Discourse Coherence

Natalie Parde, Ph.D.

Department of Computer Science

University of Illinois at Chicago

CS 521: Statistical Natural Language Processing

Spring 2020

Many slides adapted from Jurafsky and Martin (https://web.stanford.edu/~jurafsky/slp3/).

What is discourse coherence?

- The relationship (or lack thereof) between sentences in a **discourse**

I really like my class, CS 521. UIC is in Chicago. It's about statistical natural language processing.

UIC is in Chicago, and I'm taking a class there called CS 521. I really like the class. It's about statistical natural language processing.

What counts as a discourse?

- Any structured, collocated group of sentences
  - Chapter of a book
  - News article
  - Conversation
  - Twitter thread
  - Wikipedia page
  - Etc.
- Discourses are coherent, rather than random combinations of sentences

What makes a discourse coherent?

- A variety of local and global factors
  - Relations between text units
  - Inferential continuity
  - Entity salience
  - Topical salience
  - Discourse structure

I really like my class, CS 521. **UIC is in Chicago.** 😕 **It's** 😕 about statistical natural language processing.

UIC is in Chicago, **and I'm taking a class there** 🙂 called CS 521. I really like **the class** 🙂. **It's** 🙂 about statistical natural language processing.

Why do we care whether a discourse is coherent?

- Measuring discourse coherence is important for measuring the quality of a given text
- Many useful applications:
  - Automated essay grading
  - Determining which sentences to include in automatically-generated summaries
  - Measuring mental or cognitive health

So, modeling discourse coherence is very important!

- How do we do it?
  - Some key techniques:
    - Identify coherence relations
    - Determine entity salience
    - Measure lexical cohesion
    - Identify argument structure

Coherence Relations

- Connections between spans of text in a discourse
- Two commonly-used models:
  - **Rhetorical Structure Theory (RST)**
  - **Penn Discourse Treebank (PDTB)**

Rhetorical Structure Theory

- Based on a set of 23 **rhetorical relations** that can hold between spans of text within a discourse
- Most relations are between two spans:
  - **Nucleus**
    - More central to the writer's purpose
    - Interpretable independently
  - **Satellite**
    - Less central to the writer's purpose
    - Only interpretable with respect to the nucleus

Rhetorical Structure Theory

- Relations are **asymmetric**
  - Represented graphically with arrows pointing from the satellite to the nucleus
- Relations are defined by a **set of constraints** on the nucleus and satellite
- Constraints are based on:
  - **Goals and beliefs** of the writer and reader
  - **Effect** on the reader

Common RST Relations

Elaboration	Satellite gives further information about the content of the nucleus
Attribution	Satellite gives the source of attribution for an instance of reported speech in the nucleus
Contrast	Two or more nuclei contrast along some important dimension
List	A series of nuclei is given, without contrast or explicit comparison
Reason	Satellite provides the reason for the action carried out in the nucleus
Evidence	Satellite provides information with the [illegible] accept the information provided in the

Natalie told the class that there were no paper critiques due next week, reminding them that there was an exam instead.

Common RST Relations

Elaboration	Satellite gives further information about the content of the nucleus
Attribution	Satellite gives the source of attribution for an instance of reported speech in the nucleus
Contrast	Two or more nuclei contrast along some important dimension
List	A series of nuclei is given, without contrast or explicit comparison
Reason	Satellite provides the reason for the action carried out in the nucleus
Evidence	Satellite provides information with the goal of convincing the reader to accept the information provided in the nucleus

Natalie told the class that there were no paper critiques due next week.

Common RST Relations

Elaboration	Satellite gives further information about the content of the nucleus
Attribution	Satellite gives the source of attribution for an instance of reported speech in the nucleus
Contrast	Two or more nuclei contrast along some important dimension
List	A series of nuclei is given, without contrast or explicit comparison
Reason	Satellite provides the reason for the action carried out in the nucleus
Evidence	Satellite provides information with the goal of convincing the reader to accept the information provided in the nucleus

Outside was freezing, but **inside was uncomfortably warm**.

Common RST Relations

Elaboration	Satellite gives further information about the content of the nucleus
Attribution	Satellite gives the source of attribution for an instance of reported speech in the nucleus
Contrast	Two or more nuclei contrast along some important dimension
List	A series of nuclei is given, without contrast or explicit comparison
Reason	Satellite provides the reason for the action carried out in the nucleus
Evidence	Satellite provides information with the goal that the reader will believe or accept the information provided in the nucleus

In the fall, Natalie taught CS 421; in the spring, Natalie taught CS 521; in the summer, Natalie worked on research.

Common RST Relations

Elaboration	Satellite gives further information about the content of the nucleus
Attribution	Satellite gives the source of attribution for an instance of reported speech in the nucleus
Contrast	Two or more nuclei contrast along s
List	A series of nuclei is given, without c
Reason	Satellite provides the reason for the action carried out in the nucleus
Evidence	Satellite provides information with the goal of convincing the reader to accept the information provided in the nucleus

Natalie got to campus earliest on Tuesdays and Thursdays. She had to teach a class at 9:30 a.m.

Common RST Relations

Elaboration	Satellite gives further information about the content of the nucleus
Attribution	Satellite gives the source of attribution for an instance of reported speech in the nucleus
Contrast	Two or more nuclei contrast along s
List	A series of nuclei is given, without contrast or explicit comparison
Reason	Satellite provides the reason for the action carried out in the nucleus
Evidence	Satellite provides information with the goal of convincing the reader to accept the information provided in the nucleus

Natalie must be here. Her office door is cracked open.

RST relations can be hierarchically organized into discourse trees.

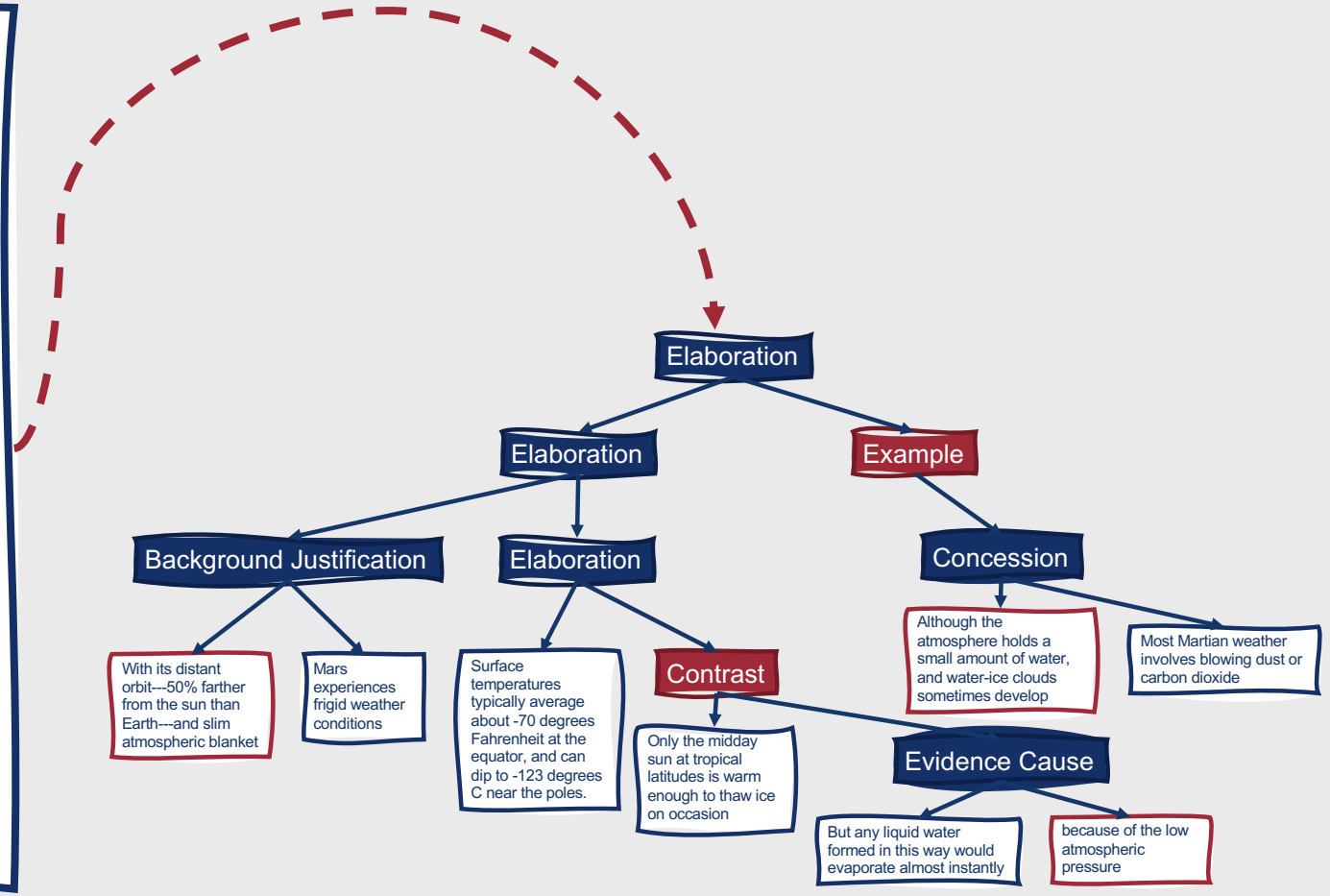
With its distant orbit-–50% farther from the sun than Earth-–and slim atmospheric blanket, Mars experiences frigid weather conditions. Surface temperatures typically average about -70 degrees Fahrenheit at the equator, and can dip to -123 degrees C near the poles.

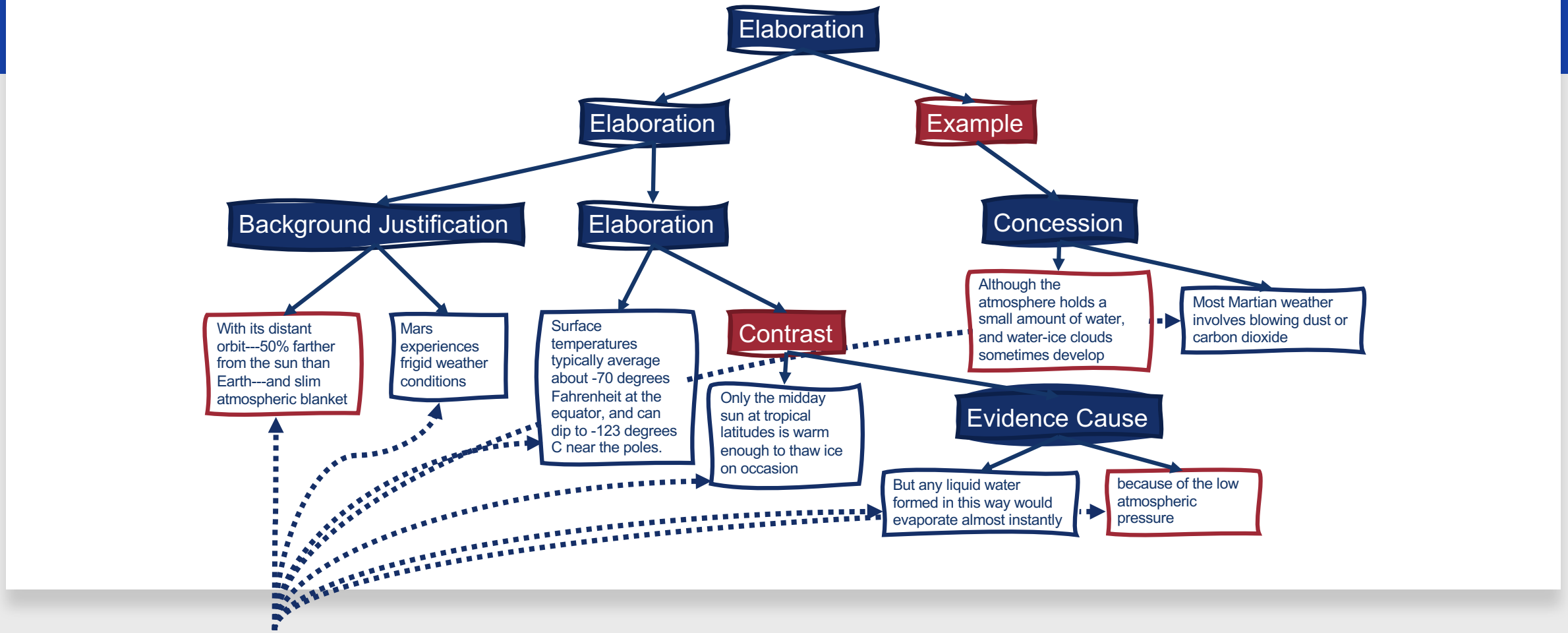
Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion, but any liquid water formed in this way would evaporate almost instantly because of the low atmospheric pressure. Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop, most Martian weather involves blowing dust or carbon dioxide.

Example Discourse Tree

With its distant orbit-–50% farther from the sun than Earth-–and slim atmospheric blanket, Mars experiences frigid weather conditions. Surface temperatures typically average about -70 degrees Fahrenheit at the equator, and can dip to -123 degrees C near the poles.

Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion, but any liquid water formed in this way would evaporate almost instantly because of the low atmospheric pressure. Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop, most Martian weather involves blowing dust or carbon dioxide.





Elementary Discourse Units (EDUs)

- Leaves in a discourse tree
  - Also referred to as discourse segments
- Determining the boundaries of EDUs is important for extracting coherence relations

RST Corpora

RST Discourse Treebank

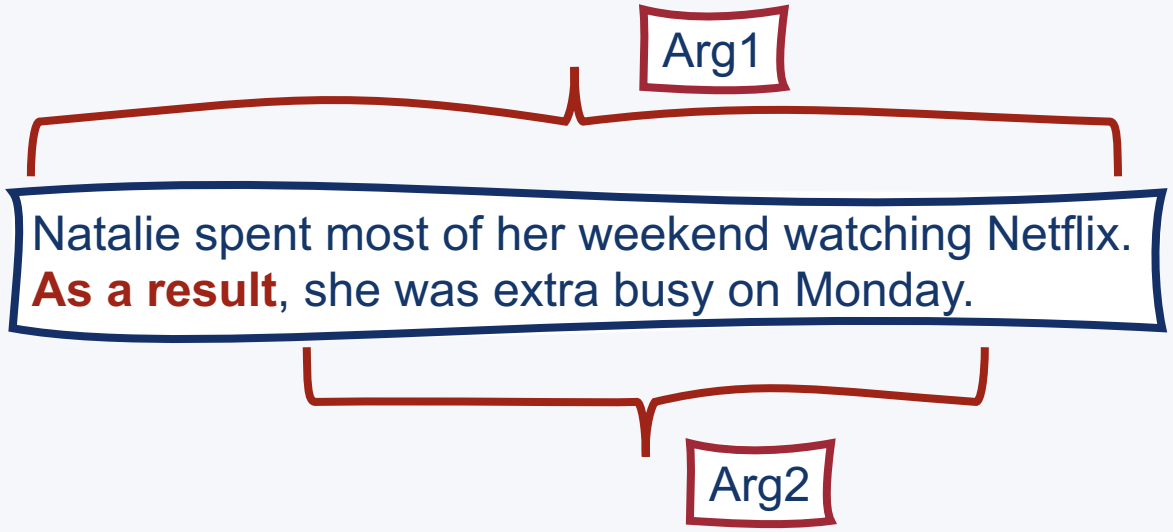
- 385 English-language documents with full RST parses
- 78 distinct relations, groups into 16 classes
- https://catalog.ldc.upenn.edu/LDC2002T07

Similar Non-English Corpora:

- CST-News (Brazilian Portuguese): http://nilc.icmc.usp.br/CSTNews/login/?next=/CSTNews/
- Rhetalho and CorpusTCC (Brazilian Portuguese): https://sites.icmc.usp.br/taspardo/Projects.htm
- Spanish RST DT (Spanish): http://corpus.iingen.unam.mx/rst/index_en.html
- Potsdam Commentary Corpus (German): http://angcl.ling.uni-potsdam.de/resources/pcc.html
- Basque RST DT (Basque): http://ixa2.si.ehu.es/diskurtsoa/en/

Penn Discourse Treebank

- **Lexically-grounded** model of coherence relations
  - Given a list of **discourse connectives** (e.g., *because*, *although*, *when*, *since*, or *as a result*) and an unlabeled document, annotators labeled:
    - Those connectives
    - The spans of text that they connected
  - In some cases, these connectives may be implicit



PDTB Sense Hierarchy

- Four main classes:
  - Temporal
  - Contingency
  - Comparison
  - Expansion
- Numerous subtypes of each

PDTB Annotations

- Only at the span-pair level!
- No hierarchical tree structure

PDTB Corpus

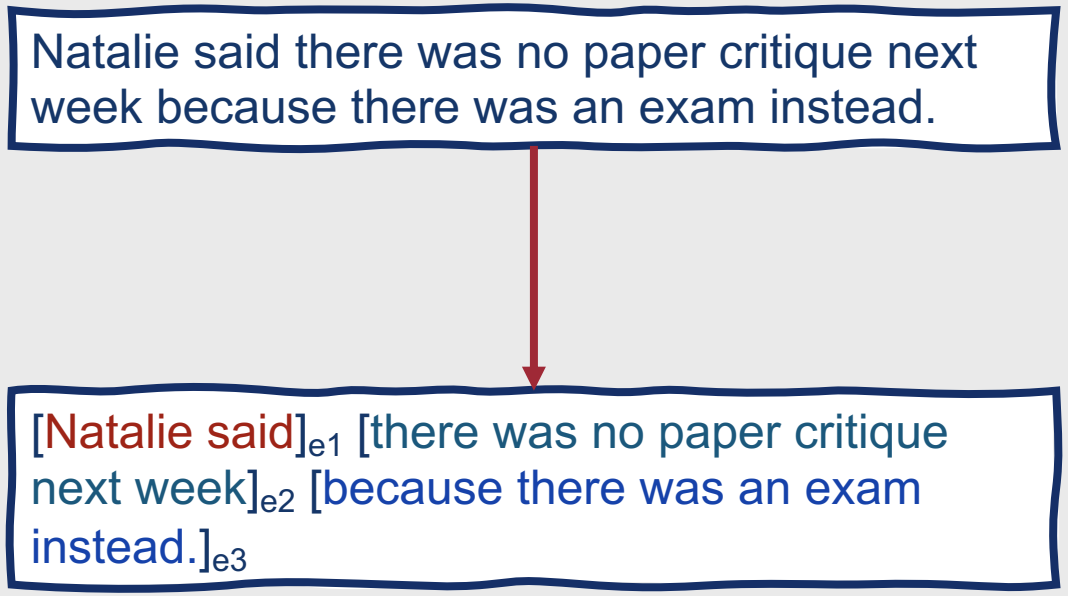
- 50k+ of annotated relations
- Built on top of the Wall Street Journal section of the Penn Treebank
- https://catalog.ldc.upenn.edu/LDC2019T05

Given a specified discourse model (e.g., RST), how do we automatically assign discourse relations to text?

- **Discourse structure parsing:** Given a sequence of sentences, automatically determine the coherence relations between them
- Discourse structure parsing can be performed similarly to constituency parsing
  - Break text into meaningful subunits
  - Organize those subunits into a set of directed (and, depending on model type, hierarchical) relations

What does this look like for RST parsing?

- **Step #1: EDU Segmentation**
  - Extract the start and end of each elementary discourse unit



EDU Segmentation

- EDUs roughly correspond to clauses
- Early EDU segmentation approaches:
  - Run a syntactic parser
  - Post-process the output
- More modern EDU segmentation approaches:
  - Usually, apply supervised neural sequence models

What does this look like for RST parsing?

- **Step #1: EDU Segmentation**
  - Extract the start and end of each elementary discourse unit
- **Step #2: Parsing Algorithm**
  - Build representations for each EDU, and apply some method to connect them using RST relations

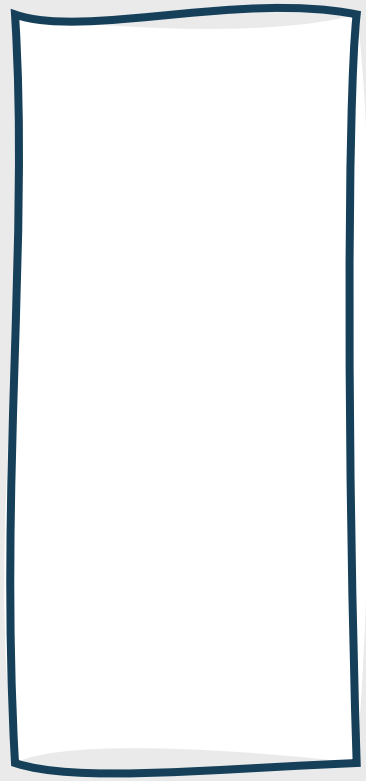
RST Parsing

- Generally based on syntactic parsing algorithms
- Common syntactic parsing approach: **Shift-reduce parser**
  - **Shift:** Push an EDU from the queue onto the stack, creating a single-node subtree
  - **Reduce:** Merge the top two subtrees (either single-node or more complex) on the stack, assigning a coherence relation label and a nuclearity direction
  - **Pop:** Remove the final tree from the stack

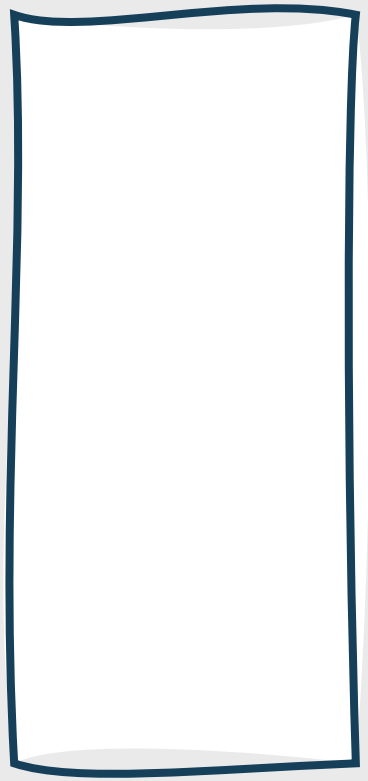
Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no paper critique next week]$_{e2}$ [because there was an exam instead.]$_{e3}$

Queue

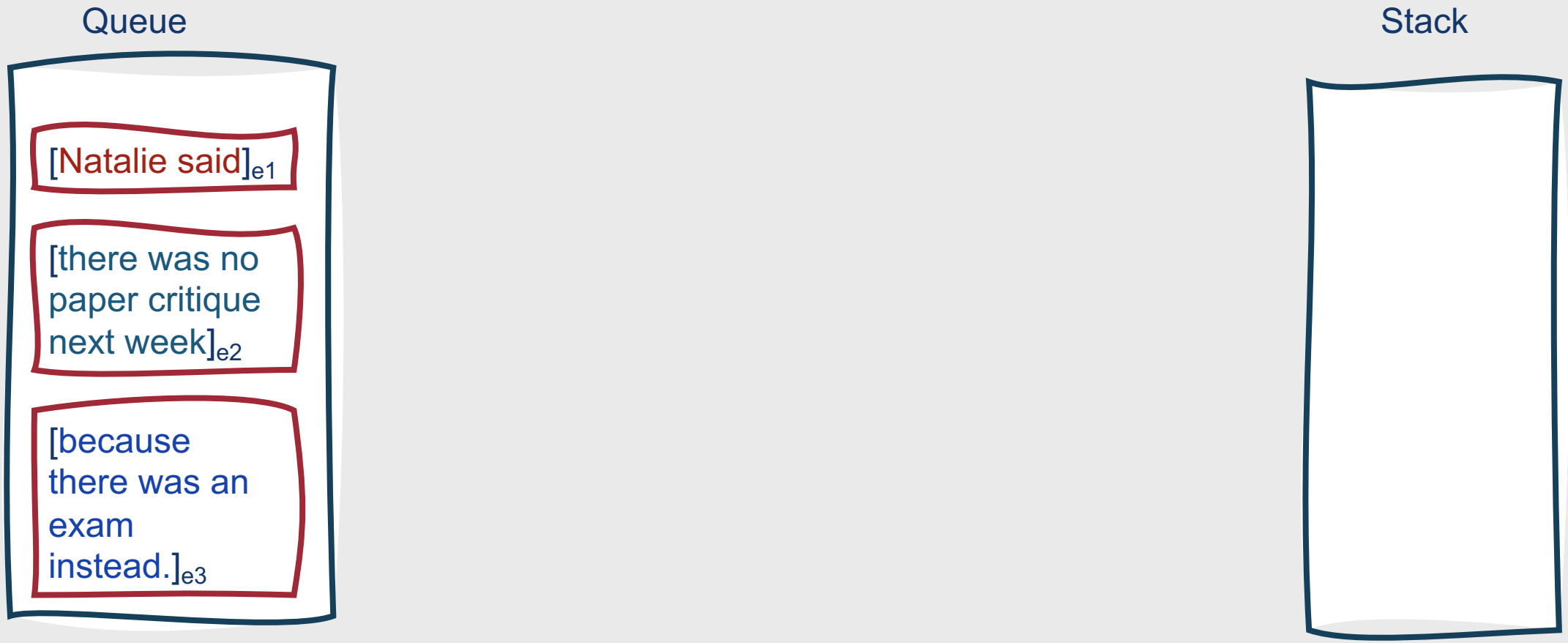


Stack



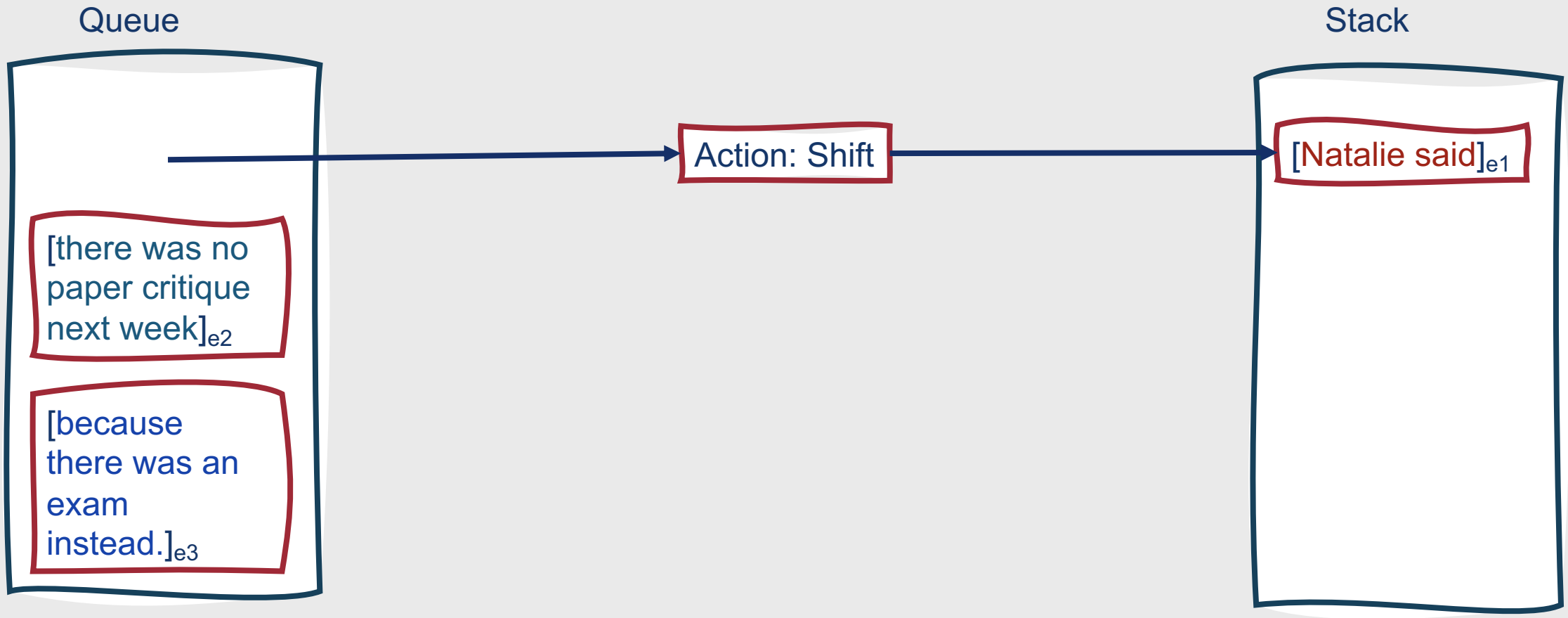
Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no paper critique next week]$_{e2}$ [because there was an exam instead.]$_{e3}$



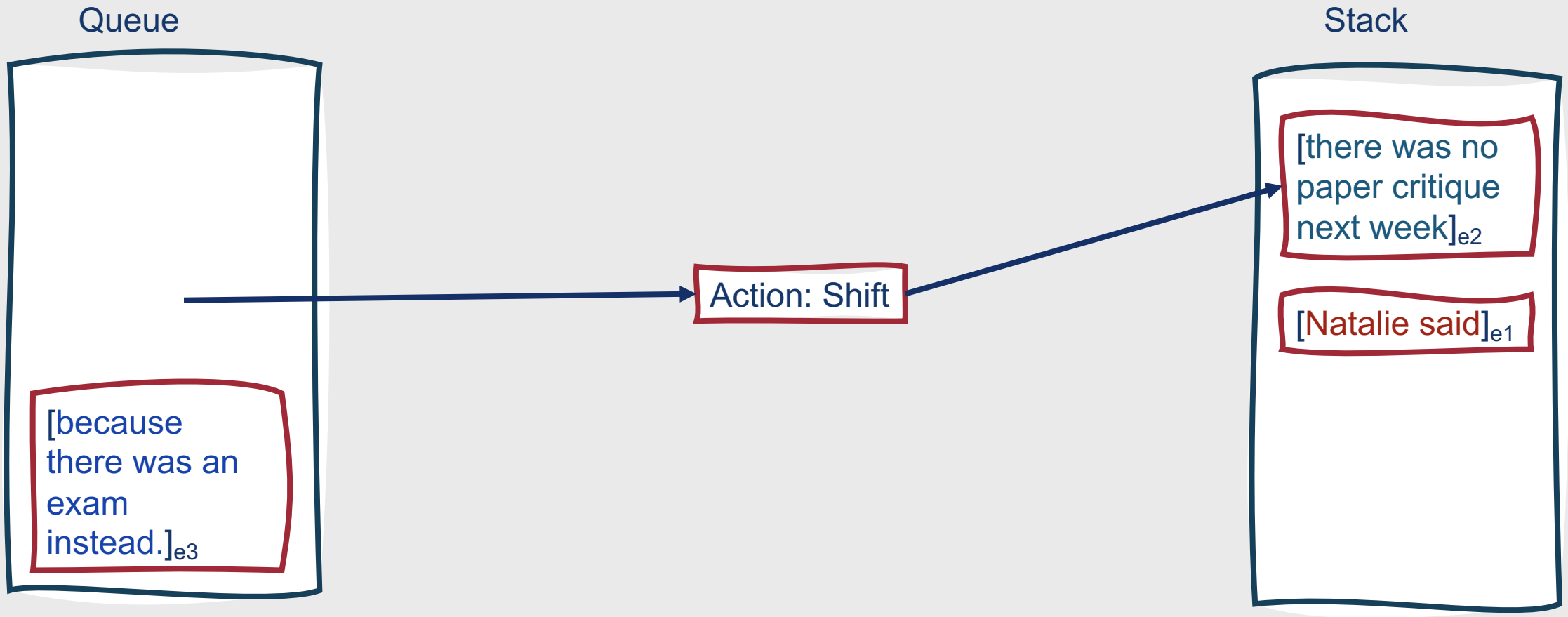
Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no paper critique next week]$_{e2}$ [because there was an exam instead.]$_{e3}$



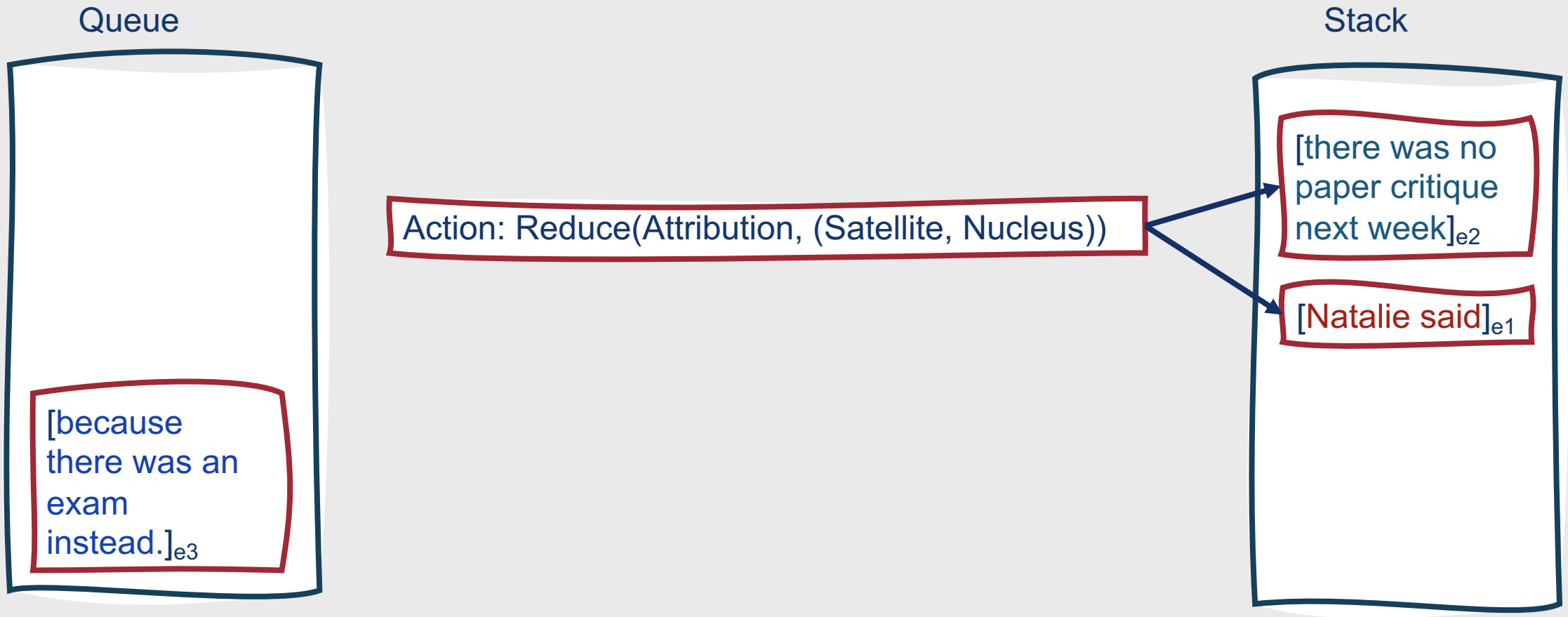
Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no paper critique next week]$_{e2}$ [because there was an exam instead.]$_{e3}$



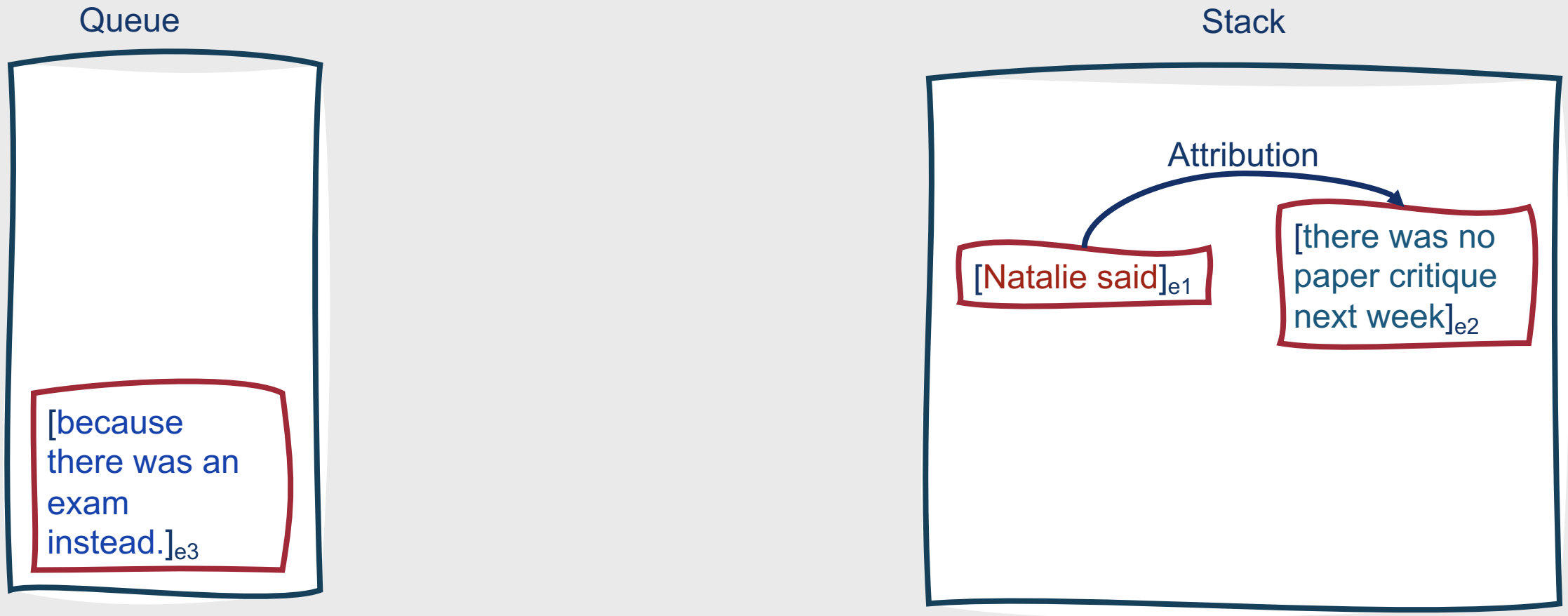
Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no paper critique next week]$_{e2}$ [because there was an exam instead.]$_{e3}$



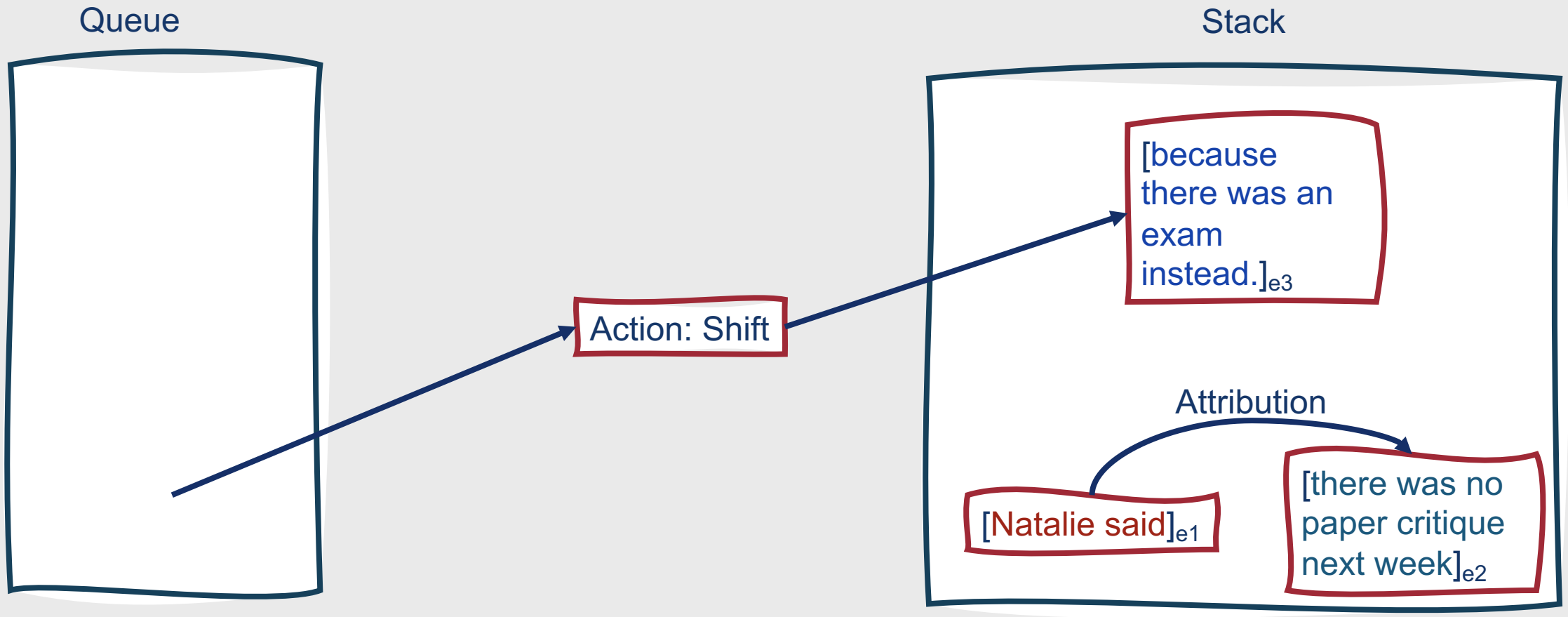
Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no paper critique next week]$_{e2}$ [because there was an exam instead.]$_{e3}$



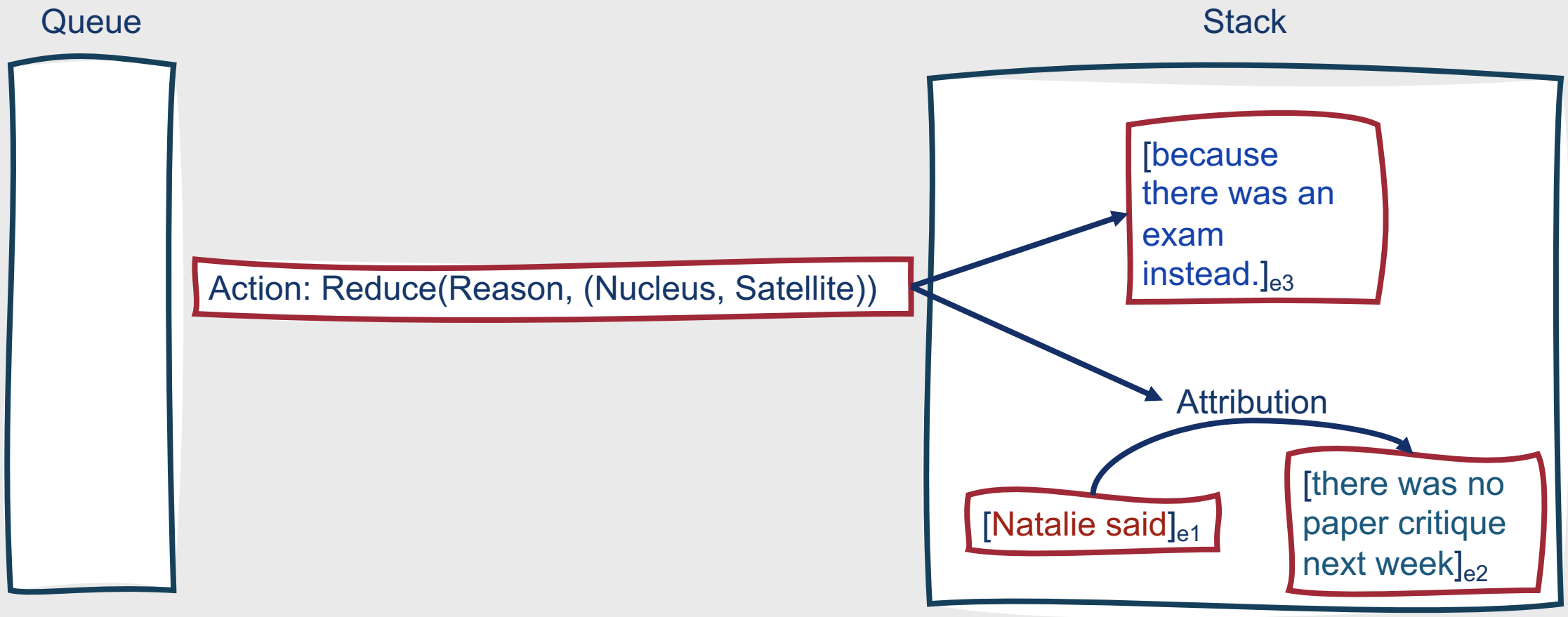
Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no paper critique next week]$_{e2}$ [because there was an exam instead.]$_{e3}$



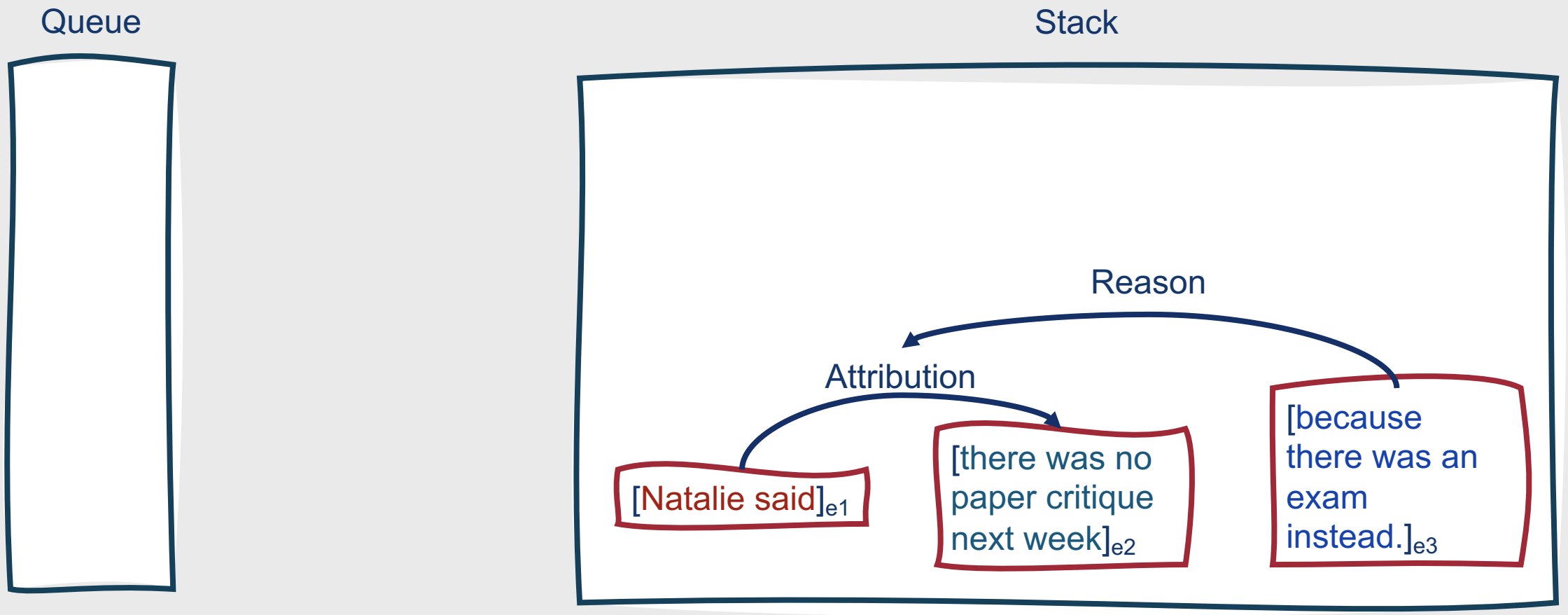
Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no paper critique next week]$_{e2}$ [because there was an exam instead.]$_{e3}$



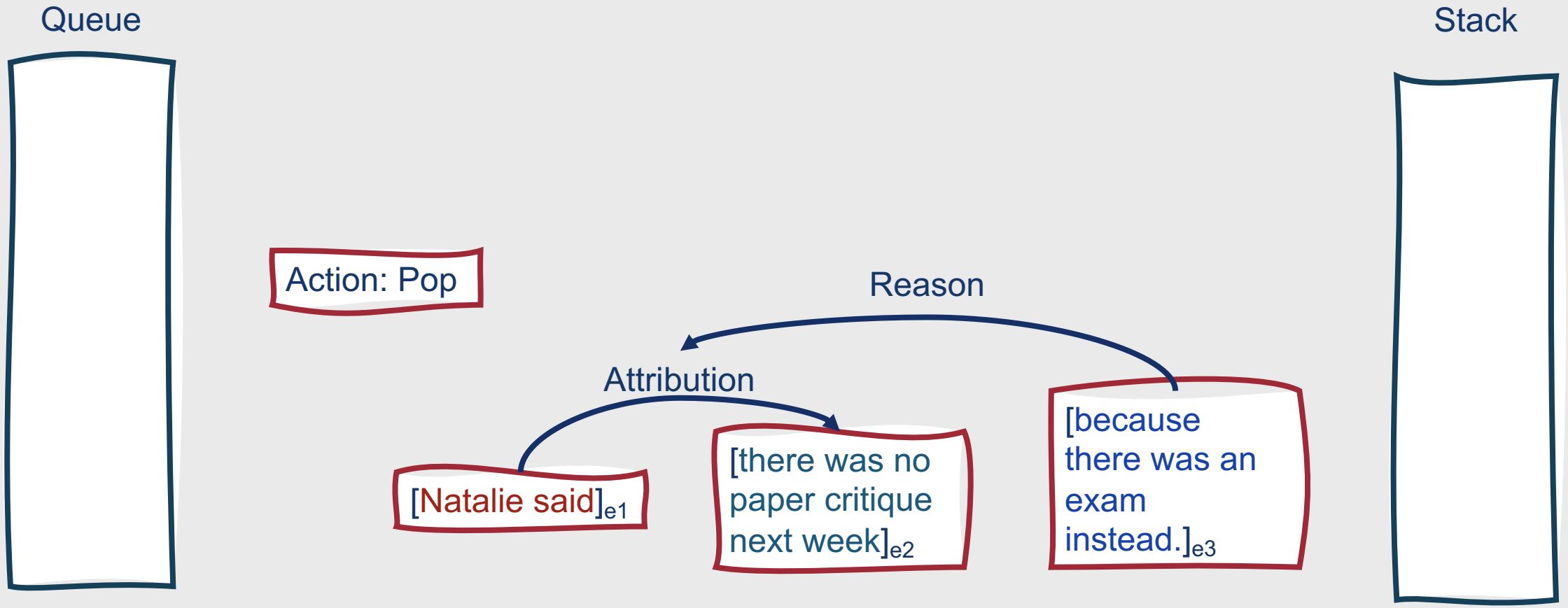
Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no paper critique next week]$_{e2}$ [because there was an exam instead.]$_{e3}$

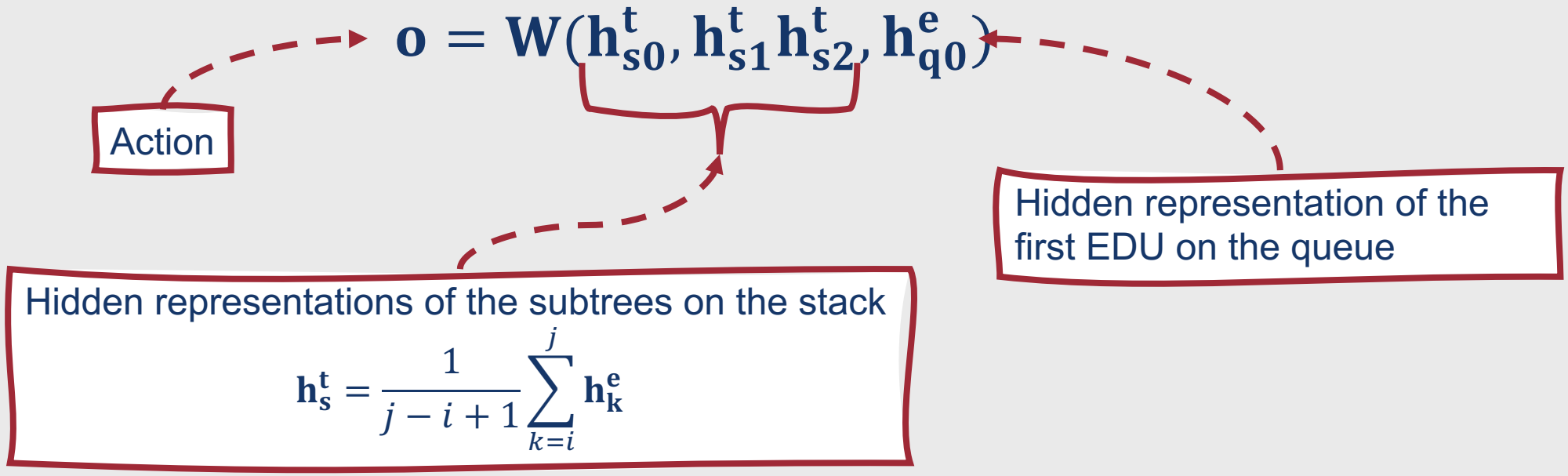


Example: Shift-Reduce Parser

[Natalie said]$_{e1}$ [there was no paper critique next week]$_{e2}$ [because there was an exam instead.]$_{e3}$



Modern RST parsers generally select actions using neural networks.



How does PDTB discourse parsing differ from this?

- **Shallow discourse parsing:** Task involves identifying relationships between text spans only, rather than full hierarchical discourse trees
- Generally four steps:
  - Find the discourse connectives
  - Find two spans for each connective
  - Label the relationship between these spans
  - Assign a relationship between adjacent pairs of sentences

Identifying discourse relations is one way to model discourse coherence….

- Another?
  - Determine **entity salience**

Entity-Based Coherence

- At each point in the discourse, some entity is salient
- A discourse is coherent by continuing to discuss the salient entity
- Two key models for entity-based coherence:
  - **Centering Theory**
  - **Entity grid model**

Centering Theory

- At any point in the discourse, one of the entities in the discourse model is salient (**being "centered" on**)
- Discourses in which adjacent sentences **continue** to maintain the same salient entity are more coherent than those which **shift** back and forth between multiple entities

Centering Theory: Intuition

- Natalie was an assistant professor at UIC.
- She taught a class there called Statistical Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- She was planning to teach the class once per year.

- Natalie was an assistant professor at UIC.
- UIC had a class that she taught called Statistical Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- The plan was that the class would be taught by Natalie once per year.

Centering Theory: Intuition

- Natalie was an assistant professor at UIC.
- She taught a class there called Statistical Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- She was planning to teach the class once per year.

- Natalie was an assistant professor at UIC.
- UIC had a class that she taught called Statistical Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- The plan was that the class would be taught by Natalie once per year.

Same propositional content, difference entity saliences

Centering Theory: Intuition

- Natalie was an assistant professor at UIC.
- She taught a class there called Statistical Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- She was planning to teach the class once per year.

Much more coherent!

- Natalie was an assistant professor at UIC.
- UIC had a class that she taught called Statistical Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- The plan was that the class would be taught by Natalie once per year.

How does Centering Theory realize this intuition?

- Maintain two representations for each utterance U_n
  - $C_b(U_n)$: Backward-looking center of U_n
    - Salient entity being focused on in the discourse after U_n is interpreted
  - $C_f(U_n)$: Forward-looking centers of U_n
    - Set of potential future salient entities (potential $C_b(U_{n+1})$)
- Set of $C_f(U_n)$ are ranked based on a variety of factors (e.g., grammatical role)
- Highest-ranked $C_f(U_n)$ is the preferred center C_p

There are four possible intersentential relationships between U_n and U_{n+1}.

- These relationships depend on $C_b(U_{n+1})$, $C_b(U_n)$, and $C_p(U_{n+1})$

	$C_b(U_{n+1}) = C_b(U_n)$ **or undefined** $C_b(U_n)$	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_{\boldsymbol{p}}(U_{n+1})$	Continue	Smooth-Shift
$C_b(U_{n+1}) \neq C_{\boldsymbol{p}}(U_{n+1})$	Retain	Rough-Shift

There are four possible intersentential relationships between U_n and U_{n+1}.

- These relationships depend on $C_b(U_{n+1})$, $C_b(U_n)$, and $C_p(U_{n+1})$

Speaker has been talking about the same entity, and is going to continue talking about the same entity

	$C_b(U_{n+1}) = C_b(U_n)$ **or undefined** $C_b(U_n)$	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_p(U_{n+1})$	Continue	Smooth-Shift
$C_b(U_{n+1}) \neq C_p(U_{n+1})$	Retain	Rough-Shift

There are four possible intersentential relationships between U_n and U_{n+1}.

- These relationships depend on $C_b(U_{n+1})$, $C_b(U_n)$, and $C_p(U_{n+1})$

Speaker plans to shift to a new entity in a future utterance, placing the current entity in a lower-rank C_f

	$C_b(U_{n+1}) = C_b(U_n)$ **or undefined** $C_b(U_n)$	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_p(U_{n+1})$	Continue	Smooth-Shift
$C_b(U_{n+1}) \neq C_p(U_{n+1})$	Retain	Rough-Shift

There are four possible intersentential relationships between U_n and U_{n+1}.

- These relationships depend on $C_b(U_{n+1})$, $C_b(U_n)$, and $C_p(U_{n+1})$

Speaker shifts to a new entity

	$C_b(U_{n+1}) = C_b(U_n)$ **or undefined** $C_b(U_n)$	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_p(U_{n+1})$	Continue	Smooth-Shift
$C_b(U_{n+1}) \neq C_p(U_{n+1})$	Retain	Rough-Shift

Based on these relationships, we can define two rules.

- If any element of $C_f(U_n)$ is realized by a pronoun in utterance U_{n+1}, then $C_b(U_{n+1})$ must be realized as a pronoun also.
- Transition states are ordered such that Continue > Retain > Smooth-Shift > Rough-Shift

	$C_b(U_{n+1}) = C_b(U_n)$ **or undefined** $C_b(U_n)$	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_{\boldsymbol{p}}(U_{n+1})$	Continue	Smooth-Shift
$C_b(U_{n+1}) \neq C_{\boldsymbol{p}}(U_{n+1})$	Retain	Rough-Shift

Why these rules?

- If any element of $C_f(U_n)$ is realized by a pronoun in utterance U_{n+1}, then $C_b(U_{n+1})$ must be realized as a pronoun also.
- Transition states are ordered such that Continue > Retain > Smooth-Shift > Rough-Shift

- If there are multiple pronouns in an utterance realizing entities from the previous utterance, one of them must realize the backward-looking center C_b
- If there's only one pronoun, it must be C_b

	$C_b(U_{n+1}) = C_b(U_n)$ **or undefined** $C_b(U_n)$	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_{\boldsymbol{p}}(U_{n+1})$	Continue	Smooth-Shift
$C_b(U_{n+1}) \neq C_{\boldsymbol{p}}(U_{n+1})$	Retain	Rough-Shift

Why these rules?

- If any element of $C_f(U_n)$ is realized by a pronoun in utterance U_{n+1}, then $C_b(U_{n+1})$ must be realized as a pronoun also.
- Transition states are ordered such that Continue > Retain > Smooth-Shift > Rough-Shift

Discourses that continue to center the same entity are more coherent than those that repeatedly shift to other centers

	$C_b(U_{n+1}) = C_b(U_n)$ **or undefined** $C_b(U_n)$	$C_b(U_{n+1}) \neq C_b(U_n)$
$C_b(U_{n+1}) = C_{\boldsymbol{p}}(U_{n+1})$	Continue	Smooth-Shift
$C_b(U_{n+1}) \neq C_{\boldsymbol{p}}(U_{n+1})$	Retain	Rough-Shift

With this in mind, we can revisit the sample texts from earlier….

- Natalie was an assistant professor at UIC.
- She taught a class there called Statistical Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- She was planning to teach the class once per year.

- Natalie was an assistant professor at UIC.
- UIC had a class that she taught called Statistical Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- The plan was that the class would be taught by Natalie once per year.

With this in mind, we can revisit the sample texts from earlier….

- Natalie was an assistant professor at UIC.
- She taught a class there called Statistical Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- She was planning to teach the class once per year.

$C_f(U_1)$: {Natalie, UIC}
$C_p(U_1)$: Natalie
$C_b(U_1)$: undefined

$C_f(U_2)$: {Natalie, class}
$C_p(U_2)$: Natalie
$C_b(U_2)$: Natalie

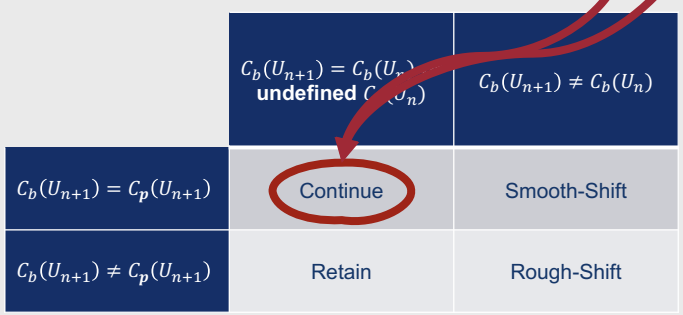
- Natalie was an assistant professor at UIC.
- UIC had a class that she taught called Statistical Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- The plan was that the class would be taught by Natalie once per year.

With this in mind, we can revisit the sample texts from earlier….

- Natalie was an assistant professor at UIC.
- She taught a class there called Statistical Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- She was planning to teach the class once per year.

$C_f(U_1)$: {Natalie, UIC}
$C_p(U_1)$: Natalie
$C_b(U_1)$: undefined

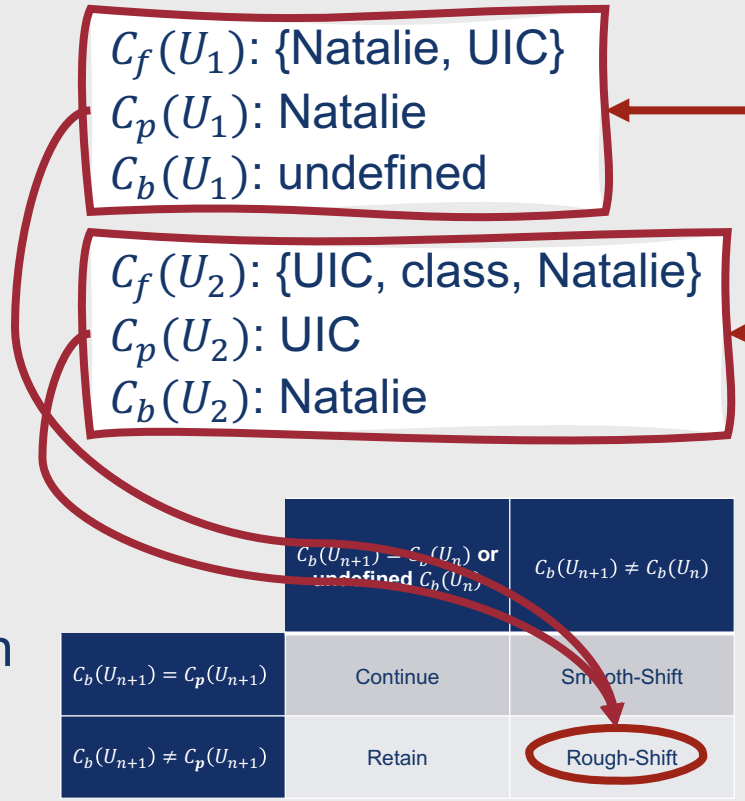
$C_f(U_2)$: {Natalie, class}
$C_p(U_2)$: Natalie
$C_b(U_2)$: Natalie



- Natalie was an assistant professor at UIC.
- UIC had a class that she taught called Statistical Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- The plan was that the class would be taught by Natalie once per year.

With this in mind, we can revisit the sample texts from earlier….

- Natalie was an assistant professor at UIC.
- She taught a class there called Statistical Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- She was planning to teach the class once per year.



- Natalie was an assistant professor at UIC.
- UIC had a class that she taught called Statistical Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- The plan was that the class would be taught by Natalie once per year.

With this in mind, we can revisit the sample texts from earlier….

- Natalie was an assistant professor at UIC.
- She taught a class there called Statistical Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- She was planning to teach the class once per year.

- Natalie was an assistant professor at UIC.
- UIC had a class that she taught called Statistical Natural Language Processing.
- She enjoyed teaching the class, because she liked NLP a lot.
- The plan was that the class would be taught by Natalie once per year.

Entity Grid Model

- Alternative way to capture entity-based coherence
- Uses machine learning to **induce patterns of entity mentioning** that make a discourse more coherent
- Based on an **entity grid**
  - 2d array representing the **distribution of entity mentions across sentences**
    - Rows = sentences
    - Columns = discourse entities
    - Values in cells = Whether the entity appears in the sentence, and its grammatical role (subject, object, neither, or absent)

	Natalie	assistant professor	UIC	class	teaching	NLP	planning	year
S1								
S2								
S3								
S4								

Example: Entity Grid Model

- $[\text{Natalie}]_s$ was an $[\text{assistant professor}]_o$ at $[\text{UIC}]_x$.
- $[\text{Natalie}]_s$ taught a $[\text{class}]_o$ at $[\text{UIC}]_x$ called Statistical Natural Language Processing.
- $[\text{Natalie}]_s$ enjoyed $[\text{teaching}]_o$ the $[\text{class}]_x$, because $[\text{Natalie}]_s$ liked $[\text{NLP}]_o$ a lot.
- $[\text{Natalie}]_s$ was $[\text{planning}]_o$ to teach the $[\text{class}]_x$ once per $[\text{year}]_x$.

	Natalie	assistant professor	UIC	class	teaching	NLP	planning	year
S1	S	O	X	-	-	-	-	-
S2								
S3								
S4								

Example: Entity Grid Model

- **$[\text{Natalie}]_s$ was an $[\text{assistant professor}]_o$ at $[\text{UIC}]_x$.**
- $[\text{Natalie}]_s$ taught a $[\text{class}]_o$ at $[\text{UIC}]_x$ called Statistical Natural Language Processing.
- $[\text{Natalie}]_s$ enjoyed $[\text{teaching}]_o$ the $[\text{class}]_x$, because $[\text{Natalie}]_s$ liked $[\text{NLP}]_o$ a lot.
- $[\text{Natalie}]_s$ was $[\text{planning}]_o$ to teach the $[\text{class}]_x$ once per $[\text{year}]_x$.

	Natalie	assistant professor	UIC	class	teaching	NLP	planning	year
S1	S	O	X	-	-	-	-	-
S2	S	-	X	O	-	-	-	-
S3								
S4								

Example: Entity Grid Model

- $[\text{Natalie}]_s$ was an $[\text{assistant professor}]_o$ at $[\text{UIC}]_x$.
- **$[\text{Natalie}]_s$ taught a $[\text{class}]_o$ at $[\text{UIC}]_x$ called Statistical Natural Language Processing.**
- $[\text{Natalie}]_s$ enjoyed $[\text{teaching}]_o$ the $[\text{class}]_x$, because $[\text{Natalie}]_s$ liked $[\text{NLP}]_o$ a lot.
- $[\text{Natalie}]_s$ was $[\text{planning}]_o$ to teach the $[\text{class}]_x$ once per $[\text{year}]_x$.

	Natalie	assistant professor	UIC	class	teaching	NLP	planning	year
S1	S	O	X	-	-	-	-	-
S2	S	-	X	O	-	-	-	-
S3	S	-	-	X	O	O	-	-
S4								

Example: Entity Grid Model

- $[\text{Natalie}]_s$ was an $[\text{assistant professor}]_o$ at $[\text{UIC}]_x$.
- $[\text{Natalie}]_s$ taught a $[\text{class}]_o$ at $[\text{UIC}]_x$ called Statistical Natural Language Processing.
- **$[\text{Natalie}]_s$ enjoyed $[\text{teaching}]_o$ the $[\text{class}]_x$, because $[\text{Natalie}]_s$ liked $[\text{NLP}]_o$ a lot.**
- $[\text{Natalie}]_s$ was $[\text{planning}]_o$ to teach the $[\text{class}]_x$ once per $[\text{year}]_x$.

	Natalie	assistant professor	UIC	class	teaching	NLP	planning	year
S1	S	O	X	-	-	-	-	-
S2	S	-	X	O	-	-	-	-
S3	S	-	-	X	O	O	-	-
S4	S	-	-	X	-	-	O	X

Example: Entity Grid Model

- $[\text{Natalie}]_s$ was an $[\text{assistant professor}]_o$ at $[\text{UIC}]_x$.
- $[\text{Natalie}]_s$ taught a $[\text{class}]_o$ at $[\text{UIC}]_x$ called Statistical Natural Language Processing.
- $[\text{Natalie}]_s$ enjoyed $[\text{teaching}]_o$ the $[\text{class}]_x$, because $[\text{Natalie}]_s$ liked $[\text{NLP}]_o$ a lot.
- **$[\text{Natalie}]_s$ was $[\text{planning}]_o$ to teach the $[\text{class}]_x$ once per $[\text{year}]_x$.**

Entity Grid Model

- Dense columns indicate entities mentioned often
- Sparse columns indicate entities mentioned rarely
- Coherence is thus measured by patterns of **local entity transition**
- Each transition ends up with a probability

Example: Entity Grid Model

	Natalie	assistant professor	UIC	class	teaching	NLP	planning	year
S1	S	O	X	-	-	-	-	-
S2	S	-	X	O	-	-	-	-
S3	S	-	-	X	O	O	-	-
S4	S	-	-	X	-	-	O	X

{x, x, -, -}

	Natalie	assistant professor	UIC	class	teaching	NLP	planning	year
S1	S	O	X	-	-	-	-	-
S2	S	-	X	O	-	-	-	-
S3	S	-	-	X	O	O	-	-
S4	S	-	-	X	-	-	O	X

Example: Entity Grid Model

{x, x, -, -}

$$p(\{x, x, -, -\}) = \frac{1}{8}$$

Example: Entity Grid Model

	Natalie	assistant professor	UIC	class	teaching	NLP	planning	year
S1	S	O	X	-	-	-	-	-
S2	S	-	X	O	-	-	-	-
S3	S	-	-	X	O	O	-	-
S4	S	-	-	X	-	-	O	X

{O, -}

$$p(\{o, -\}) = \frac{3}{24}$$

Entity Grid Model

- These transitions and their probabilities can be used as features for a machine learning model, trained to predict coherence scores
- These models can be trained in a **self-supervised** manner:
  - Learn to distinguish the natural order of sentences in a discourse (expected to be coherent) from a modified order (e.g., randomized order)

How do we evaluate entity-based coherence models?

- Best option: Compare human coherence ratings with predicted coherence ratings from the model
- However, collecting human labels is expensive!
- Alternate option:
  - Take a naturally-occurring document, and use this as a positive sample
  - Mess up the order of its sentences in some way, and use this as a negative sample
    - Random permutation
    - Or, move one of the sentences to a different position
  - Evaluate the frequency with which the model predicts the naturally-occurring document to be more coherent than the messed-up version(s)

We've talked about identifying coherence relations and entity salience …what about topical salience?

- Discourses are more coherent when they discuss a consistent set of topics
- This can be modeled using measures of **lexical cohesion**
  - **Lexical cohesion:** The sharing of identical or semantically-related words across nearby sentences

Latent Semantic Analysis (LSA)

- Early model of lexical cohesion
- First to use word embeddings!
- Models the coherence between two sentences *i* and *j* as the cosine between their embedding vectors (traditionally, dimensionality-reduced TF*IDF vectors)
  - $\text{sim}(i,j) = \cos(i,j) = \cos(\sum_{w \in i} \mathbf{w}, \sum_{w \in j} \mathbf{w})$
- The overall coherence of a text is thus the average similarity over all pairs of adjacent sentences s_i and s_{i+1}
  - $coherence(t) = \frac{1}{n-1} \sum_{i=1}^{n-1} \text{sim}(s_i, s_{i+1})$

More modern models make use of this intuition as well.

- **Local coherence discriminator (LCD)**
  - Computes the coherence of a text as the average of coherence scores between adjacent sentences
  - Learns to discriminate between naturally-occurring adjacent sentences and those in a messed-up order in a neural, self-supervised manner
- Neural models of topical coherence may implicitly learn that sentences having causal/temporal relations, and sentences exhibiting higher entity coherence, tend to be overall more coherent as well

Coherence relations, entity salience, and topical salience all focus on local coherence.

- However, discourses must be globally coherent as well!
  - Stories have an overall narrative structure
  - Persuasive essays follow specific argument structure
  - Scientific papers are characterized by a structure common across research publications

Argumentation Structure

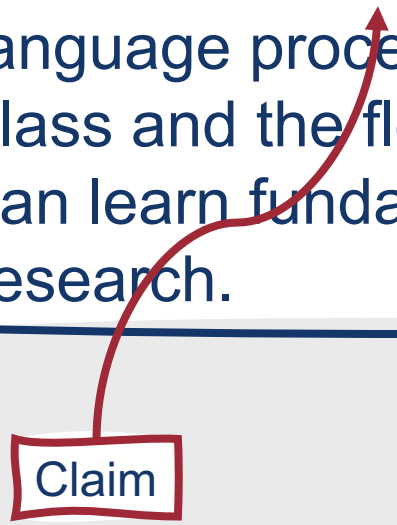
- **Argumentation mining:** The computational analysis of rhetorical strategy
- Persuasive arguments generally contain well-defined argumentative components:
  - **Claim:** The central, controversial, component of the argument
  - **Premise:** A persuasive support or attack of the claim or another premise

Example: Argumentation Structure

CS 521 is the best class at UIC. It covers a very exciting topic: natural language processing. It also offers both the structure of a lecture-based class and the flexibility of a seminar course. This mix is nice because you can learn fundamental principles but also get up to speed on contemporary research.

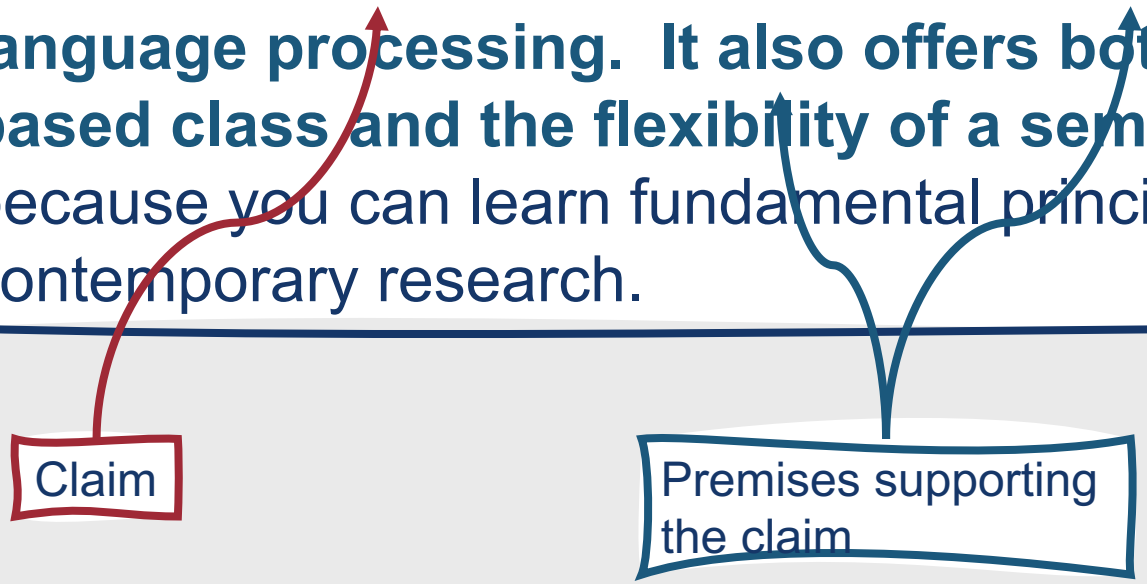
Example: Argumentation Structure

CS 521 is the best class at UIC. It covers a very exciting topic: natural language processing. It also offers both the structure of a lecture-based class and the flexibility of a seminar course. This mix is nice because you can learn fundamental principles but also get up to speed on contemporary research.

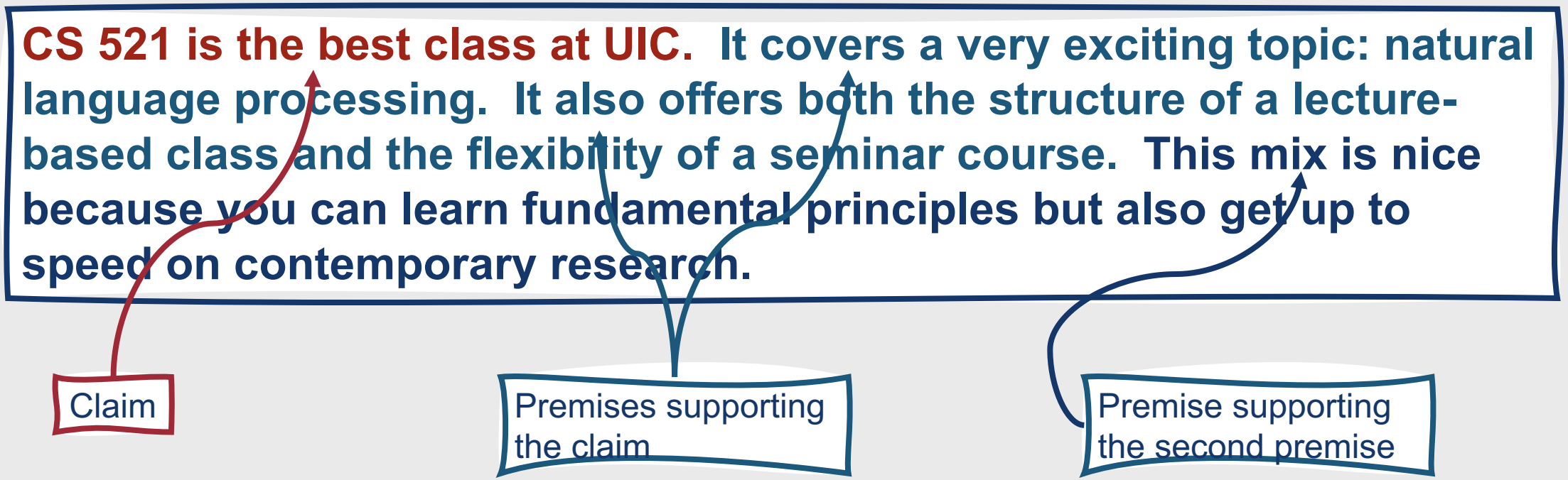


Example: Argumentation Structure

CS 521 is the best class at UIC. It covers a very exciting topic: natural language processing. It also offers both the structure of a lecture-based class and the flexibility of a seminar course. This mix is nice because you can learn fundamental principles but also get up to speed on contemporary research.



Example: Argumentation Structure

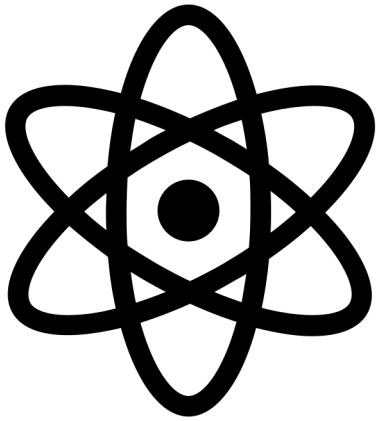


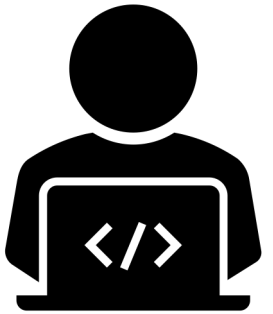
How can we detect argumentation structure?

- Classifiers to identify claims, premises, and non-argumentation
- Methods to detect specific argumentation schemes
  - Argument from example
  - Argument from cause to effect
  - Argument from consequences
  - Etc.
- Related research: Studying how components of argument structure are associated with persuasive success

We can apply similar methods to scientific discourse.

- In scientific papers, authors need to:
  - Indicate a scientific goal
  - Develop a method for reaching that goal
  - Provide evidence for the solution
  - Compare to prior work
- Parallel to argumentation structure: Each paper tries to make a **knowledge claim**!
- Modeling scientific discourse is an active research problem, as is modeling other global discourse structures (e.g., stories)





Summary: Discourse Coherence

- **Discourse coherence** is the relationship (or lack thereof) between sentences in a discourse
- It is influenced by a variety of factors:
  - **Coherence relations**
  - **Entity salience**
  - **Topical salience**
  - **Global structure**
- Common models of discourse relation include **Rhetorical Structure Theory** and the **Penn Discourse Treebank**
- **Entity salience** can be modeled using **Centering Theory** or the **entity grid** model
- **Lexical cohesion** may be measured using **latent semantic analysis** or other word embedding-based methods
- **Argumentation structure** captures **global coherence**, and may be applied to a variety of domains including persuasive essays and scientific discourse