

# Toward Interactive Dictation

Belinda Z. Li, Jason Eisner, Adam Pauls, Sam Thomson  
ACL 2023



Microsoft Semantic Machines

# Problem Overview



*Just wanted to ask about the event on the 23rd.*

Just wanted to ask about the event on the 23rd.



*—on **Friday** the 23rd.*

Just wanted to ask about the event on Friday the 23rd.



*Is the event still on?*

Just wanted to ask about the event on Friday the 23rd. Is the event still on?



*Replace “the event” in the last sentence with “it”.*

Just wanted to ask about the event on Friday the 23rd. Is it still on?

# Existing Speech-to-Text Systems

Most current systems do not support editing through voice.

Ones that do:



**Nuance Dragon NaturallySpeaking**



**Microsoft Word, Dictate**

# Limitation: Inflexible natural language for commanding

- Relies on wake words to activate command mode
- Users must memorize a list of commands



## Control the microphone

Go to sleep | Stop listening  
Wake up | Listen to me  
Microphone off

## Get help

Give me help  
"Search the help for ..."  
What can I say  
– Show navigation commands  
– Show correction commands  
– Show formatting commands  
– Show punctuation commands  
– Open help

## Basic dictation and editing

### Add lines and spaces

New line | Press enter  
New paragraph  
Press Tab key | Tab key | Tab  
Tab <n> times

## Format

Underline <xyz>, Capitalize <xyz>  
All caps on | off  
Quote that  
Bracket that

## Copy and paste

Cut | Copy that  
Cut | Copy <text>  
Cut | Copy from <text> to <text>  
Paste that

## Spell out

Spell that  
Spell <cap b a hyphen 5>  
Spell <Charlie alpha papa>  
Switch to Spell mode

## Move the insertion point

Insert before <xyz>  
Go back  
Go to top | bottom

## Dictating punctuation

Period  
Comma  
Question mark  
Exclamation mark

## Fixing mistakes

Undo | Undo that  
Scratch that  
Scratch that <n> times  
Delete line  
Delete last <n> words  
Delete <text>  
Resume with <xyz>  
Backspace <n>  
Correct <xyz>  
Correct that

## Select text

Select all  
Select <xyz>  
Select next <n> words  
Select <start> through <end>  
Select previous paragraph  
Select document  
Unselect that

Move down <n> lines  
Go to end of line  
Move left <n> characters  
Page up | down

## Move in a list

Move down <n>  
Go to bottom | top  
Press Enter  
Press right arrow

## Add new words or commands

Add new word  
Add new command  
Open vocabulary editor  
Open command browser

## Search the computer

Search the computer for...  
Search documents for...  
Search e-mail for...

## Searching the web

Search the web for <text>  
Search eBay for <text>  
Open top sites for <text>  
Search video for <text>

# Limitation: Inflexible natural language for commanding

- Relies on wake words to activate command mode
- Users must memorize a list of commands



## Control the microphone

Go to sleep | Stop listening  
Wake up | Listen to me  
Microphone off

## Get help

Give me help  
"Search the help for ..."  
What can I say  
– Show navigation commands  
– Show correction commands  
– Show formatting commands  
– Show punctuation commands  
– Open help

## Basic dictation and editing

### Add lines and spaces

New line | Press enter  
New paragraph  
Press Tab key | Tab key | Tab  
Tab <n> times

## Format

Underline <xyz>, Capitalize <xyz>  
All caps on | off  
Quote that  
Bracket that

## Copy and paste

Cut | Copy that  
Cut | Copy <text>  
Cut | Copy from <text> to <text>  
Paste that

## Spell out

Spell that  
Spell <cap b a hyphen 5>  
Spell <Charlie alpha papa>  
Switch to Spell mode

## Move the insertion point

Insert before <xyz>  
Go back  
Go to top | bottom

## Dictating punctuation

Period  
Comma  
Question mark  
Exclamation mark

## Fixing mistakes

Undo | Undo that  
Scratch that  
Scratch that <n> times  
Delete line  
Delete last <n> words  
Delete <text>  
Resume with <xyz>  
Backspace <n>  
Correct <xyz>  
Correct that

## Select text

Select all  
Select <xyz>  
Select next <n> words  
Select <start> through <end>  
Select previous paragraph  
Select document  
Unselect that

Move down <n> lines  
Go to end of line  
Move left <n> characters  
Page up | down

## Move in a list

Move down <n>  
Go to bottom | top  
Press Enter  
Press right arrow

## Add new words or commands

Add new word  
Add new command  
Open vocabulary editor  
Open command browser

## Search the computer

Search the computer for...  
Search documents for...  
Search e-mail for...

## Searching the web

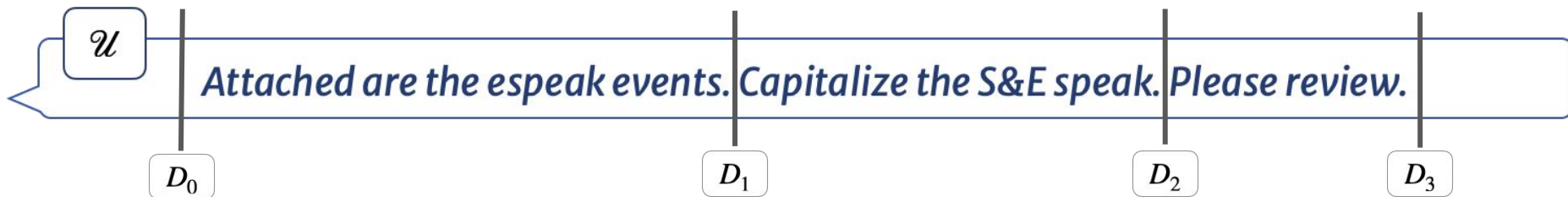
Search the web for <text>  
Search eBay for <text>  
Open top sites for <text>  
Search video for <text>

**Want *natural and intuitive*  
dictation and commanding**

# New Task!: Interactive Dictation

## 1. Flexible interleaving of dictation and editing

- No reserved trigger words for invoking commands
- Challenge: Predicting *segmentation* between dictation and editing commands



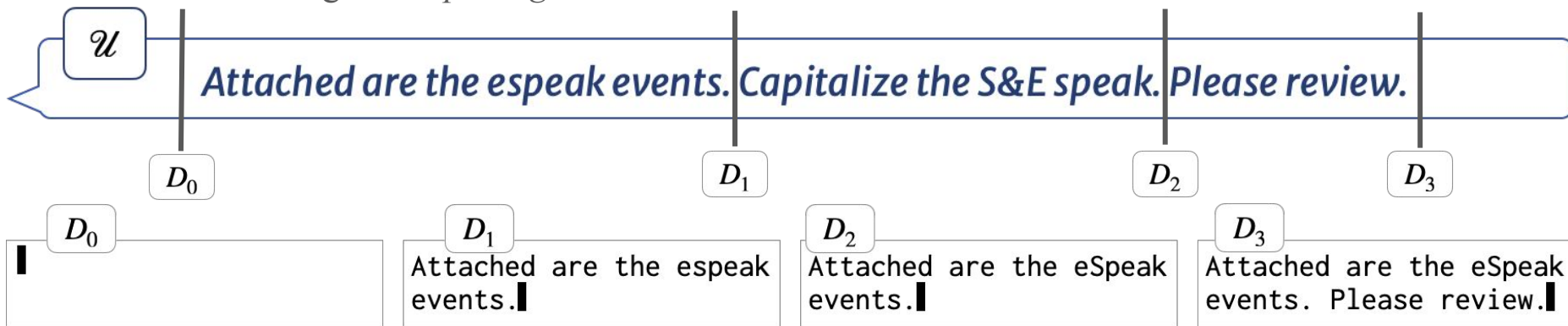
# New Task!: Interactive Dictation

## 1. Flexible interleaving of dictation and editing

- No reserved trigger words for invoking commands
- Challenge: Predicting *segmentation* between dictation and editing commands

## 2. Intuitive and open-ended natural language for editing

- No fixed templates for different types of command
- Challenge: *Interpreting* which command to invoke and where/how



# Our Contributions

1. Introducing and formalizing a new task, **Interactive Dictation**
2. Designing a data collection interface and build a dataset for this task
3. Creating a baseline system for the task



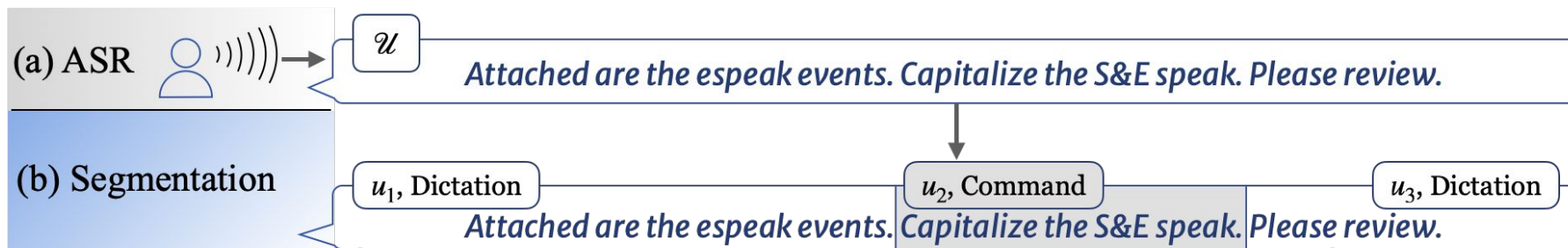
# Interactive Dictation: Basic Procedure



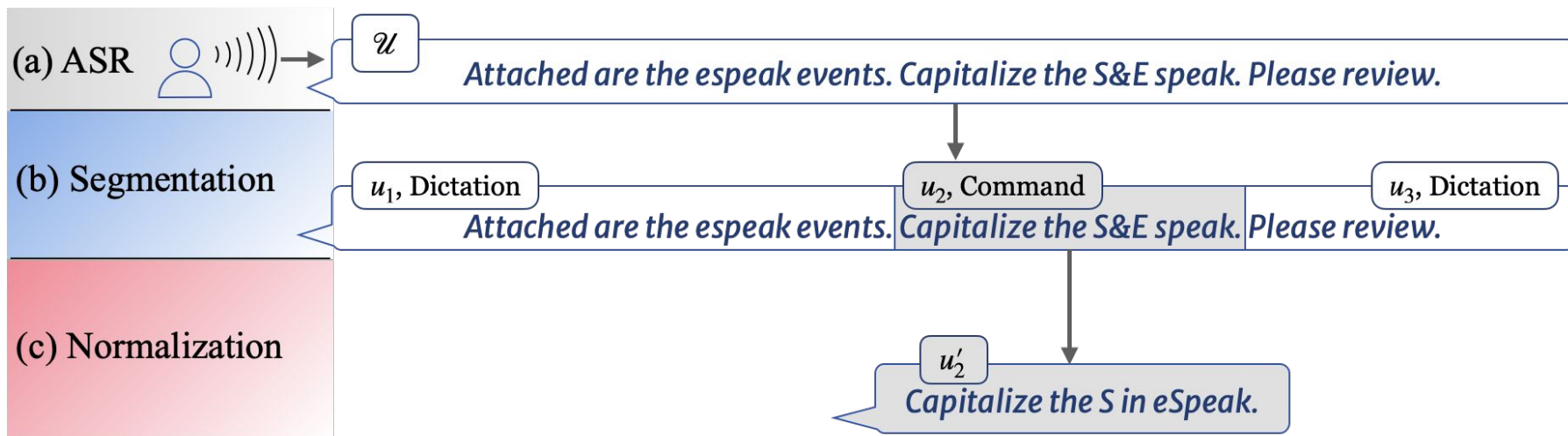
# Interactive Dictation: Building a System



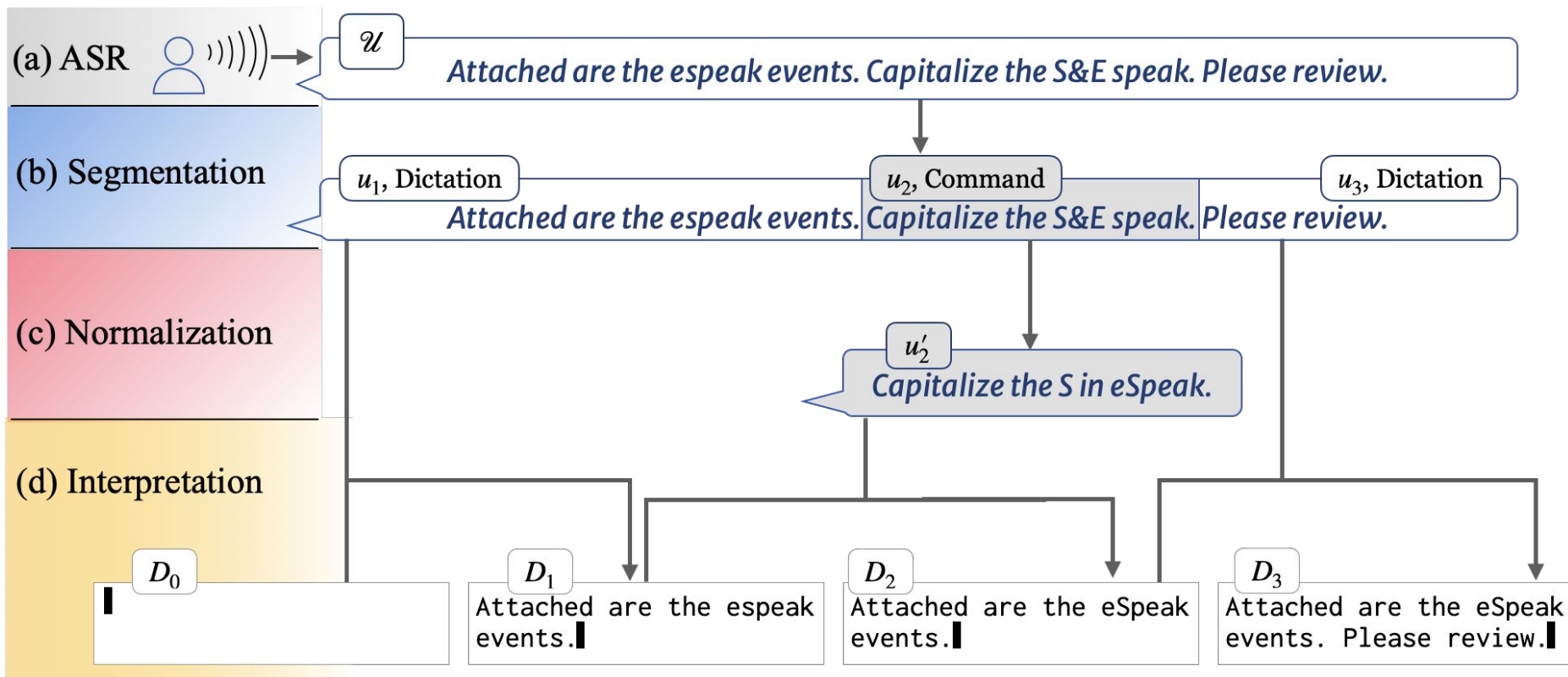
# Interactive Dictation: Building a System



# Interactive Dictation: Building a System



# Interactive Dictation: Building a System



# Annotating Commands & Transcriptions

## Annotation Interface

The screenshot displays the Annotation Interface, which is divided into two main sections: Dictation Segments and Command Segments.

**Dictation Segments:**

- Begin transcription.** Hold 'ctrl' to go into command mode.
- Pause recording.**
- Replicate the following text:** ☐ See Remaining Changes
- Attached are the eSpeak events currently scheduled for 2000. Please review at your convenience.** (Target document state  $D_n$ )
- Have a great day!** (The eThink Team)
- Transcription output appears here... (click to edit)**
- Update Cursor Position**
- Attached are the espeak events currently scheduled for 2000.** (Document state after selected segment  $D_i$ )
- Output diff appears here...** (Change in document state from segment  $\Delta(D_{i-1}, D_i)$ )
- Attached are the espk events currently scheduled for 2000.** (Document state before selected segment  $D_{i-1}$ )
- Preceding output appears here... Copy**
- Attached are the espeak events currently scheduled for 2000.**

**Command Segments:**

- Command log:**
- insertText: |-** (ASR: Attached are the espeak events currently scheduled for 2000.)
- editText: |-** (ASR: Respo espi as he speak. (Literal Utterance  $u_i$ ))
- Gold ASR: (click to edit)** (Normalized Utterance  $u'_i$ )
- Respell espi as espeak.** (Selected segment to edit  $O_i$ )
- insertText: |-** (ASR: Please review at your convenience. (Actual Literal Utterance))
- editText: |-** (ASR: Capitalize the S&E speak.)
- Gold ASR: (click to edit)** (Capitalized the S in eSpeak.)

**Buttons:**

- Delete Selected Command & Afterwards**
- Reset**
- Submit**

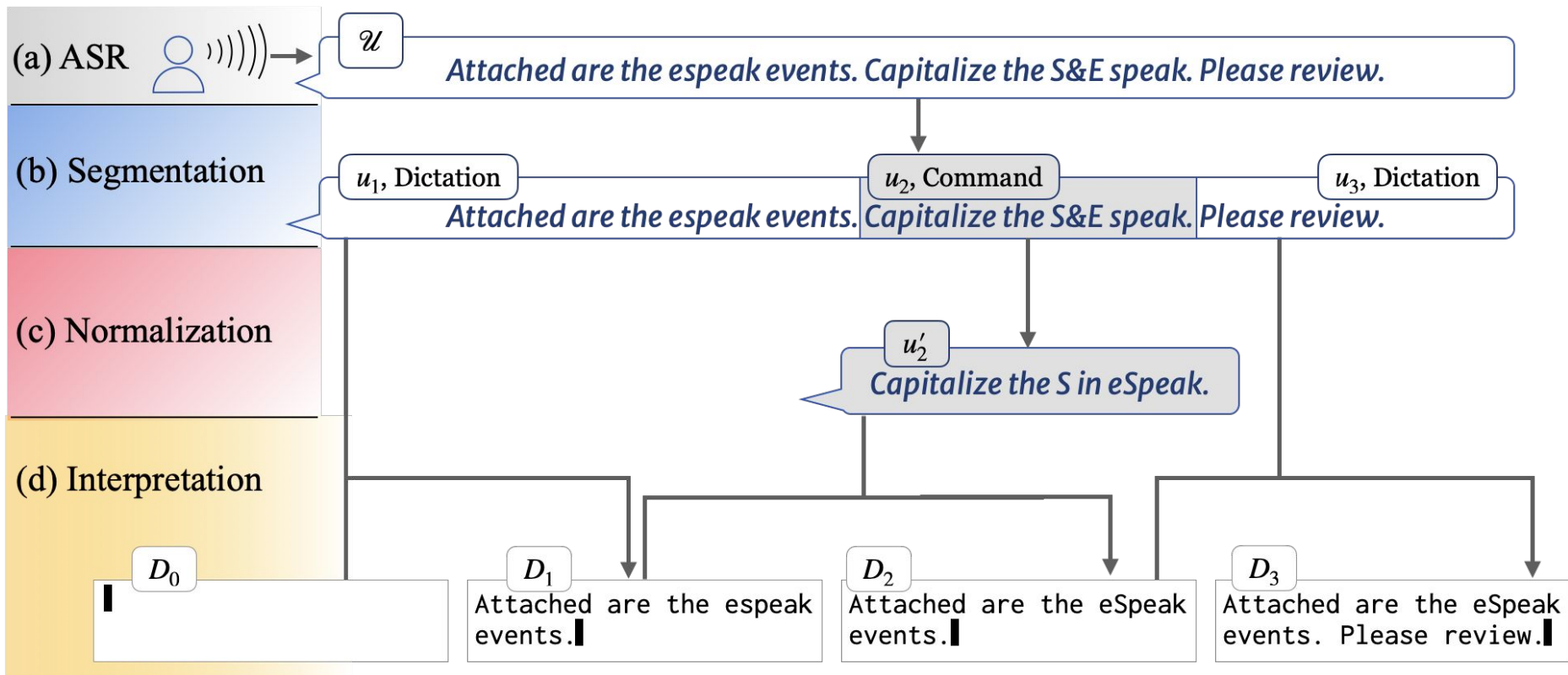
# Dataset: TERTIUS

11 annotators were instructed to do one of the following:

1. *Replicate doc*: exactly recreate an email from the Enron Email Dataset
2. *Elaborate doc*: expand a terse description of an email into a full email
3. *Replicate segment*: exactly recreate the effect of a single command segment sampled from annotations on the previous 2 objectives

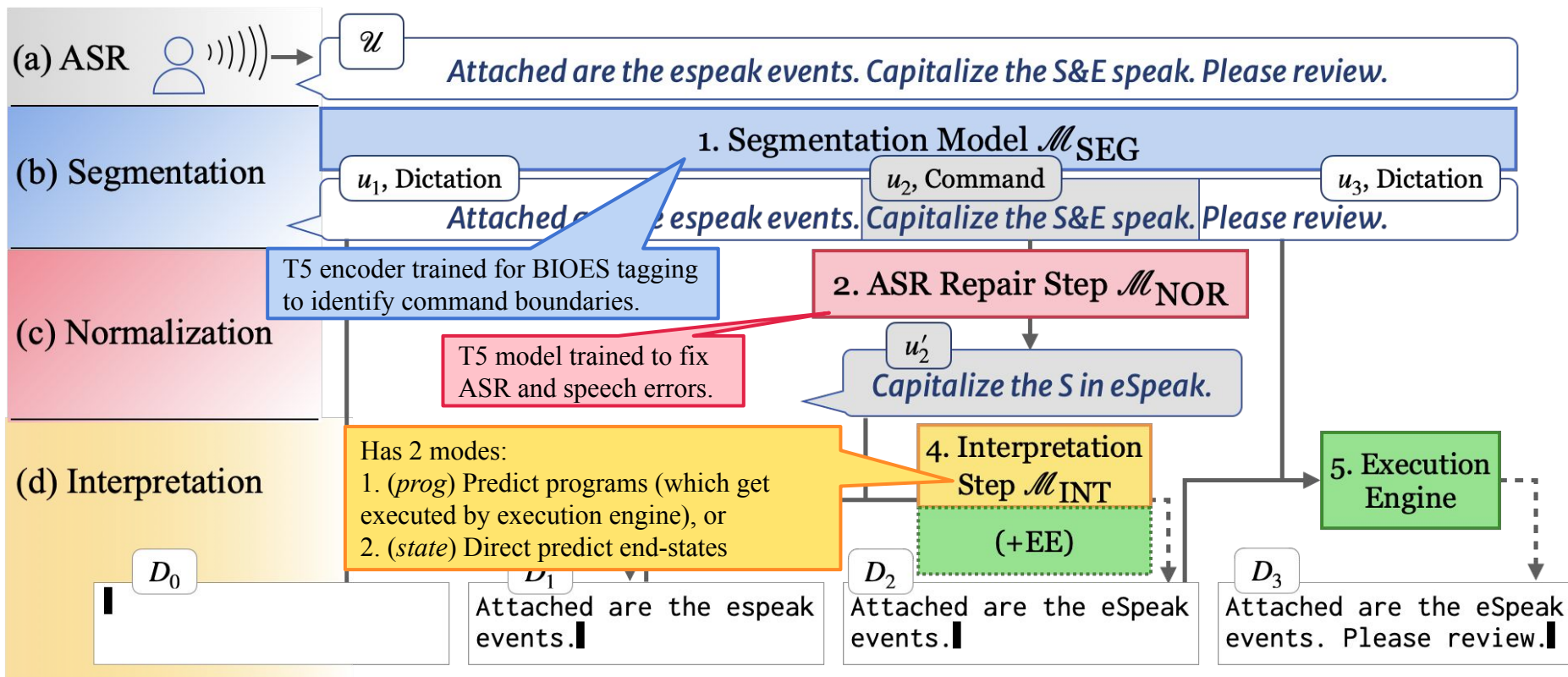
Trajectories	Segments		
	Dictation	Command	Total
1320	959	3225	4184

# Interactive Dictation: Building a System





# Interactive Dictation: Instantiating Models



## Results: Segmentation model

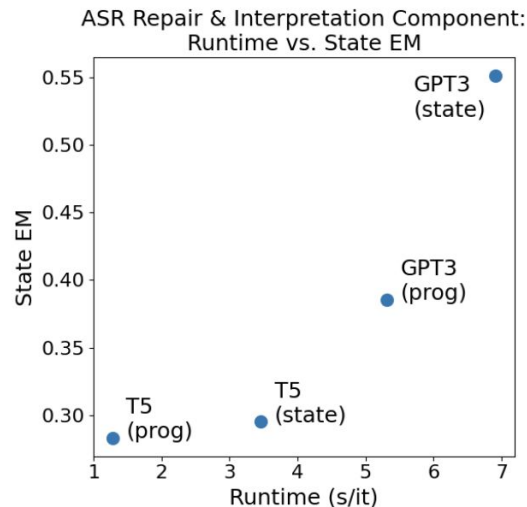
**Exact Match:** # of dialogues in which command boundary are exactly correct

Model	Segmentation Exact-Match (Per dialogue)	Per-sample runtime (A100)
T5-base Encoder only	85.3%	0.097 s/it

# Results: ASR Repair + Interpretation Models

**State Exact Match:** # of commands for which the end-state is correctly predicted, whereby correctness is evaluated with exact string match.

Model	State EM	Per-command Runtime (s/it)
T5 (prog)	28.3%	1.28
T5 (state)	29.5%	3.46
GPT3 (prog)	38.6%	3.52
GPT3 (state)	55.1%	6.92



# Future Work

- Better evaluation of models
  - Human Evaluation
  - Partial credit
- Taking advantage of incrementality
- Include timing & prosody information in models
- Greater diversity in prompts and human voices
  - More open-ended prompts for more natural interactions
- Better/more flexible execution engine
- Model-in-the-loop annotation
  - Allows for data on clobber commands/redo/undo

# Thank you!

- Code & Data: <https://aka.ms/tertius>

# Conclusion

- We introduce a new task, **interactive dictation**, whereby:
  - 1. Users can naturally interleave dictation and commanding, and
  - 2. Users can flexibly invoke commands with a wide variety of utterances
- We construct a dataset **TERTiUS** for the task
- We build a baseline system for the task, discovering a tradeoff between speed and accuracy
  - We explore different choices and sizes of model architecture (T5 vs. GPT3)
  - We explore generating programs vs. generating document states directly