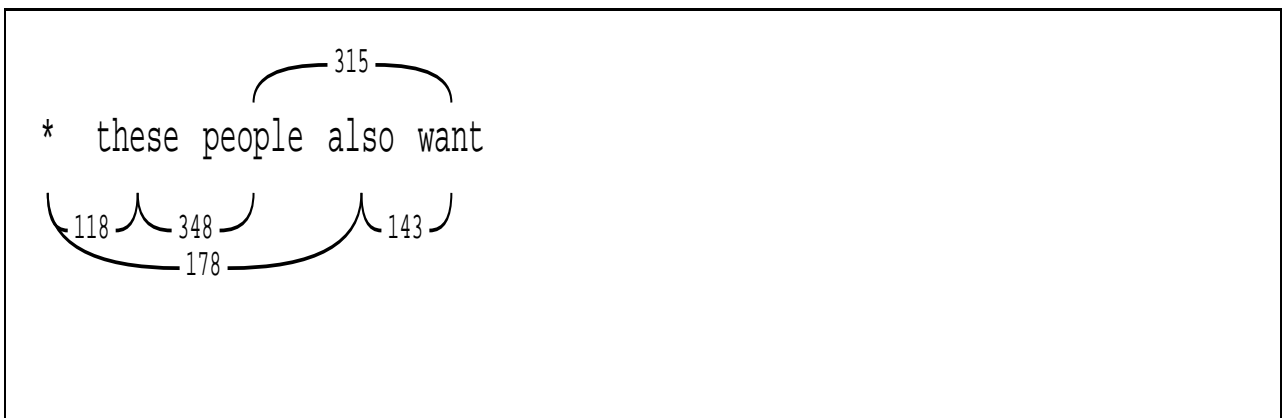
- Link with negative value not accepted.

\* these people also

118 348

-164

- Link crossing not allowed.
- Link with minimum score gets eliminated.

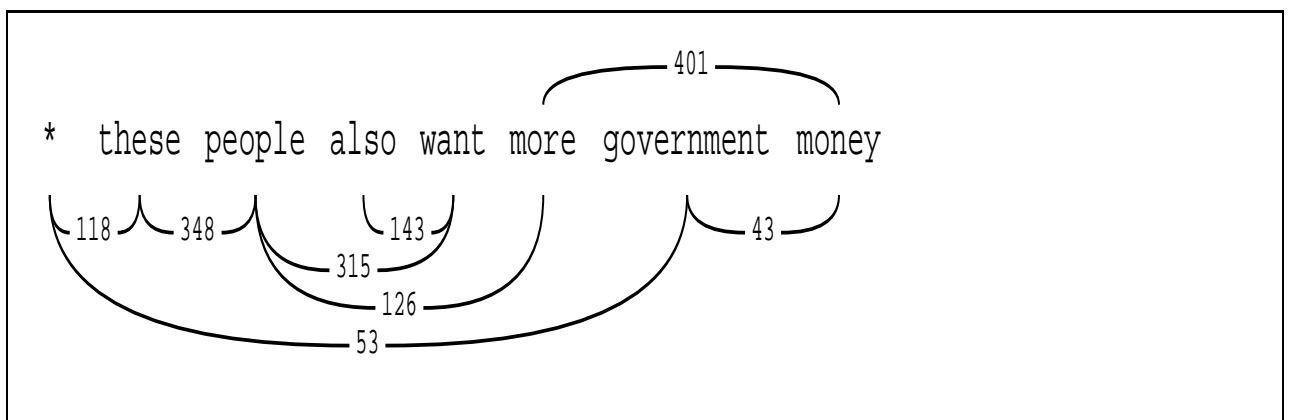




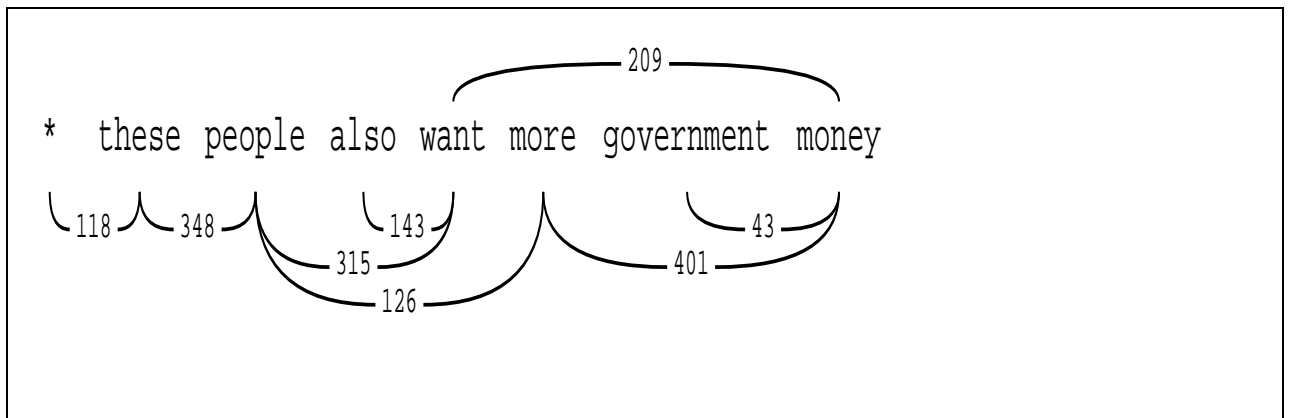
\* these people also want

118 348 143  
315

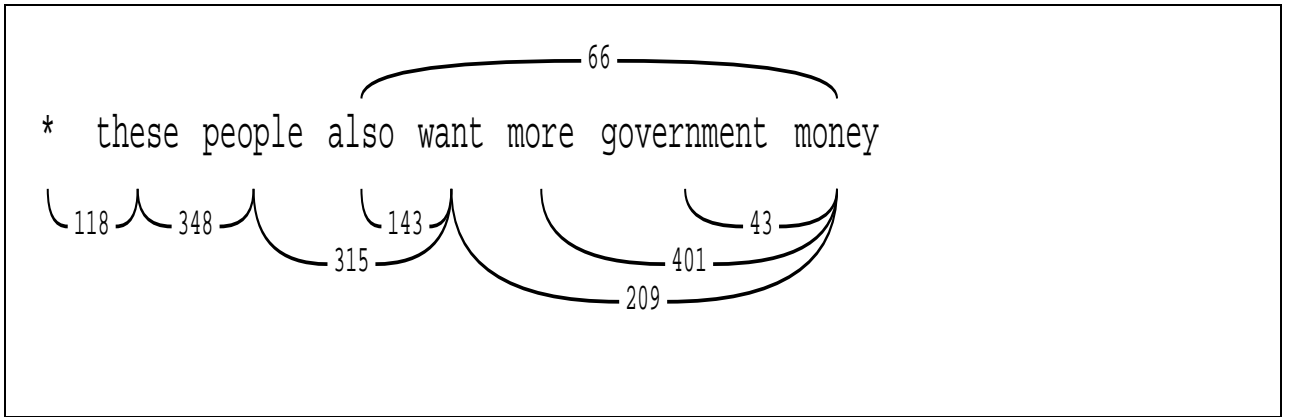
- The two constraints straighten out previous mistakes by eliminating bad links.



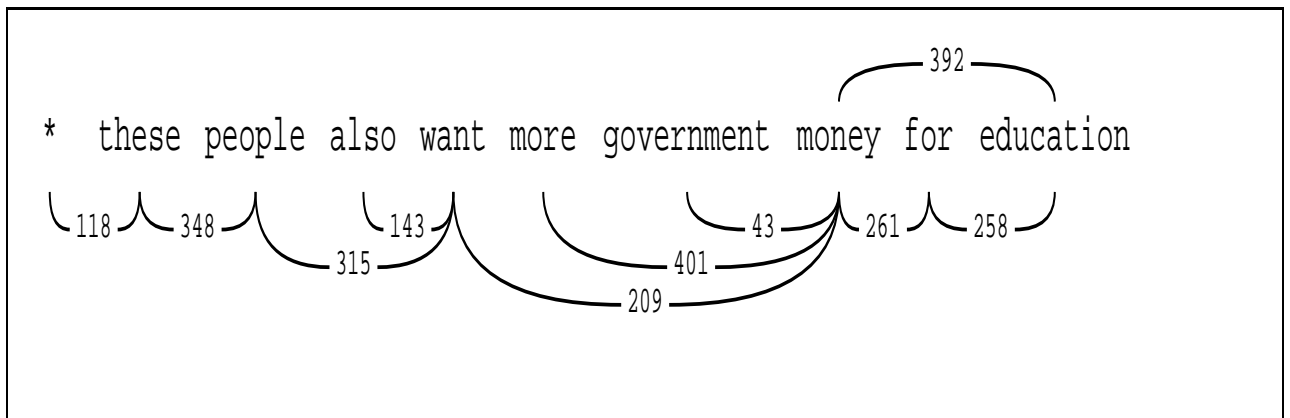
- Eliminating bad links 2/3



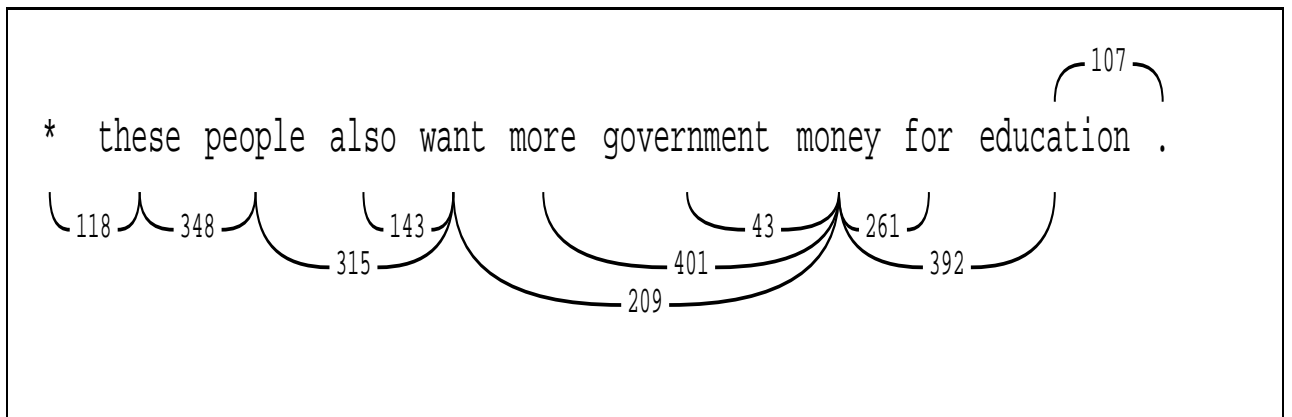
- Eliminating bad links 3/3



- 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)

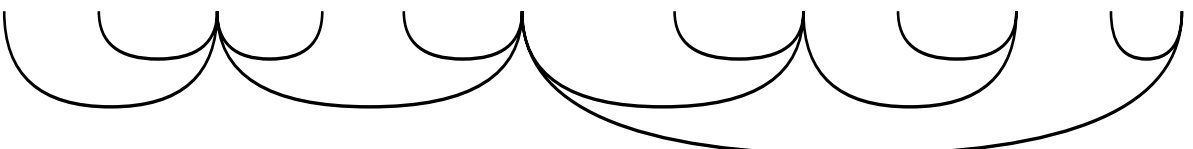
\* the cause of his death friday was not given . \*





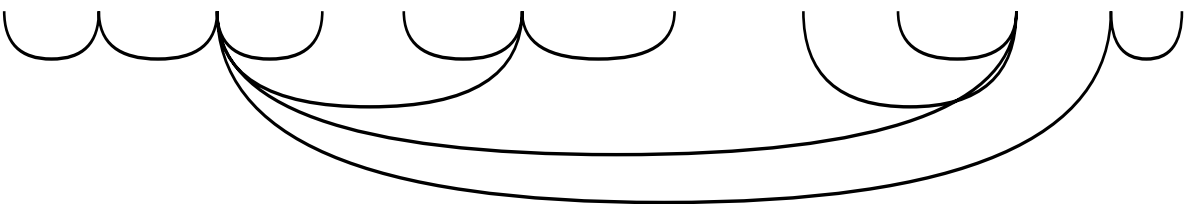
Long Distance Link 2/3  
(After 100,000 words of training)

\* the cause of his death friday was not given . \*




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

\* the cause of his death friday was not given . \*



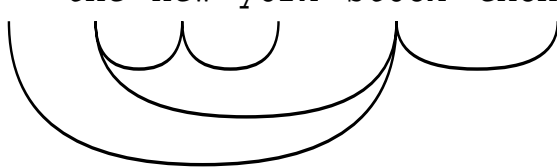
Complex Noun Phrase 1/4  
(After 10,000 words of training)

\* the new york stock exchange composite index fell . \*

The image shows the sentence "\* the new york stock exchange composite index fell . \*" with several hand-drawn arcs below it. A large arc spans from the start of the sentence to the end of the word "index". Two smaller arcs are drawn under "new york" and "stock exchange". A single arc is drawn under the period ".". The asterisks at the beginning and end of the sentence are not underlined.

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

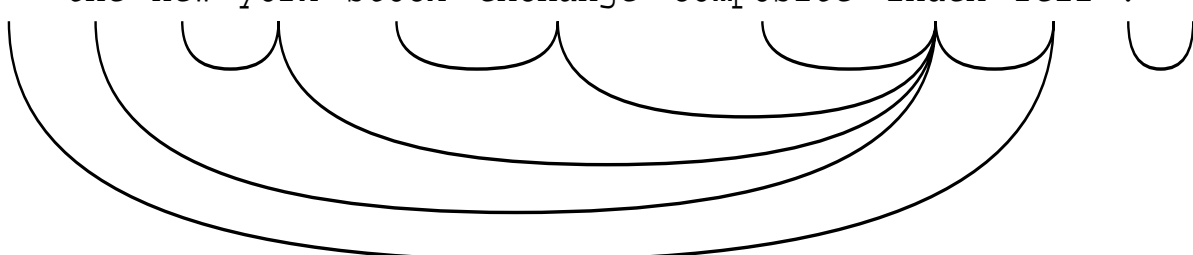
\* the new york stock exchange composite index fell . \*



The diagram illustrates the hierarchical structure of the noun phrase "the new york stock exchange composite index" using nested brackets. The outermost bracket encompasses the entire phrase. Inside it, a second-level bracket groups "new york stock exchange" together, with "new york" and "stock exchange" each having their own smaller brackets underneath. A third-level bracket groups "composite index" together, with "composite" and "index" each having their own smallest brackets underneath. This structure shows how the model has learned to parse the complex noun phrase into its constituent parts.


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

\* the new york stock exchange composite index fell . \*



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

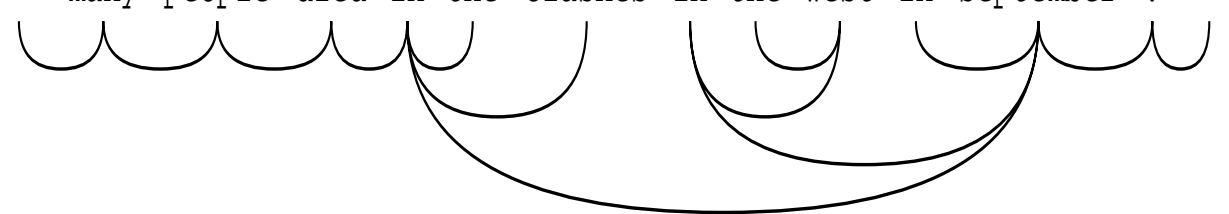
\* the new york stock exchange composite index fell . \*



The diagram illustrates the hierarchical structure of the noun phrase 'the new york stock exchange composite index' in the sentence '\* the new york stock exchange composite index fell . \*'. It features a series of nested, downward-curving arcs (brackets) positioned below the words. The innermost arcs group individual words: 'the', 'new', 'york', 'stock', 'exchange', 'composite', and 'index'. Successive, wider arcs group these words into larger units, representing the nested structure of the noun phrase. The outermost arc encompasses the entire noun phrase 'the new york stock exchange composite index', which is followed by the verb 'fell' and a period. The asterisks at the beginning and end of the sentence serve as markers for the overall phrase structure.

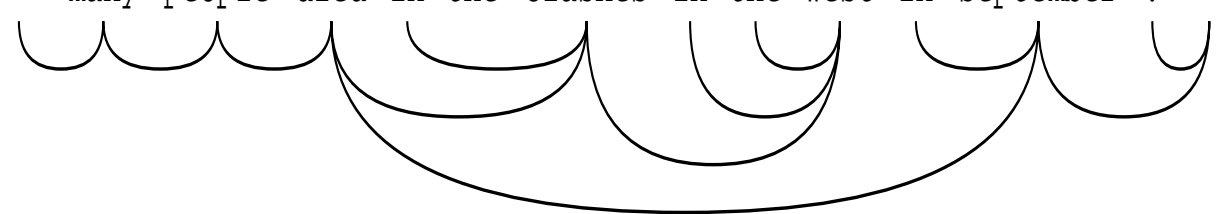
Syntactic Ambiguity 1/3  
(After 1,000,000 words of training)

\* many people died in the clashes in the west in september . \*



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

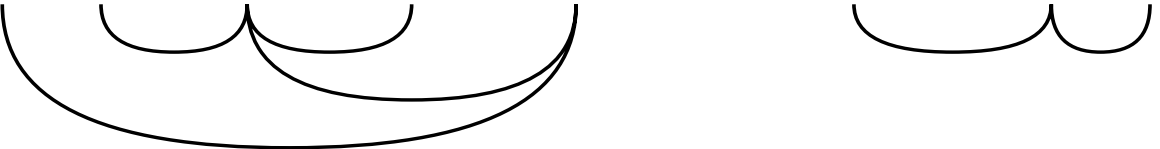
\* many people died in the clashes in the west in september . \*






Syntactic Ambiguity 2/3  
(After 500,000 words of training)

\* a number of people protested . \*



The diagram illustrates syntactic ambiguity for the sentence "a number of people protested .". It features two sets of arcs. The first set consists of three nested arcs: the innermost arc connects "a" and "number", the middle arc connects "number" and "of", and the outermost arc connects "a" and "of". The second set consists of two arcs: the inner arc connects "people" and "protested", and the outer arc connects "a" and "protested".

\* the number of people increased . \*



The diagram illustrates syntactic ambiguity for the sentence "the number of people increased .". It features two sets of arcs. The first set consists of three nested arcs: the innermost arc connects "the" and "number", the middle arc connects "number" and "of", and the outermost arc connects "the" and "of". The second set consists of two arcs: the inner arc connects "people" and "increased", and the outer arc connects "the" and "increased".

Syntactic Ambiguity 2/3  
(After 5,000,000 words of training)

\* a number of people protested . \*

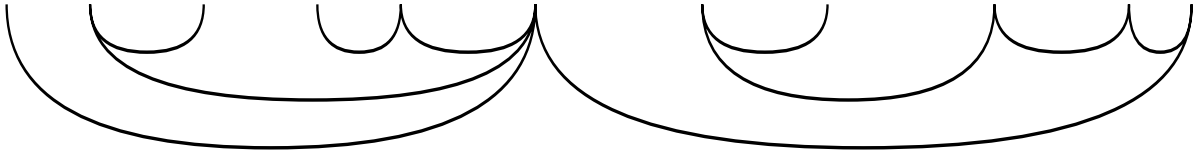
The diagram illustrates syntactic ambiguity for the sentence "a number of people protested". It features a long outer arc under the entire phrase "a number of people protested" and a shorter arc under the words "a number of". This structure suggests two possible interpretations: one where "a number" is the subject and "of people protested" is a modifier, and another where "a number of people" is the subject and "protested" is the verb.

\* the number of people increased . \*


The diagram illustrates syntactic ambiguity for the sentence "the number of people increased". It features a long outer arc under the entire phrase "the number of people increased" and a shorter arc under the words "the number of". This structure suggests two possible interpretations: one where "the number" is the subject and "of people increased" is a modifier, and another where "the number of people" is the subject and "increased" is the verb.

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

\* the driver saw the airplane flying over washington . \*




\* the pilot saw the train flying over washington . \*

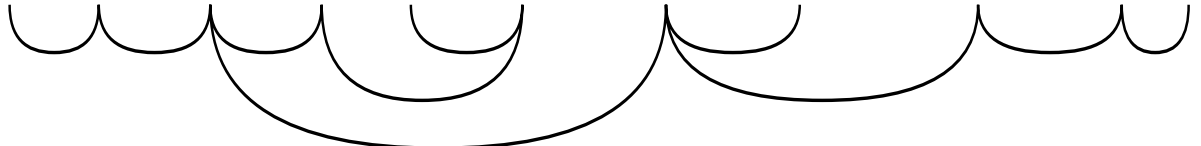


Syntactic Ambiguity 3/3  
(After 10,000,000 words of training)

\* the driver saw the airplane flying over washington . \*



\* the pilot saw the train flying over washington . \*

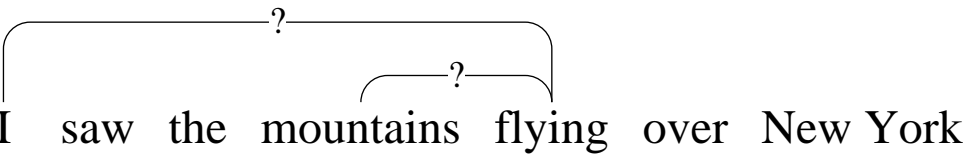


## Results

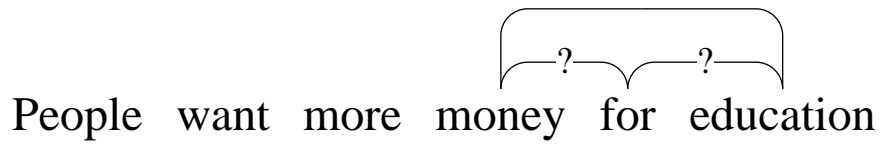
- Evaluation criteria
- Upper and lower bounds
- Link accuracy
- Related work

## Evaluation criteria: Content-word links

I saw the mountains flying over New York



People want more money for education



## 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:

$n_1$  = human links

$n_2$  = program links

$n_{12}$  = common links

- Precision =  $n_{12} / n_2$

- Recall =  $n_{12} / n_1$



Lower bound:

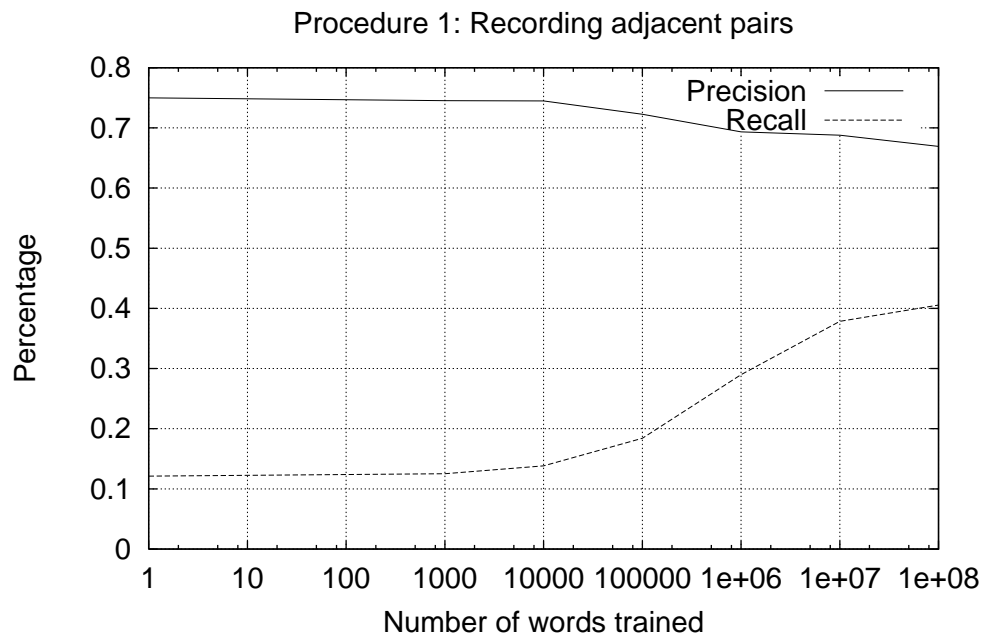
Random lexical attraction → 8.9% precision,  
5.4% recall

Linking every adjacent word → 41% recall

Upper bound:

85% of syntactically related pairs have positive lexical attraction

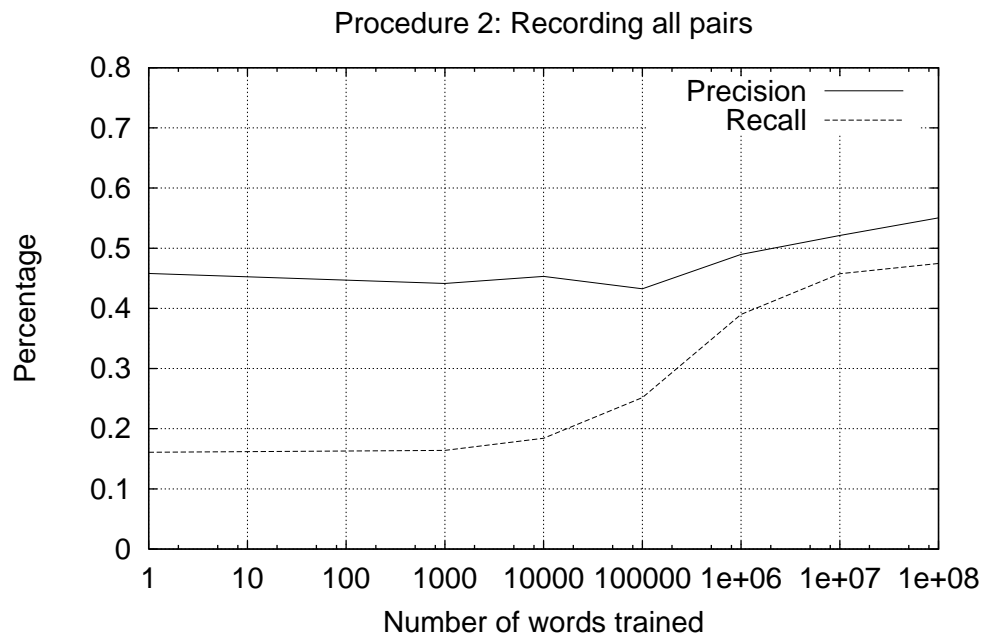
## Recording adjacent pairs



Precision = 67%

Recall = 41%

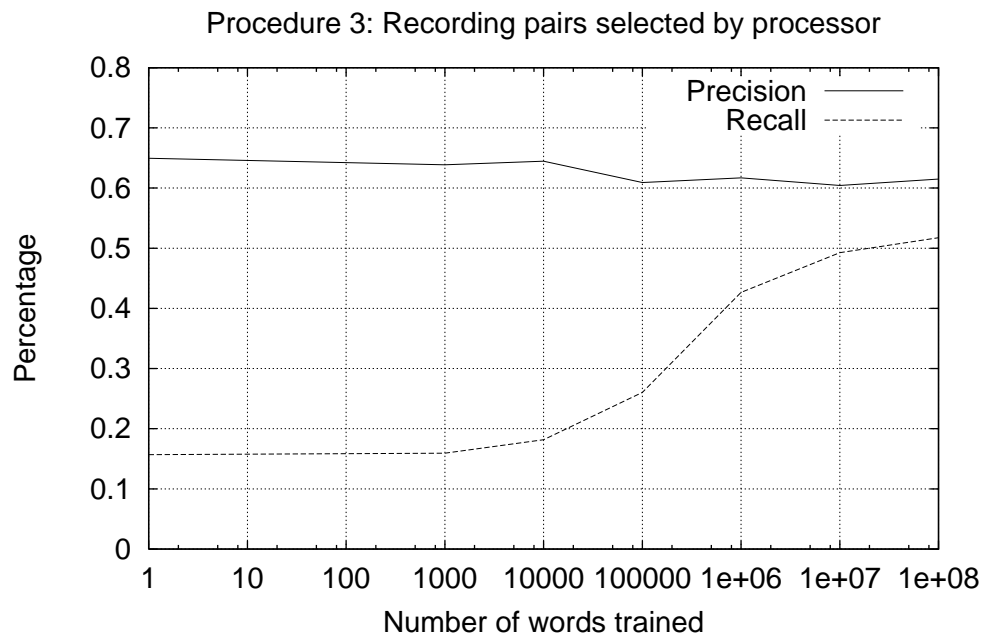
# Recording all pairs



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



## 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