Pragmatic and Discourse Analysis

6.	**Pragmatic and Discourse Analysis** Reference resolution, co reference resolution, pronoun resolution, approaches to pronoun resolution	**04**

Discourse Analysis

- **Discourse Analysis** is a field of linguistics that studies the ways in which language is used in real-world contexts.
- It examines how language is used to create meaning, structure social interactions, and shape our understanding of the world.
- Language does not normally consist of isolated, unrelated sentences, but instead of collocated, related groups of sentences. Such a group of sentences is **discourse.**
    - **"John went to Bill's car dealership to check out an Acura Integra. He looked at it for about an hour."**
        - What do pronouns such as he and it denote?
        - Can we build a computational model for the resolution of referring expressions?

Need for coreference resolution

- Look at the following examples
  - "There is a pile of inflammable trash next to your car. You will have to get rid of it.
  - If these shoes don't fit your feet, exchange them
  - If the baby does not thrive on raw milk, boil it.

Cohesion

Cohesion is a crucial concept in discourse analysis because it focuses on **how different parts of a text or conversation are connected**, making it easier for participants to follow and understand.

In discourse analysis, cohesion is studied to understand how speakers or writers use linguistic elements to link ideas, creating a sense of unity and logical flow within communication.

For example, in the sentence "**The cat chased the mouse, and the mouse ran away," the conjunction "and" is used to connect the two sentences and create a coherent narrative.**

- Task: summarize the following paragraph

First Union Corp is continuing to wrestle with severe problems. According to industry insiders at Paine Webber, their president, John R. Georgius, is planning to announce his retirement tomorrow.

Summary is

First Union President John R. Georgius is planning to announce his retirement tomorrow.

Coherence Relations

- From the example of summarization above we understand that to build such a summary we need to know
  - Second sentence in the above paragraph is more important
  - First sentence is just a subordinate to sentence 2
- Relationships of this sort between the sentences are called **coherence relations**
- Determining coherence structures between the sentences is an important task in discourse analysis
- Coherence property is useful in tasks such as automatic essay grading
- Coherence is also useful in evaluating the output quality of the NLP generation systems

Understanding coherence

- Consider the following example

(21.4) John hid Bill's car keys. He was drunk.

(21.5) ?? John hid Bill's car keys. He likes spinach.

Which sentence seems odd??

Coreference

- It occurs when two or more expressions in a text refer to the same person or thing; they have the same referent.

Bill said he would come;

The proper noun Bill and the pronoun he refer to the same person, namely to Bill.

*Diya said that **she** will call you later.*

Diya and **she** refer to the **same person**.

*The teacher told the students that **the teacher** would help them after class.*

Both mentions of **the teacher** refer to **one person**, so this is **coreference** too (even though the same noun phrase is repeated).

*Mr. Sharma is our principal. **The head of the school** is very strict.*

“Mr. Sharma” and “the head of the school” → both refer to the same person → **coreference**.

Coreference resolution

- Coreference resolution is the task of clustering mentions in text that refer to the same underlying real world entities.
- Coreference resolution, is the task of finding all expressions that are coreferent with any of the entities found in a given text.
- Coreference resolution is the task of resolving noun phrases to the entities that they refer to.
- For example, in the sentence, “Andrew said he would buy a car” the pronoun “he” refers to the same person, namely to “Andrew”.

Where can coreference resolution be useful?

- Coreference resolution is one of the most challenging issues in the field of natural language processing
- It is an essential key in
  - text comprehension
  - information extraction
  - Summarization
  - questions answering
  - machine translation
  - Conversational agents

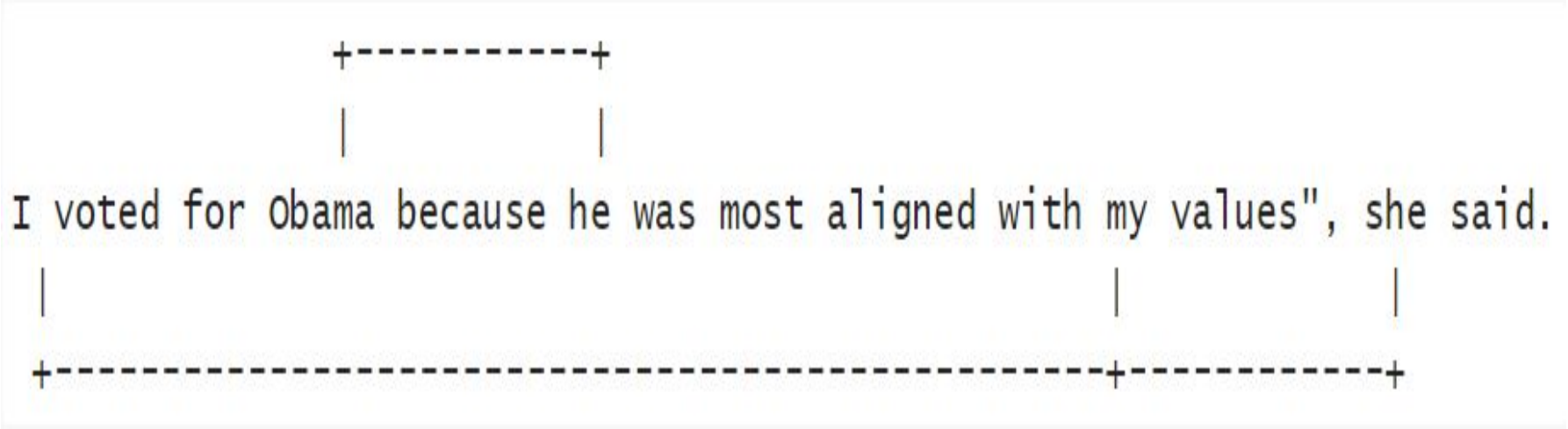
Coreference :	Two or more expressions refer to the same entity	*Riya said she will call.*
Anaphora:	Coreference where the reference points backward	*John arrived. He was tired.*
Cataphora:	Coreference where the reference points forward	*Before he spoke, John paused.*

- Reference
  - to mention or to refer
  - refer to an entity
- Referent
  - the thing in the world that a word or phrase denotes or stands for (Oxford dictionary)
  - Entity that is referred to
- Antecedent
  - a thing that existed before or logically precedes another (Oxford dictionary)
- Coreference
  - Expressions referring to the same entity
- Anaphora
  - Reference to an antecedent entity
- Cataphora
  - Reference to a descendent entity
- Referring expression
  - Expression used to perform reference

Approach to reference resolution system

- Rule based
  - Hobb's algorithm √
- Mention pair models √
- Mention ranking models
- Entity mention models
- Clustering based models

Coreference resolution



“I”, “my”, and “she” belong to the same cluster and “Obama” and “he” belong to the same cluster.

Coreference resolution

Paul Allen was born on January 21, 1953, in Seattle, Washington, to Kenneth Sam Allen and Edna Faye Allen. Allen attended Lakeside School, a private school in Seattle, where he befriended Bill Gates, two years younger, with whom he shared an enthusiasm for computers. Paul and Bill used a teletype terminal at their high school, Lakeside, to develop their programming skills on several time-sharing computer systems.

RUN

0 Paul Allen was born on January 21 , 1953 , in 1 Seattle , Washington , to Kenneth Sam Allen and Edna Faye Allen .

0 Allen attended Lakeside School , a private school in 1 Seattle , where 0 he befriended 2 Bill Gates , two years younger , with whom 0 he shared an enthusiasm for computers . 3 0 Paul and 2 Bill used a teletype terminal at 3 their high school , Lakeside , to develop 3 their programming skills on several time - sharing computer systems .

Coreference resolution

Document

The legal pressures facing Michael Cohen are growing in a wide-ranging investigation of his personal business affairs and his work on behalf of his former client, President Trump. In addition to his work for Mr. Trump, he pursued his own business interests, including ventures in real estate, personal loans and investments in taxi medallions.

RUN

The legal pressures facing [0 Michael Cohen] are growing in a wide - ranging investigation of [0 his] personal business affairs and [0 his] work on behalf of [1 [0 his] former client , President Trump] . In addition to [0 his] work for [1 Mr. Trump] , [0 he] pursued [0 his] own business interests , including ventures in real estate , personal loans and investments in taxi medallions .

Binding Theory

- Binding theory concerns syntactic restrictions on nominal reference
- It particularly focuses on the **possible coreference relationships** between a pronoun and its antecedent
- Example:
  - He voted for himself
  - She voted for her

The Three Principles of Binding Theory

Principle A – Reflexive pronouns Anaphors (reflexive pronouns like *himself, herself, themselves*) **must refer to an antecedent within the same clause.**

Example:

John$_1$ *hurt* ***himself***$_1$.

(“himself” refers to “John” → allowed)

Example:

John$_1$ *said that Mary hurt* ***himself***$_1$.

(“himself” is not in the same clause as “John” → **not allowed**)

The Three Principles of Binding Theory

Principle B – Pronouns

A pronoun (he, she, them, etc.) **must NOT refer to an antecedent within the same clause.**

Example:

John$_1$ said that Mary likes ***him$_1$***.

(“him” and “John” are in different clauses → allowed)

Example:

John$_1$ likes ***him$_1$***.

(“him” and “John” in the same clause → **not allowed**)

The Three Principles of Binding Theory

Principle C – Names and referring expressions

A proper noun or R-expression (like “John”, “the boy”, “Mary”) must be free — it cannot be bound by another NP.

Example:

*He₁ said that **John₁** is smart.*

(“John” refers independently — allowed)

Example:

***He₁** likes **John₁**.*

(“John” is bound by “He” → not allowed) *considering He and John refer to the same person- coreference.*

Approach to coreference resolution

Coreference Resolution in Two Steps :

1. Detect the mentions (easy)

“[I] voted for [Nader] because [he] was most aligned with [[my] values],” [she] said

2. Cluster the mentions (hard)

“[I] voted for [Nader] because [he] was most aligned with [[my] values],” [she] said

Hobb's Algorithm

- Hobbs' algorithm was one of the earliest approaches to pronoun resolution. The algorithm is mainly based on the **syntactic parse tree of the sentences**.
- It is a **syntactic, rule-based algorithm** used to identify the **antecedent** of a **pronoun** — that is, the noun phrase (NP) that the pronoun refers to — by **traversing the parse tree** of a sentence or set of sentences.
- It uses **syntactic structure (tree relationships)** rather than world knowledge or semantics.
- Example:
  - *John went to the store. He bought milk.*

    Here, Hobbs' algorithm identifies that **"He" → "John."**

Hobb's Algorithm

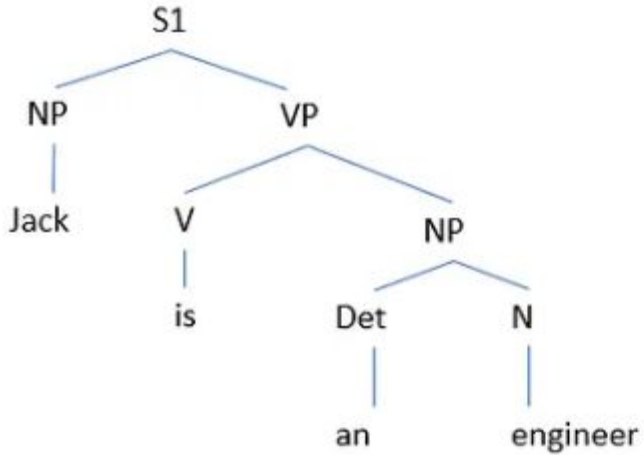
1. Start with the pronoun node in the parse tree
2. **Move up the tree to the first NP or S (sentence) node encountered.** Label this node as **X**, and note the **path (P)** from the pronoun to X.
3. **From node X, traverse the tree left-to-right, breadth-first, below X, skipping:**
    a. Any nodes along path P, and
    b. Any subtrees to the **right** of the path P.
4. Each NP node you encounter that **can be a possible antecedent** (matches gender/number, etc.) is a **candidate**.
    a. If one matches, **select it as the antecedent.**
    b. If not, move upward.
5. If you reach the top of the current sentence and still haven't found a match:
    a. Move to **previous sentences**, and
    b. Traverse their parse trees **left-to-right, breadth-first**, proposing all NP nodes that agree in features (number, gender).

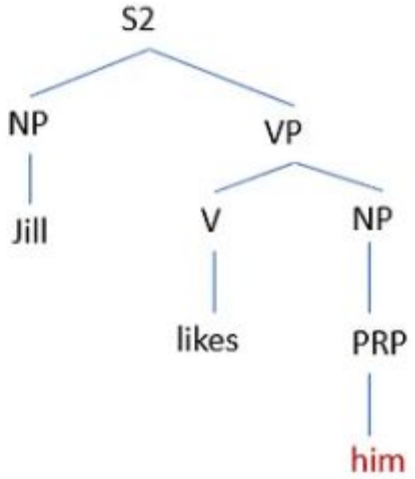
Consider two sentences:

Sentence 1(S1): Jack is an engineer.

Sentence 2 (S2): Jill likes him.

So here, we have the **syntactic parse tree** of the two sentences as shown.





The algorithm starts with the target pronoun and walks up the parse tree to the root node ‘S’. For each noun phrase or ‘S’ node that it finds, it does the **breadth first left to right search** of the node’s children to the left of the target.

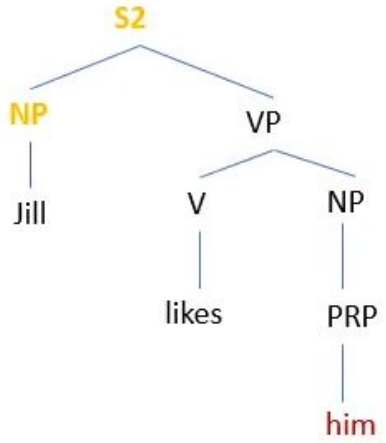
So in our example, the algorithm starts with the parse tree of the sentence 2 and climbs up to the root node S2.

Then it does a breadth first search to find the noun phrase (NP). Here the algorithm, finds its first noun phrase for noun ‘Jill’.

But it does not explore that branch because of the syntactic constraint of **Binding theory: Principle B – Pronouns**

> **A pronoun** (he, she, them, etc.) **must NOT refer to an antecedent within the same clause. Slide 20**

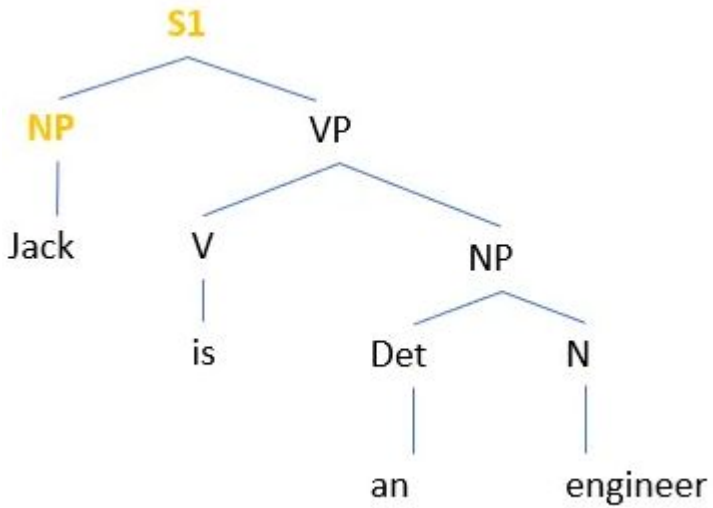
Also because of the **gender agreement constraint** even if the branch was explored, Jill won’t be the accepted referent for pronoun ‘him’.

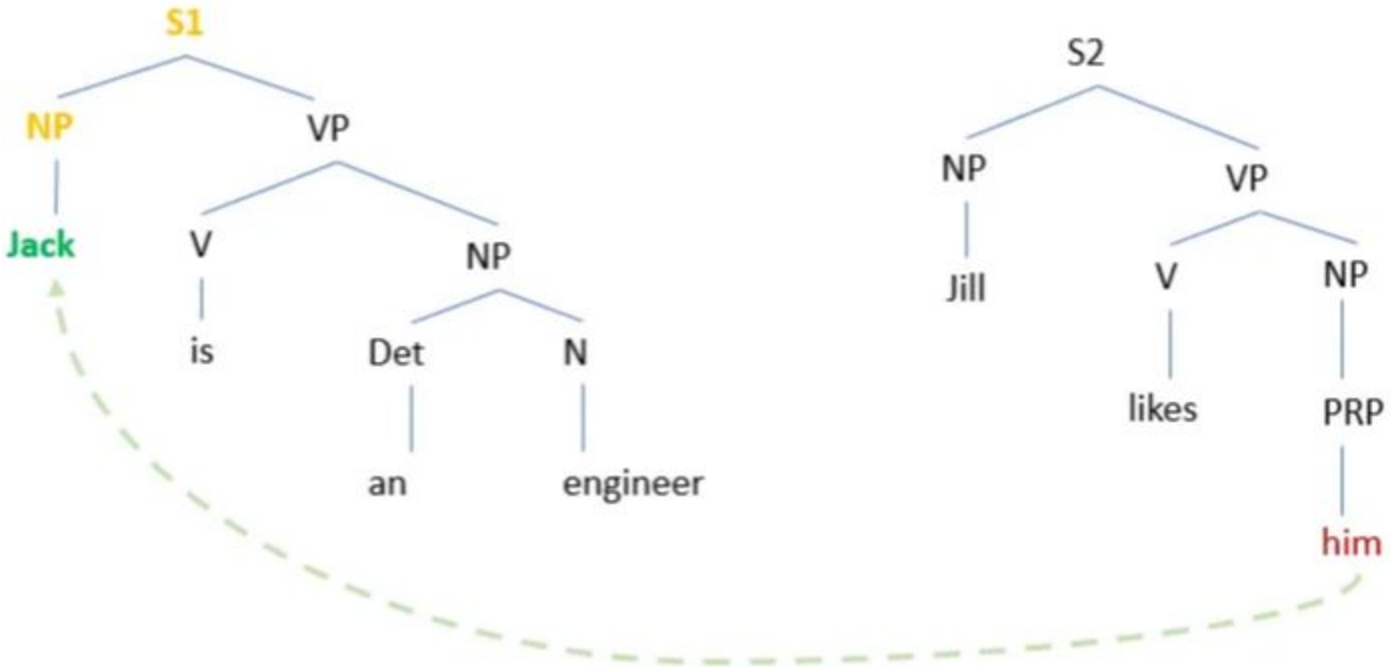


Hence the algorithm now starts the search in the syntax tree of the previous sentence.

For each noun phrase that it finds it does a breadth first **left to right** search of the node's children.

And hence the **subject Jack** in the sentence, Jack is an engineer, is **explored before** the **object engineer** and finally Jack is the resolved referent for the pronoun him.




S1
NP
VP
Jack
V
NP
is
Det
N
an
engineer
S2
NP
VP
Jill
V
NP
likes
PRP
him

Jack and Jill went up the hill

To fetch a pail of water

Jack fell down and broke his crown

And Jill came tumbling after

Jack and Jill went up the hill

To fetch a pail of water

Jack fell down and broke his crown

And Jill came tumbling after

Search for the referent is always restricted to the left of the target

Jack and Jill went up the hill

To fetch a pail of water

Jack fell down and broke his crown

And Jill came tumbling after

Gender agreement which eliminates the possibilities of Jill, hill and water.

Jack and Jill went up the hill

To fetch a pail of water

Jack fell down and broke his crown

And Jill came tumbling after

Recency property-Pronouns can only go a few sentences back

Example where Hobb's algorithm will fail

- *The city council refused the women a permit because they feared violence*
- *The city council refused the women a permit because they advocated violence*

- Hobb's algorithm will fail
- Need knowledge of the world

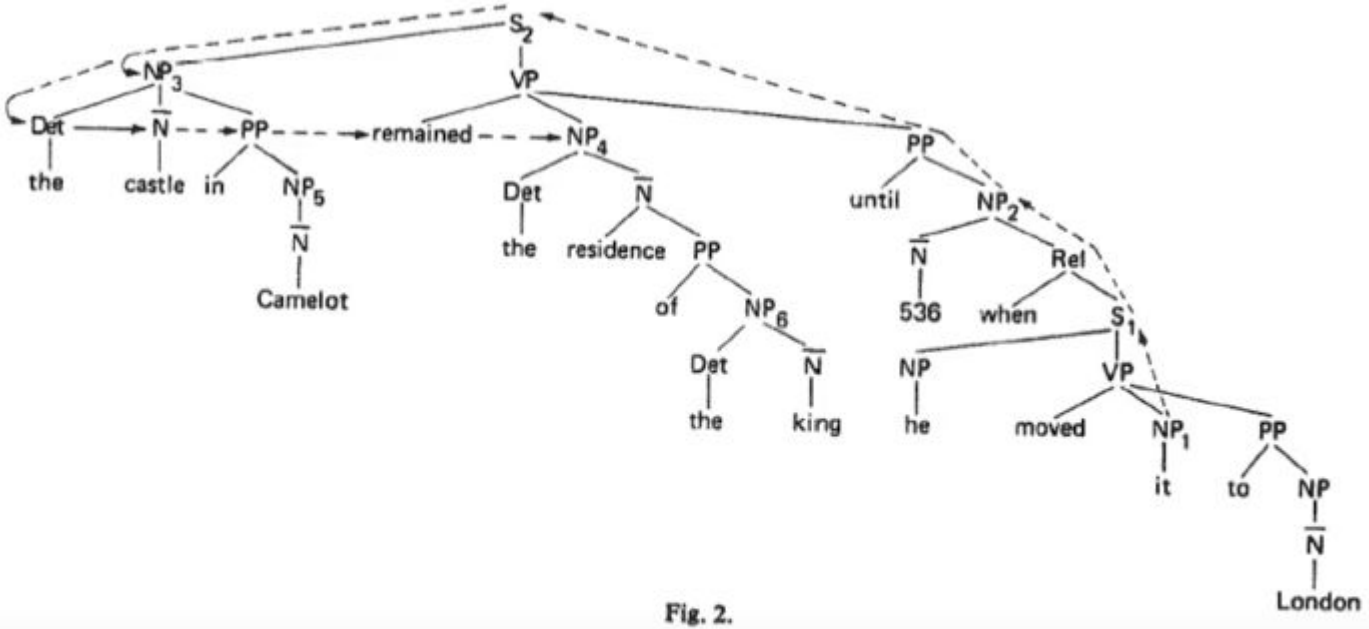
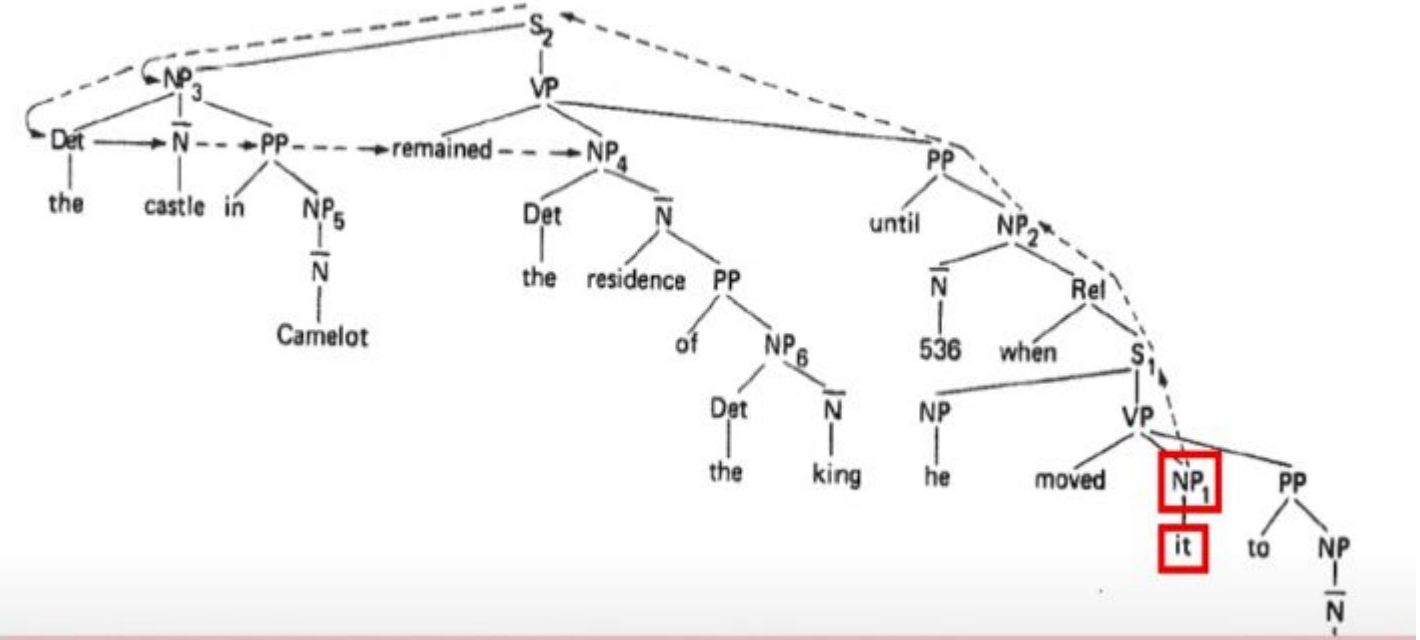


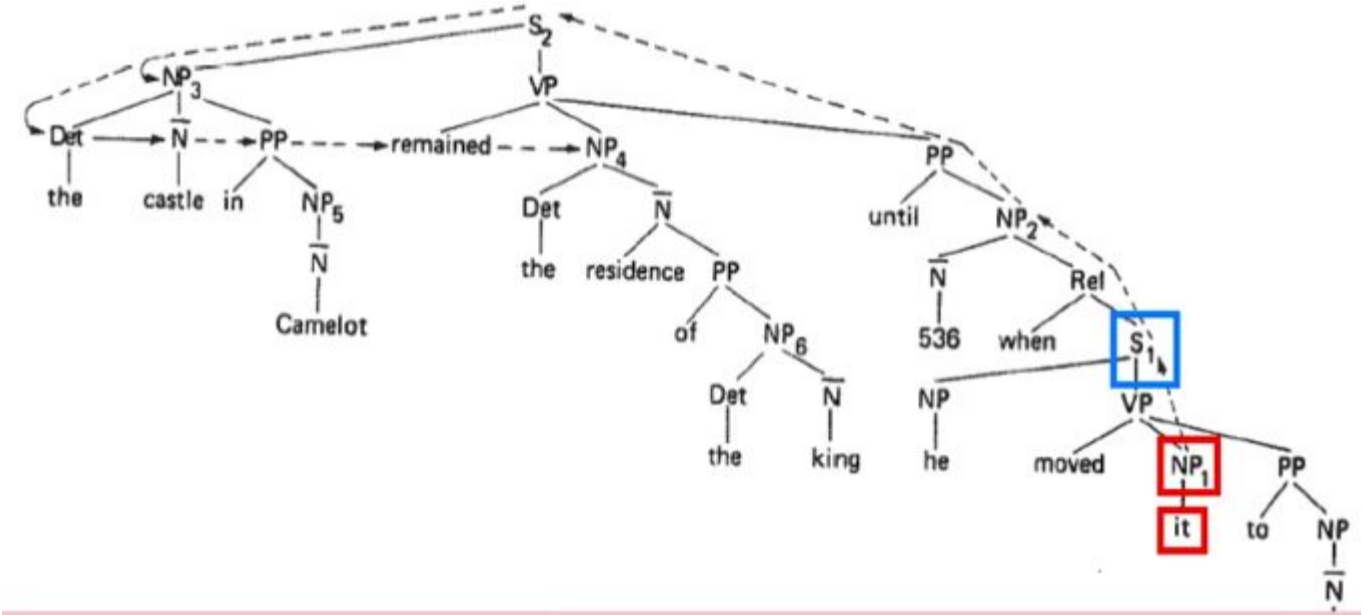
Fig. 2.

Setup

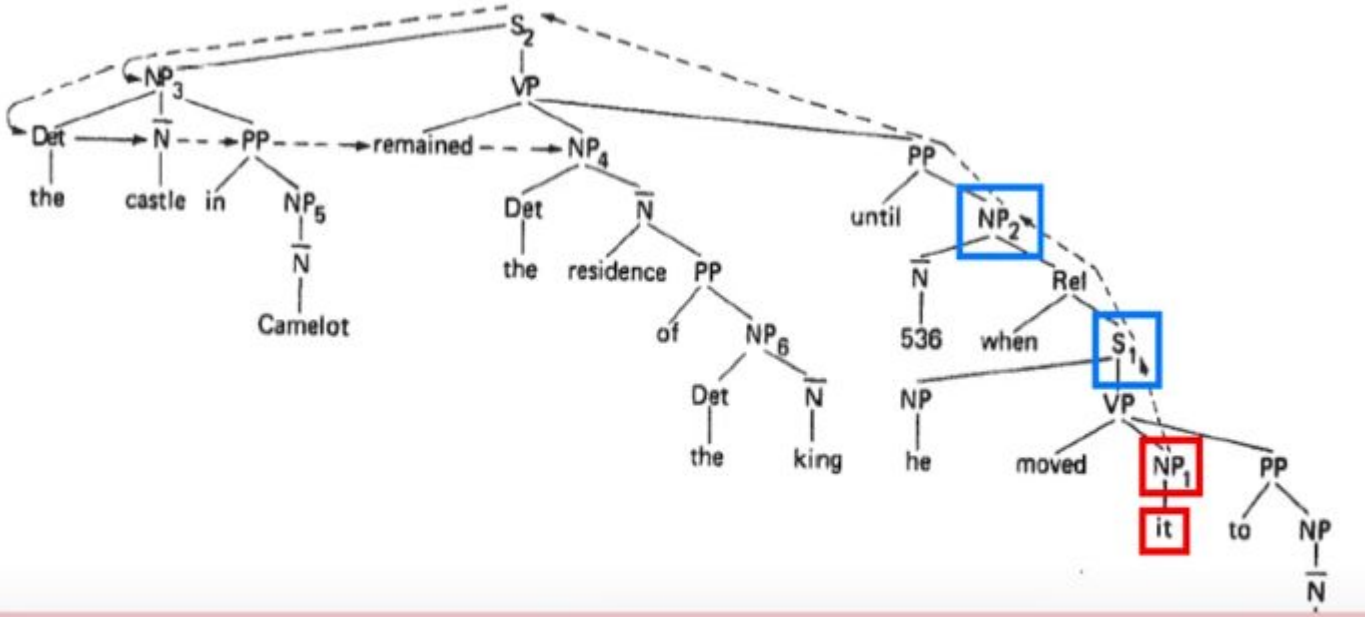
In this sentence, “it” refers to “the residence of the king”



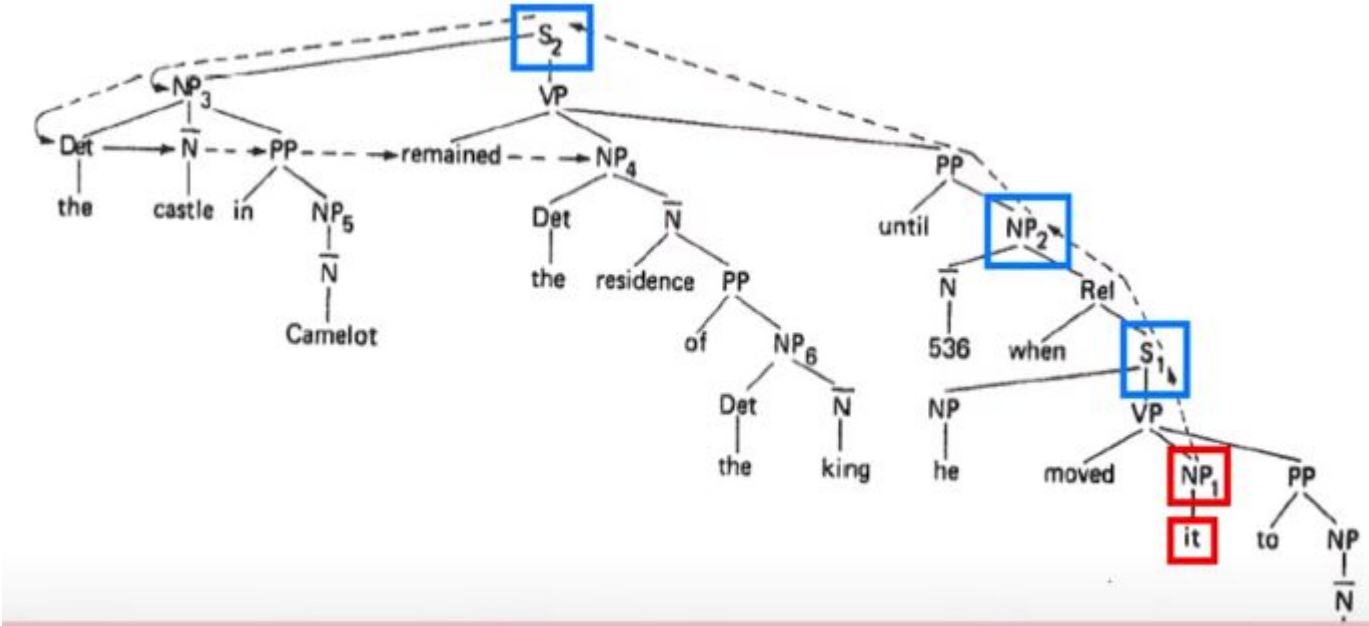
Begin at the NP immediately dominating the pronoun.



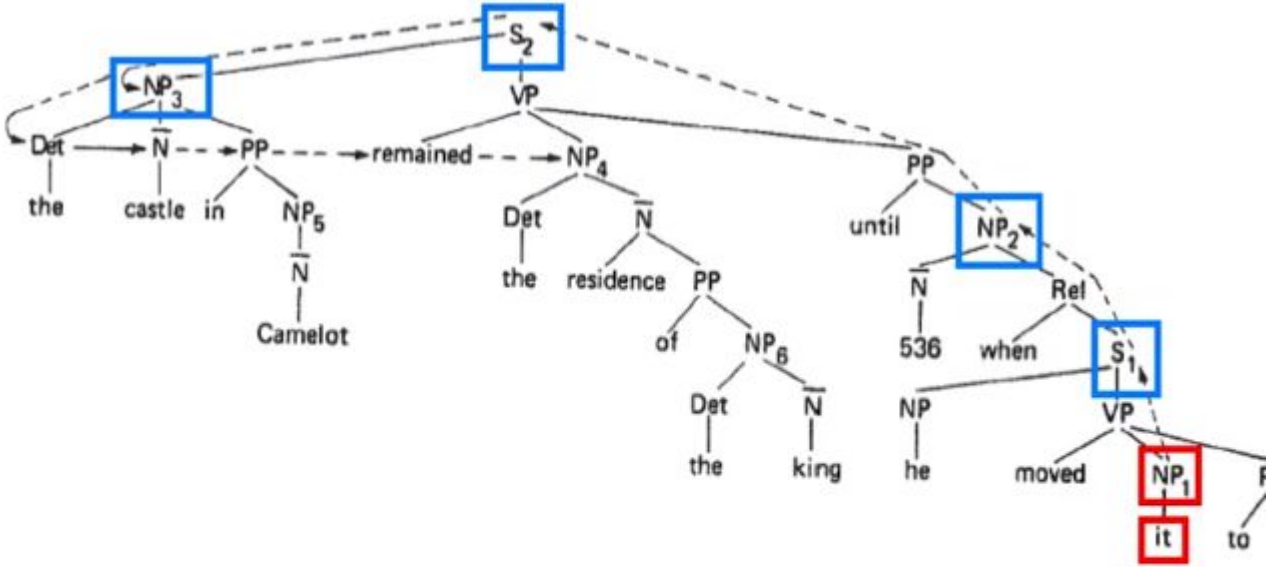
Go up tree to first NP or S encountered. Call node X, and path to it, p.
Search left-to-right below X and to left of p, proposing any NP node which has an NP or S between it and X.
S1 has no candidates.



NP2 is proposed. Rejected by selectional restrictions (dates can't move).

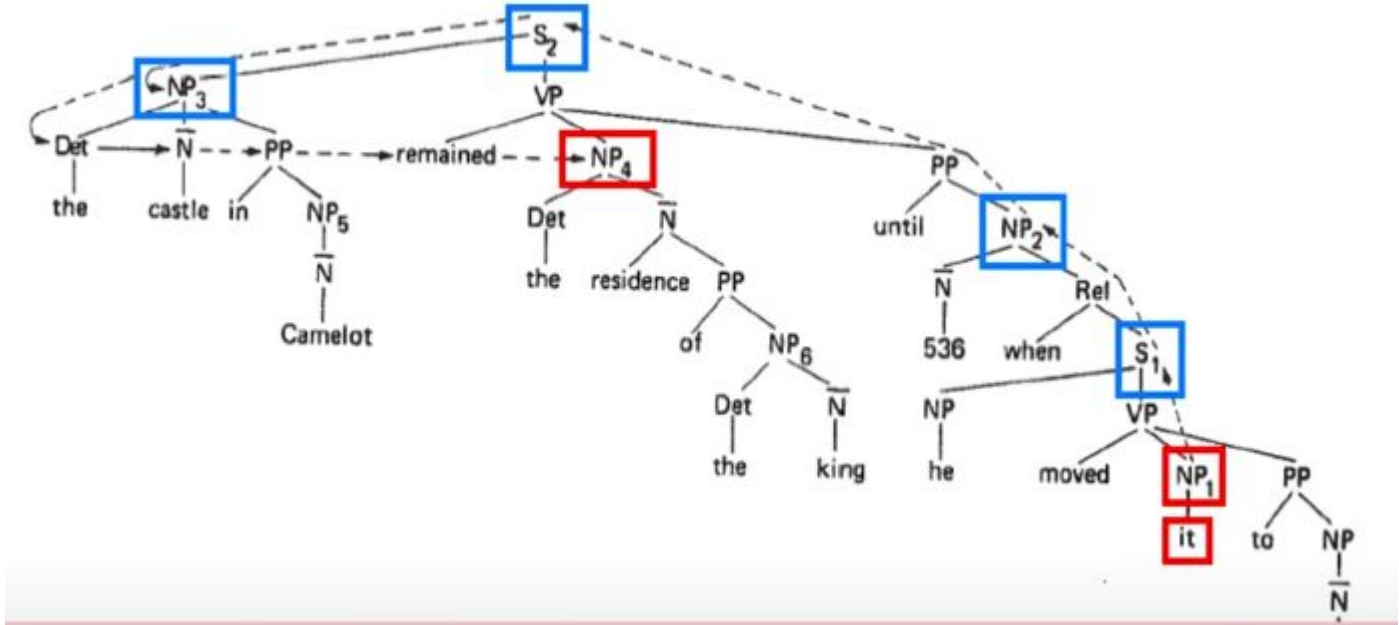


From X, go up to first NP or S node encountered.

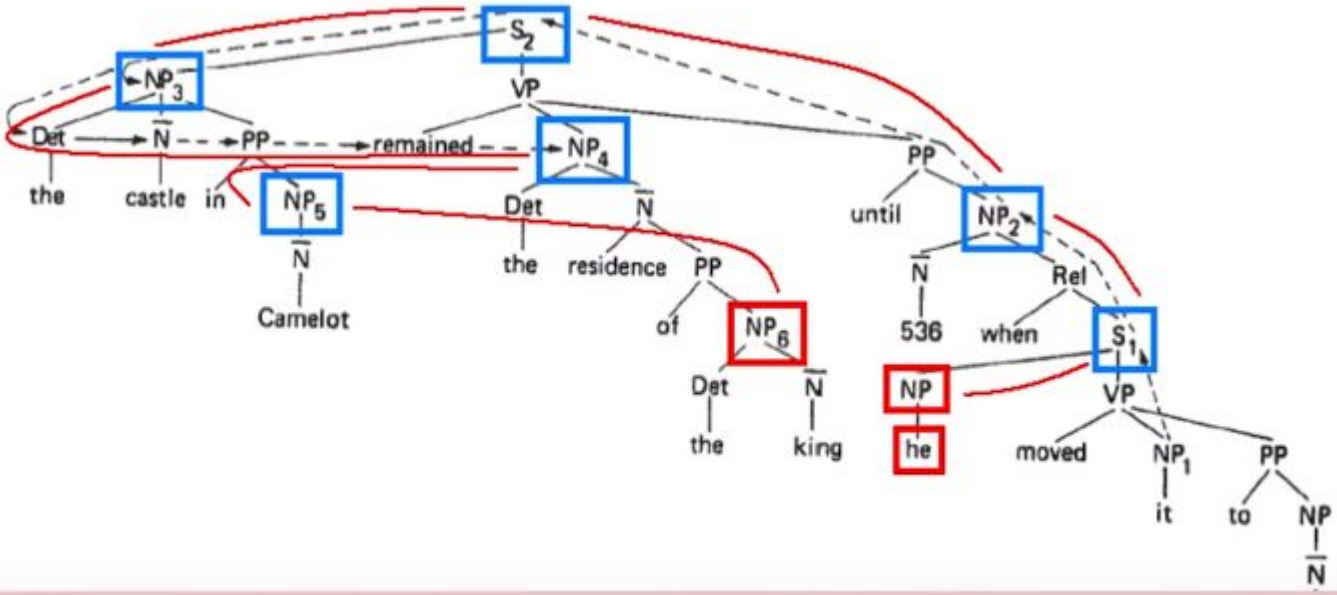


Search left to right.

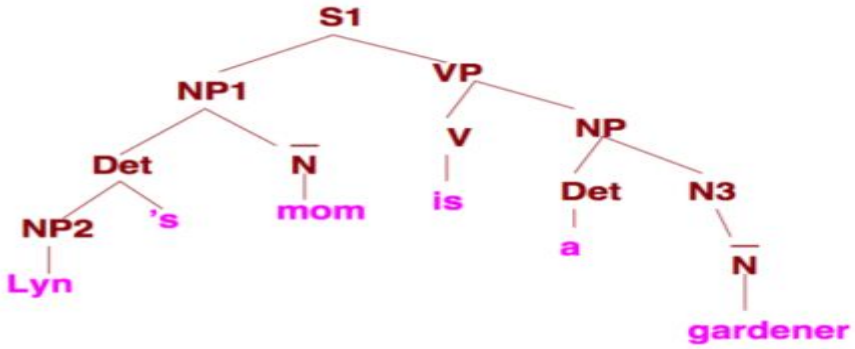
NP3: proposed. Rejected by rejected by selectional restrictions (ca
move large fixed objects.)

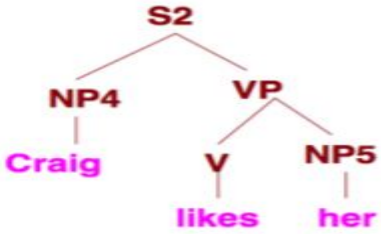


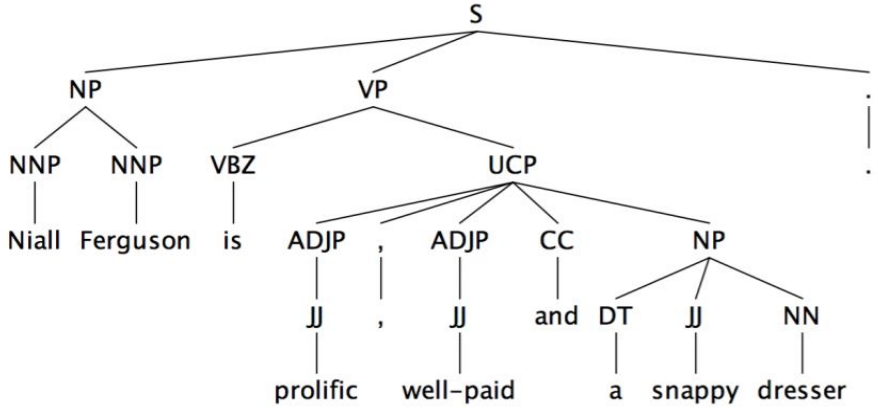
NP4: proposed. Accepted.

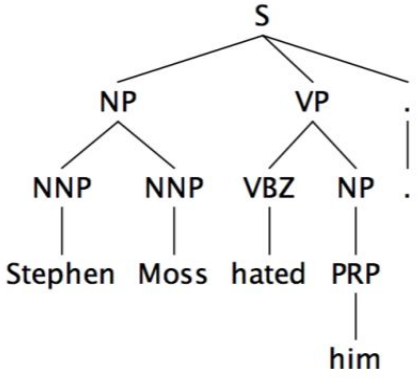


The referent for “he”: we follow the same path, get to the same place, but reject NP4, then reject NP5. Finally, accept NP6.


S1
NP1
VP
Det
N̄
V
NP
NP2
's
mom
is
Det
N3
Lyn
a
N̄
gardener


S2
NP4
VP
Craig
V
NP5
likes
her


S
NP
VP
.
NNP
NNP
VBZ
UCP
.
Niall
Ferguson
is
ADJP
,
ADJP
CC
NP
JJ
,
JJ
and
DT
JJ
NN
prolific
well-paid
a
snappy
dresser


S
NP
VP
.
NNP
NNP
VBZ
NP
.
Stephen
Moss
hated
PRP
him