

Table of Contents

Chapter 1: Introduction to Discourse Analysis	3
---	---

1.1 Beyond the Sentence: What is Discourse?

1.2 The Importance of Discourse in NLP

1.3 Key Concepts: Cohesion and Coherence

Chapter 2: Discourse Segmentation	6
---	---

2.1 Defining Discourse Segments

2.2 Approaches to Segmentation

2.3 TextTiling Algorithm

2.4 Machine Learning Approaches for Segmentation

Chapter 3: Coherence and Coherence Relations	9
--	---

3.1 What is Coherence?

3.2 Rhetorical Structure Theory (RST)

3.3 Centering Theory

3.4 Coherence vs. Cohesion

Chapter 4: Reference Phenomena and Anaphora 12

4.1 Introduction to Reference

4.2 Types of Referring Expressions

4.3 Anaphora, Cataphora, and Exophora

4.4 The Challenge of Anaphora for NLP

Chapter 5: Hobbs' Algorithm for Anaphora Resolution 15

5.1 The Naive Approach and its Limitations

5.2 Hobbs' Algorithm: Intuition and Steps

5.3 Walkthrough Example

5.4 Strengths and Weaknesses

Chapter 6: The Centering Algorithm for Anaphora Resolution 18

6.1 Theoretical Foundations: Centers and Utterances

6.2 The Centering Model: C_p , C_f , C_b

6.3 Constraints and Rules of the Centering Algorithm

6.4 Walkthrough Example

6.5 Comparison with Hobbs' Algorithm

Chapter 7: Coreference Resolution	21
---	----

7.1 From Anaphora to Coreference

7.2 The Coreference Resolution Task

7.3 Mention Detection

7.4 Machine Learning and Neural Approaches

7.5 Evaluation: MUC, B-CUBED, CEAF

Chapter 8: Essential NLP Resources & Corpora	25
--	----

8.1 Lexical Resources: WordNet

8.2 Pre-processing Tools: Porter Stemmer, Lemmatizer, Brill's Tagger

8.3 Syntactic & Semantic Corpora: Penn Treebank, PropBank, FrameNet

8.4 General Corpora: Brown Corpus, British National Corpus (BNC)

Chapter 1: Introduction to Discourse Analysis

1.1 Beyond the Sentence: What is Discourse?

Traditional NLP models often operate at the sentence level. However, human language meaning is constructed across sentences, paragraphs, and entire documents. Discourse is any coherent sequence of utterances—written or spoken—that is longer than a single sentence.

Discourse Analysis is the study of this structure, focusing on how sentences connect to form a meaningful whole. It answers questions like:

How do we interpret pronouns like "he" or "it"?

What is the logical connection between two consecutive sentences?

How does a conversation maintain its topic?

1.2 The Importance of Discourse in NLP

Ignoring discourse structure leads to shallow and often incorrect understanding. Consider these NLP applications:

Machine Translation: Correct pronoun translation depends on knowing the antecedent.

Text Summarization: Identifying the most important sentences requires understanding discourse structure.

Question Answering: Answers often span multiple sentences and require resolving references.

Dialogue Systems: Maintaining a coherent conversation is impossible without tracking discourse state.

1.3 Key Concepts: Cohesion and Coherence

These are two fundamental principles that make a discourse "hang together."

Cohesion: The surface-level, linguistic "glue" that connects sentences. It is realized through specific grammatical and lexical elements.

Example: "John bought a car. It was red." (The pronoun "it" is a cohesive tie).

Coherence: The underlying logical consistency and conceptual "sense" that the discourse makes. It resides in the mind of the reader/listener.

Example: "John bought a car. The sky is blue." These sentences are cohesive (no obvious grammatical errors) but are not coherent because there's no logical connection.

The rest of these notes will explore the mechanisms of cohesion (like reference) and the theories that model coherence.

Chapter 2: Discourse Segmentation

2.1 Defining Discourse Segments

Just as sentences are segments of text, discourses can be broken down into smaller, topically coherent units called discourse segments. A segment is a contiguous sequence of clauses or sentences that are about a common subtopic or are connected by a common rhetorical relation.

Example: A news article can be segmented into: [Lead], [Background], [Main Event], [Consequences], [Conclusion].

2.2 Approaches to Segmentation

There are two primary approaches:

Cohesion-based: Looks for shifts in lexical cohesion. If the vocabulary changes significantly, it might indicate a segment boundary.

Cue-based: Uses discourse markers (also called cue words or connectives) like "however," "in conclusion," "firstly," "on the other hand" to signal boundaries or relationships.

2.3 TextTiling Algorithm

A classic lexical cohesion algorithm for segmenting text. The core idea is to compute a "cohesion score" between adjacent blocks of text and hypothesize a segment boundary where this score reaches a local minimum.

Steps:

Tokenize & Stopword Removal: Clean the text.

Create Token Sequence: [w1, w2, w3, ..., wn].

Define a Pseudosentence: A sliding window of k-words (e.g., k=20).

Compute Lexical Score: For each gap between pseudosentences, compute a similarity score (e.g., cosine similarity) between the vocabulary of the left block and the right block.

Find Boundaries: Smooth the similarity scores and identify the deepest valleys (minima) as candidate segment boundaries.

2.4 Machine Learning Approaches for Segmentation

Modern approaches treat it as a supervised classification task. For each sentence boundary, the model classifies it as either a "segment boundary" or "not a segment boundary."

Features used can include:

Lexical features (word unigrams/bigrams)

Syntactic features (POS tags, parses)

Discourse markers

Entity-based features (change in named entities)

Semantic similarity between adjacent sentences

Chapter 3: Coherence and Coherence Relations

3.1 What is Coherence?

As introduced earlier, coherence is the mental representation of the discourse that makes it meaningful. A coherent discourse is one where the reader can easily infer the logical and functional connections between its parts.

3.2 Rhetorical Structure Theory (RST)

RST is a influential theory that describes the hierarchical, rhetorical structure of text. It posits that clauses in a text are linked by rhetorical relations.

Nucleus and Satellite: A relation holds between a core unit (the Nucleus) and a supporting unit (the Satellite).

Common Relations:

Elaboration: Satellite provides additional detail about the Nucleus.

Explanation: Satellite explains the reason for the Nucleus.

Contrast: Nucleus and Satellite are contrasted.

Condition: Satellite states a condition for the Nucleus.

Sequence: Presents multiple Nuclei in a sequence.

Example:

text

[Nucleus: John loves skiing.]

[Relation: Explanation]

[Satellite: He goes to the Alps every winter.]

3.3 Centering Theory

Centering Theory is a model of the local coherence of discourse and the attentional state of participants. It focuses on how entities are mentioned from one utterance to the next within a discourse segment. We will explore this in great detail in Chapter 6, as it forms the basis for a powerful anaphora resolution algorithm.

3.4 Coherence vs. Cohesion

It is critical to distinguish these concepts finally:

Feature	Cohesion	Coherence
---------	----------	-----------

Nature	Linguistic, surface-level	Cognitive, conceptual
Location	In the text	In the mind of the reader
Mechanism	Grammatical & lexical ties (reference, conjunction)	Inferred logical relations
Example	"John fell. He cried." (cohesive)	"John fell. Therefore, he cried." (cohesive & coherent)

Chapter 4: Reference Phenomena and Anaphora

4.1 Introduction to Reference

Reference is the linguistic process of using expressions to denote an entity in the world or in the discourse. The expression used is called a referring expression, and the entity it refers to is its referent.

4.2 Types of Referring Expressions

Indefinite Noun Phrases: Introduce a new entity into the discourse.

Example: "I saw a cat yesterday."

Definite Noun Phrases: Refer to an entity assumed to be known to the listener.

Example: "The cat was black." (Assumes you know which cat).

Pronouns: Short, reduced forms that are highly context-dependent.

Example: "It ran away."

Demonstratives: "this," "that," "these," "those."

Names: "John," "London," "NASA."

4.3 Anaphora, Cataphora, and Exophora

Anaphora: When a referring expression (the anaphor) points back to a previously mentioned entity (the antecedent).

Example: "Mary arrived. She was happy." ("She" is the anaphor, "Mary" is the antecedent).

Cataphora: When a referring expression points forward to a subsequent mention.

Example: "After he finished work, John went home." ("he" is the cataphor, "John" is the antecedent).

Exophora: When a referring expression points to an entity in the external context, not in the text.

Example: (Pointing) "Look at that!"

4.4 The Challenge of Anaphora for NLP

Resolving anaphora (finding the correct antecedent for an anaphor) is a core task in discourse analysis. It is difficult because:

There can be multiple candidate antecedents.

"John told Bill that he lost his key." (Who does "he" and "his" refer to?)

It requires world knowledge.

"We gave the babies their bottles because they were hungry." (Who is hungry?)

"We gave the babies their bottles because they were made of glass." (What is made of glass?)

Chapter 5: Hobbs' Algorithm for Anaphora Resolution

5.1 The Naive Approach and its Limitations

A naive approach might be to simply pick the most recent noun phrase. This fails frequently, as seen in the "babies/bottles" example. We need a more structured algorithm.

5.2 Hobbs' Algorithm: Intuition and Steps

Hobbs' algorithm (1978) is a simple, syntax-based algorithm for pronominal anaphora resolution. It uses the syntactic parse tree of the sentence to search for the antecedent in a specific order.

Assumptions:

The antecedent and anaphor agree in number and gender.

The algorithm is applied to a parsed text.

Algorithm Steps (Simplified):

Start from the noun phrase (NP) node dominating the pronoun.

Go up the parse tree to the first NP or S (sentence) node encountered. Call this node X, and the path from the pronoun to X, p.

Traverse all branches below node X to the left of path p in a left-to-right, breadth-first fashion. Propose any NP node encountered that has an NP or S between it and X as the antecedent.

If no antecedent is found, move up the tree to the next NP or S node and call it X. The path p is now from the original pronoun to this new X.

Traverse all branches below node X to the left of path p (as in step 3). If X is the root S node, traverse all branches from X to the right of the path p, but do not go below any NP or S encountered.

Repeat steps 4 and 5 until an antecedent is found or the root node is passed.

5.3 Walkthrough Example

Sentence: "John said he did it."

Parse the sentence: (S (NP John) (VP said (S (NP he) (VP did it))))).

Pronoun: "he". Start at its NP node.

Go up to the first S node that contains "he" (the embedded S). This is X.

Traverse left from X: We find the NP "John".

Check agreement: "John" (singular, male) agrees with "he" (singular, male). Antecedent found: John.

5.4 Strengths and Weaknesses

Strengths: Simple, intuitive, requires only a syntactic parser. Surprisingly effective for its simplicity.

Weaknesses: Purely syntactic; fails on semantic or world knowledge constraints (the "babies/bottles" problem). Performance depends entirely on parse tree accuracy.

Chapter 6: The Centering Algorithm for Anaphora Resolution

6.1 Theoretical Foundations: Centers and Utterances

Centering Theory models the local coherence of a discourse by tracking the "center of attention" as the discourse unfolds. It operates within a discourse segment, which is a sequence of utterances ($U_1, U_2, \dots, U_n$).

For each utterance U_n , we define:

$C_f(U_n)$: The forward-looking centers list. This is a ranked list of all entities mentioned in U_n , ordered by their prominence (often based on grammatical role: Subject > Object > Others).

$C_p(U_n)$: The preferred center. This is the highest-ranked element in $C_f(U_n)$, i.e., the most likely candidate to be referred to next.

$C_b(U_n)$: The backward-looking center. This is the entity in U_n that links it to the previous utterance, i.e., the current focus of attention.

6.2 The Centering Model: C_p , C_f , C_b

The state of the discourse is defined by the relationship between $C_b(U_n)$ and $C_b(U_{n-1})$, and between $C_b(U_n)$ and $C_p(U_n)$.

6.3 Constraints and Rules of the Centering Algorithm

Constraints:

There is exactly one C_b per utterance.

Every element in $C_f(U_n)$ must be realized in U_n .

The $C_b(U_n)$ is the highest-ranked element of $C_f(U_{n-1})$ that is realized in U_n .

Rules:

If any element of $C_f(U_{n-1})$ is realized as a pronoun in U_n , then so is $C_b(U_n)$.

Transition States are ordered by coherence:

CONTINUE: $C_b(U_n) = C_b(U_{n-1})$ and $C_b(U_n) = C_p(U_n)$. (Most coherent)

RETAIN: $C_b(U_n) = C_b(U_{n-1})$ but $C_b(U_n) \neq C_p(U_n)$.

SMOOTH-SHIFT: $C_b(U_n) \neq C_b(U_{n-1})$ but $C_b(U_n) = C_p(U_n)$.

ROUGH-SHIFT: $C_b(U_n) \neq C_b(U_{n-1})$ and $C_b(U_n) \neq C_p(U_n)$. (Least coherent)

6.4 Walkthrough Example

Discourse:

U1: John went to the store.

U2: He bought milk.

U3: It was expensive.

Step 1 - U1:

Cf(U1): {John, store} [Subject > Object]

Cp(U1): John

Cb(U1): undefined (no previous utterance)

Step 2 - U2:

Cf(U2): {He, milk}

Cp(U2): He

Find Cb(U2): The highest-ranked element from Cf(U1) that is in U2 is "John". So Cb(U2) = John.

Check Rule 1: "John" is realized as the pronoun "He". Good.

Transition: Cb(U2) = Cb(U1)? N/A. But Cb(U2) = John = Cp(U1). This is a CONTINUE-like transition.

Step 3 - U3:

Cf(U3): {It}

Cp(U3): It

Find $Cb(U3)$: The highest-ranked element from $Cf(U2)$ that is in $U3$ is "milk". So $Cb(U3) = \text{milk}$.

Transition: $Cb(U3) (\text{milk}) \neq Cb(U2) (\text{John})$. But $Cb(U3) = Cp(U3) (\text{It})$. This is a SMOOTH-SHIFT.

Anaphora Resolution: In $U3$, "It" is resolved to "milk".

6.5 Comparison with Hobbs' Algorithm

Hobbs is a search procedure for a single pronoun.

Centering is a model of discourse coherence that simultaneously resolves all pronouns in an utterance and predicts the coherence of the discourse. It is more powerful but also more complex to implement.

Chapter 7: Coreference Resolution

7.1 From Anaphora to Coreference

Anaphora Resolution is a subset of Coreference Resolution.

Anaphora Resolution: Finds the antecedent for a single anaphoric expression (like a pronoun).

Coreference Resolution: Finds all expressions in a text that refer to the same real-world entity. This creates coreference chains.

Example Coreference Chain:

$[\text{John Smith}]_1 \rightarrow [\text{He}]_1 \rightarrow [\text{The CEO}]_1 \rightarrow [\text{Smith}]_1$

7.2 The Coreference Resolution Task

The task is to cluster all mentions in a text that refer to the same entity.

Input: "John arrived. He sat down. Mary saw him. She smiled."

Output: Coreference Chains:

Chain 1: {John, He, him}

Chain 2: {Mary, She}

7.3 Mention Detection

The first step is to identify all potential mentions in the text. Mentions can be:

Pronouns (he, she, it, they)

Named Entities (John, London)

Noun Phrases (the man, a big company)

Definite Noun Phrases (the CEO)

7.4 Machine Learning and Neural Approaches

Modern coreference resolution is dominated by machine learning models.

The Mention-Pair Model: A binary classifier for every pair of mentions: "Do these two mentions refer to the same entity?"

The Entity-Centric Model: Builds clusters incrementally. For a new mention, decide which existing cluster it belongs to, or if it starts a new cluster.

End-to-End Neural Coreference (Lee et al., 2017): A state-of-the-art approach that jointly learns mention detection and coreference resolution using span representations (representations for every possible text span).

7.5 Evaluation: MUC, B-CUBED, CEAF

Evaluating coreference is complex because it's a clustering task.

MUC (Vilain et al., 1995): Measures the number of links needed to align the system's chains with the gold-standard chains. Prone to over-rewarding systems that produce large clusters.

B-CUBED (Bagga and Baldwin, 1998): Computes precision and recall for each mention and then averages. More balanced.

CEAF (Luo, 2005): Uses an alignment between system and gold clusters based on entity similarity.

Chapter 8: Essential NLP Resources & Corpora

This chapter provides a reference for the key tools and datasets mentioned.

8.1 Lexical Resources: WordNet

What it is: A large lexical database of English. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept.

Use in Discourse: Used for computing semantic similarity between words, which can be a feature in coreference resolution (e.g., "car" and "vehicle" are similar) or segmentation.

8.2 Pre-processing Tools

Porter Stemmer: A simple, rule-based algorithm for stemming (crudely reducing words to a root form). "running" -> "run".

Lemmatizer: A more sophisticated tool that uses a vocabulary and morphological analysis to return the base dictionary form (lemma). "running" -> "run", "better" -> "good". (e.g., from NLTK or spaCy).

Brill's Tagger: A transformation-based rule-learning tagger. It starts with a simple baseline (e.g., unigram tagger) and learns a set of context-sensitive rules to correct mistakes. "He can can a can." -> PRP/MD/VB/DT/NN/.

8.3 Syntactic & Semantic Corpora

Penn Treebank: A corpus of over 4.5 million words of American English, annotated with part-of-speech tags and syntactic parse trees. Essential for training parsers used by algorithms like Hobbs'.

PropBank (Proposition Bank): A corpus annotated with verbal propositions and their arguments (Who did what to whom). "John gave Mary the book" is tagged as [Arg0: John] [Rel: gave] [Arg1: the book] [Arg2: Mary].

FrameNet: A corpus based on Frame Semantics. It annotates texts against "frames" (script-like scenarios like Commerce_buy), which have frame-specific roles (Buyer, Seller, Goods).

8.4 General Corpora

Brown Corpus: The first million-word electronic corpus of English, compiled in 1961. Contains text from 500 sources across 15 genres. Used for empirical study of language.

British National Corpus (BNC): A 100-million-word corpus of modern English, both spoken and written. A standard resource for linguistic research and tool development