Carnegie Mellon University
Language
Technologies
Institute

11-324/11-624/11-724 Human Language for AI

Introduction to Discourse

David R. Mortensen

November 15, 2021

Language Technologies Institute
Carnegie Mellon University

Homework Assignment Road Map

There *were* two remaining homework assignments:

1. Homework 6: Discourse Segmentation
    - Use linguistic features to add paragraph boundaries back into Wikipedia articles from which they have been eliminated
    - Supervised sequence labeling task (labeling sentences as first in segment (B) or internal to segment (I))
2. Homework 7: Variationist Sociolinguistic Study
    - Identify a data set in which you can identify both social and linguistic variables (such as a labeled social media set)
    - Select one or more pairs of social variables and linguistic variables that you think may be related
    - Test whether there is a statistically significant relationship between them.

We will **not** have time to do both. Which would you prefer to do?

Learning Objectives

At the end of this lecture, students will understand the importance of discourse structure in language use.

- The will recognize that context is **fundamental** (not ancillary) to language
- They will understand the difference between cohesion and coherence

- Cohesion
  - What formal factors contribute to cohesion
  - How cohesion relates to discourse segmentation
- Coherence
  - What kinds of meaning relations are involved in coherence
  - How coherence relates to discourse parsing

Introduction

Discourse Matters

Consider the following sentences:

- However, the incompetence of his managers insures him a steady, six-figure income.
- He only knows Java.
- Worse still, he always optimizes the outermost loop first.
- Eric is a pathetic programmer.

Discourse Matters

There are $4! = 24$ ways to permute these sentences, but only 2 of them are acceptable as discourses:

1. Eric is a pathetic programmer.
2. He only knows Java.
3. Worse still, he always optimizes the outermost loop first.
4. However, the incompetence of his managers insures him a steady, six-figure income.

and

1. Worse still, he always optimizes the outermost loop first.
2. Eric is a pathetic programmer.
3. However, the incompetence of his managers insures him a steady, six-figure income.
4. He only knows Java.

Discourse Matters

Contrast this, for example, with the following:

1. Eric is a pathetic programmer.
2. Worse still, he always optimizes the outermost loop first.
3. However, the incompetence of his managers insures him a steady, six-figure income.
4. He only knows Java.

Or this:

1. However, the incompetence of his managers insures him a steady, six-figure income.
2. Worse still, he always optimizes the outermost loop first.
3. Eric is a pathetic programmer.
4. He only knows Java.

Language is Situated at All Levels

All language is situated in a context:

- A phonological context (sounds before and after)
- A morphosyntactic context (structurally-related morphemes/words)
- A spatial and temporal context (what words like *here* and *now* point to)
- A discourse context (other sentences in a paragraph, conversation, speech, etc.)
- A social context
  - The purpose of an utterance (what the utterance **does**)
  - The identity of the speaker and hearer
  - The common ground between speaker and hearer

Language Has Structure above the Level of the Sentence

Conversations, paragraphs, poems, lectures, and books can seem amorphous and unstructured. However, they actually have an *intricate* and *interleaved* structure that allows them to function as social media. This structure (discourse) goes all the way up to Discourse—the broad social narratives that underlay much of our interaction as humans.

Cohesion and Discourse Segmentation

Cohesion

- Cohesion is closely related to “topic.”
- Utterances or sentences that are about the same topic typically show a high degree of cohesion (if they are neighbors).
- Cohesion and “topic” are hierarchical
    - Sentences on the same topic are often grouped together as paragraphs
    - Paragraphs that are more closely related to one another than to surrounding paragraphs may be grouped into subsections
    - Topically related subsections may be grouped into sections
    - Documents tend to have a single overarching topic that means the sections in a document are more topically related to one another than to sections from a random document
- While measures of cohesion do not necessarily match the document structure imposed by an author or editor through structural formatting, they do correlate with them and they do share the same tree-like patterning, at least to some degree.

What does it mean to ramble?

Factors in Cohesion

There are a number of different formal devices that contribute to cohesion. These include the following:

- **Lexical chains**. The use of the same word or similar words in utterances that are near one another.

  (1) I didn't say he was a poor coder. I said he was a bad programmer.

- **Anaphora chains** (or more generally, **coreference chains**). The use of coreference between pronouns or other expressions.

  (2) I found the bug in your code. It was an off-by-one error.

- **Discourse markers** (or **cue words**). Expressions or words, often conjunctions, that signal continuity in topic or a change in topic explicitly.

  (3) George had confidence. However, he lacked gravitas.

Segmenting Discourse

How would you divide a discourse into segments (by topic, for example)? One approach:

Segmenting Discourse

How would you divide a discourse into segments (by topic, for example)? One approach:

- Extract features for each potential boundary
  - Lexical chains crossed
  - Coreference chains crossed
  - Discourse markers

Segmenting Discourse

How would you divide a discourse into segments (by topic, for example)? One approach:

- Extract features for each potential boundary
  - Lexical chains crossed
  - Coreference chains crossed
  - Discourse markers
- Label as boundaries points where lexical chain crossings, co-reference chain crossings, and discourse makers indicating continuity are minimal and where discourse markers indicating discontinuity are maximal

Segmenting Discourse

How would you divide a discourse into segments (by topic, for example)? One approach:

- Extract features for each potential boundary
  - Lexical chains crossed
  - Coreference chains crossed
  - Discourse markers
- Label as boundaries points where lexical chain crossings, co-reference chain crossings, and discourse makers indicating continuity are minimal and where discourse markers indicating discontinuity are maximal
- This can be done heuristically in a rule-based framework

Segmenting Discourse

How would you divide a discourse into segments (by topic, for example)? One approach:

- Extract features for each potential boundary
  - Lexical chains crossed
  - Coreference chains crossed
  - Discourse markers
- Label as boundaries points where lexical chain crossings, co-reference chain crossings, and discourse makers indicating continuity are minimal and where discourse markers indicating discontinuity are maximal
- This can be done heuristically in a rule-based framework
- This task can also be operationalized as a supervised sequence labeling task

Coherence and Discourse Parsing

Coherence

Compare the following examples:

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

b. John hid Bill's car keys. He likes spinach.

Result Infer that the state or event asserted by S_0 causes or could cause the state or event asserted by S_1.
Lakshmi had an original thought. She was fired.

Explanation Infer that the state or event asserted by S_1 causes or could cause the state or event asserted by S_0.
Lakshmi was fired. She had had an original thought.

Parallel Infer $p(a_1, a_2, ...)$ from the assertion of S_0 and $p(b_1, b_2, ...)$ from the assertion of S_1, where a_i and b_i are similar, for all i.
Lakshmi was fired; Krishna was promoted.

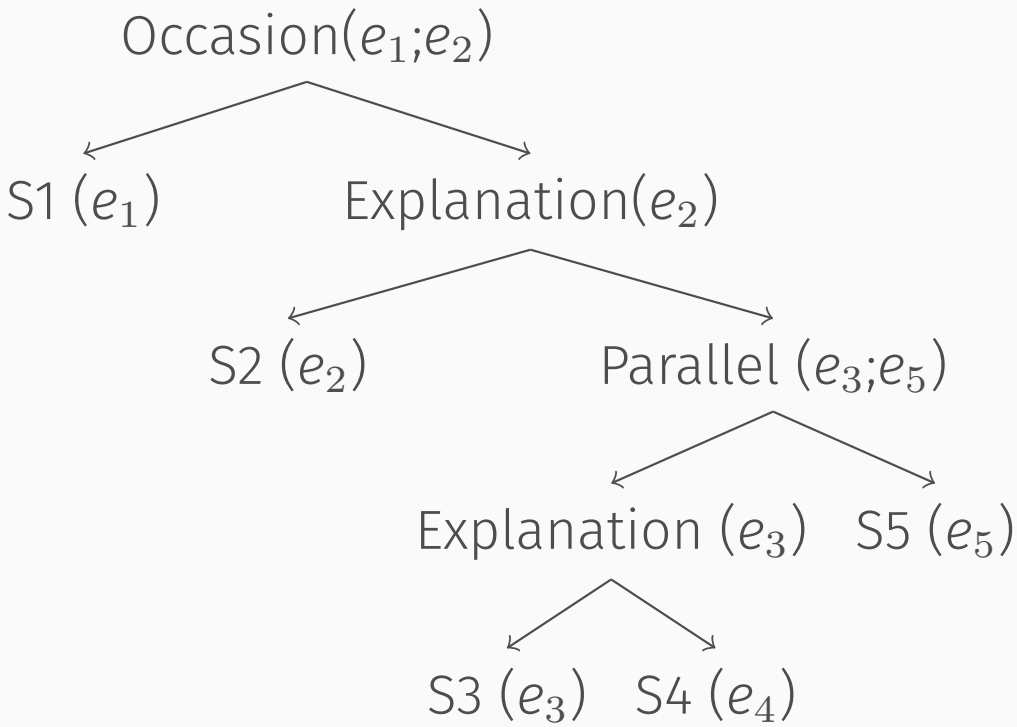
Elaboration Infer the same proposition P from the assertions of S_0 and S_1.
Lakshmi was fired. Her employment was terminated abruptly and without warning.

Occasion A change of state can be inferred from the assertion of S_0, whose final state can be inferred from S_1, or a change of state can be inferred from the assertion of S_1, whose initial state can be inferred from S.
Xudong opened a text editor. He starting pounding out a script similar to those he had written a thousand times before.

A Discourse for Analysis

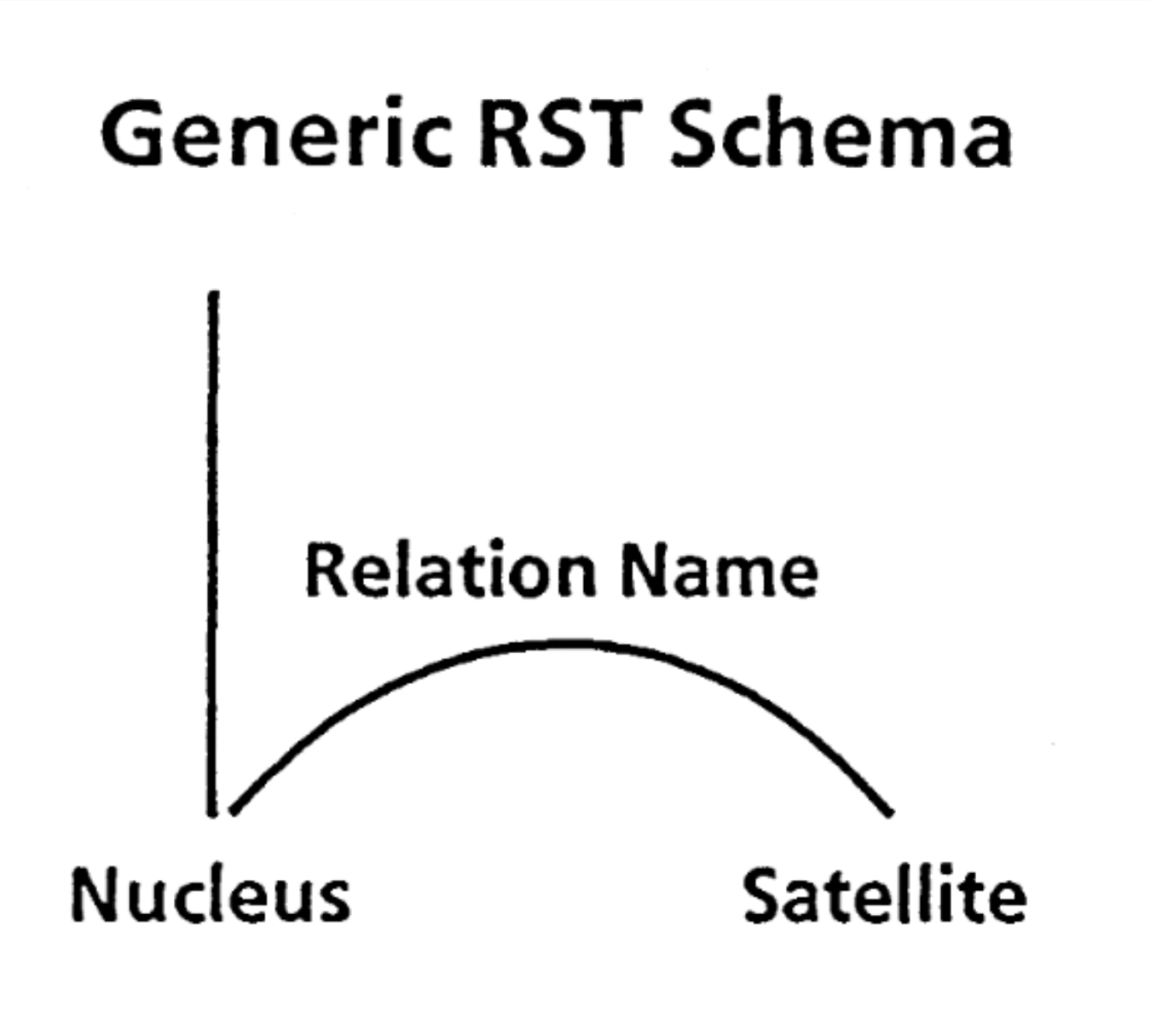
John went to the bank to deposit his paycheck. (S1)
He then took a train to Bill's car dealership. (S2)
He needed to buy a car. (S3)
The company he works for now isn't near any public transportation. (S4)
He also wanted to talk to Bill about their softball league. (S5)

Discourse Parsing



John went to the bank to deposit his paycheck. (S1)
He then took a train to Bill's car dealership. (S2)
He needed to buy a car. (S3)
The company he works for now isn't near any public transportation. (S4)
He also wanted to talk to Bill about their softball league. (S5)

Rhetorical Structure Theory (RST)



Constraints on an RST Relation

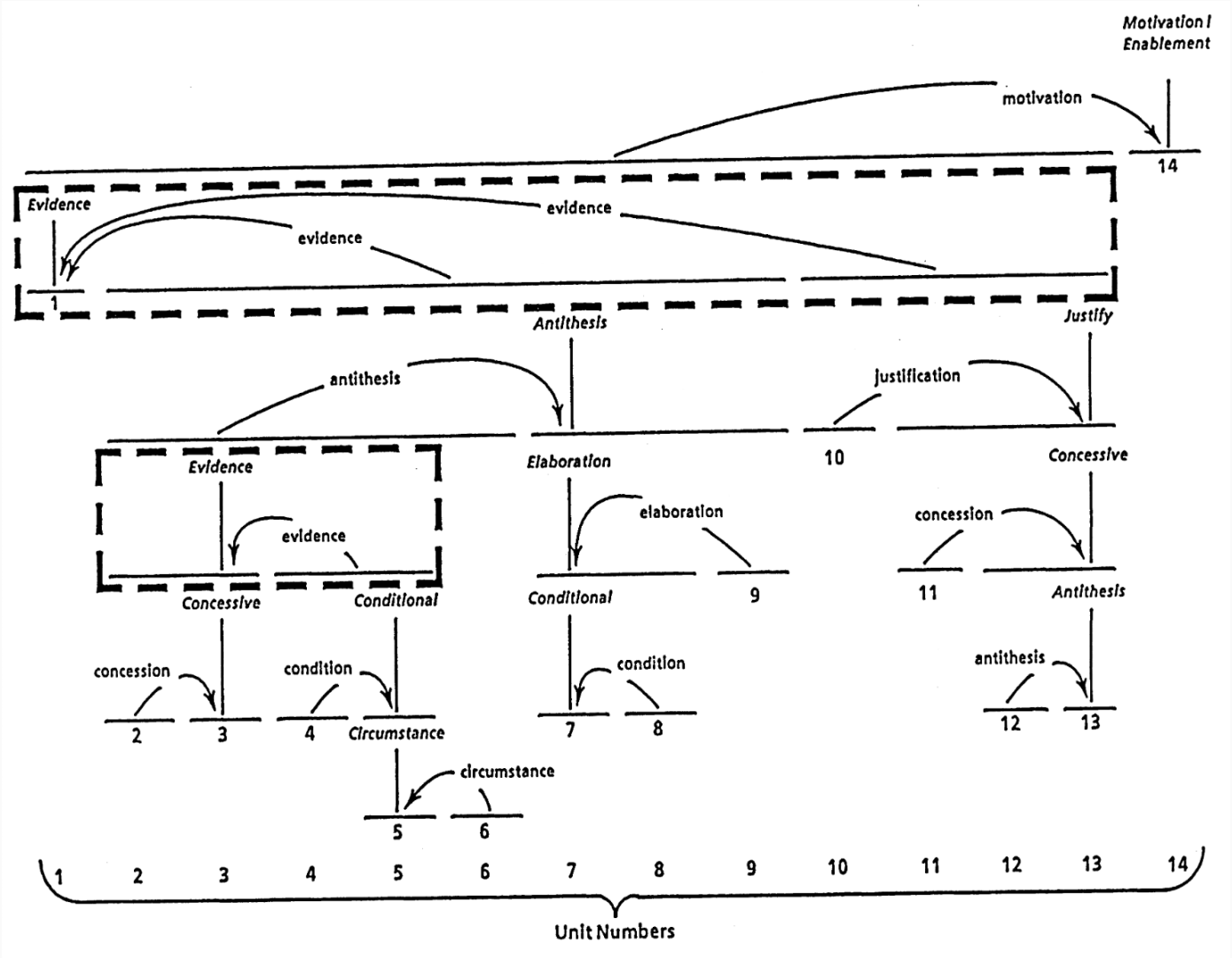
An RST relation has four parts:

1. Constraints on the Nucleus
2. Constraints on the Satellite
3. Constraints on the combination of Nucleus and Satellite
4. The Effect

Take the Evidence relation as an example:

1. **Constraints on the Nucleus (the claim)**: the reader possibly does not already believe the claim.
2. **Constraints on the Satellite (the evidence)**: The reader either already believes the satellite or will find it credible.
3. **Constraints on the combination of the nucleus and satellite**: Comprehending the evidence will increase the reader's belief in the claim
4. **The Effect**: the reader's belief in the claim is increased

Rhetorical Structure Theory (RST)



RST diagram of an advocacy text

Comparing Hobbs (1979) and RST

- Hobbs-style discourse parsing results in a constituency tree (like a phrase-structure parse)
- RST discourse parsing results in a kind of dependency tree (with labeled nucleus-satellite relations)
- Hobbs identified many discourse relations; RST posits many more (64, in total)

Conclusion

Cohesion and discourse segmentation: **Form**

Coherence and discourse parsing: **Function**

Questions?