UWNLP

TOYOTA TECHNOLOGICAL INSTITUTE CHICAGO

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Parsing Speech: A Neural Approach to Integrating Lexical and Acoustic-Prosodic Information

Trang Tran*[1], Shubham Toshniwal*[2], Mohit Bansal[3], Kevin Gimpel[2], Karen Livescu[2], Mari Ostendorf[1]

[1]Electrical Engineering, University of Washington
[2]Toyota Technological Institute at Chicago
[3]Computer Science, UNC Chapel Hill

*Equal Contribution

Challenges in Parsing Speech

- Voice-based HCI more widely used → parsing speech (and NLP for speech) more important
- Speech vs. text:
    - Speech lacks clues for conventional parsing (punctuation, case, …)
    - ASR (and human) errors in transcribed speech are common
    - Speech has disfluent components (filled pauses, *[edits]*, …)

Wall Street Journal:

Pierre Vinken , 61 years old , will join the board as a non executive director Nov. 29 .

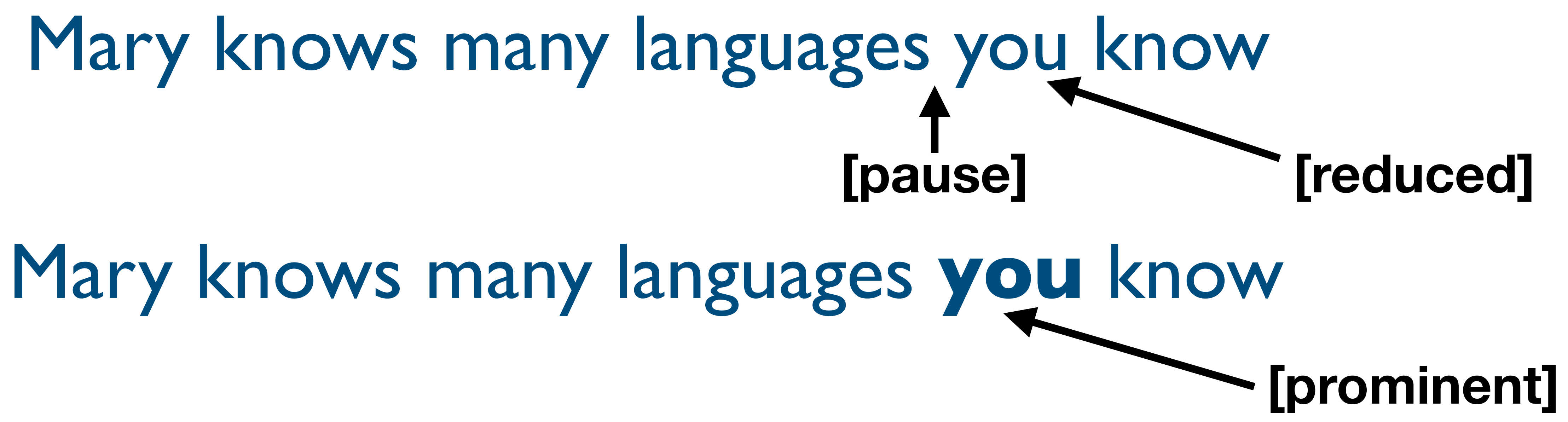
Switchboard:

and uh *[we were]* i was fortunate in that i was personally acquainted with the uh people who uh ran the nursing home in our little hometown

Prosody and Parsing

- Prosody
  - Symbolic level: phrase boundaries (constituents) and prominence (stress, pitch accent)
  - Acoustic cues: pauses, word/syllable lengthening, pitch (f0) contour, energy, voice quality

- Prosodic information in the acoustic signal can help parsing
  - Prosodic cues signal disfluencies (interruption points)
  - Prosodic boundaries align with constituent boundaries (Grosjean et al., 1979)
  - Boundary and prominence help resolve ambiguities (Price et al., 1991)



Mary knows many languages you know

[pause]

[reduced]

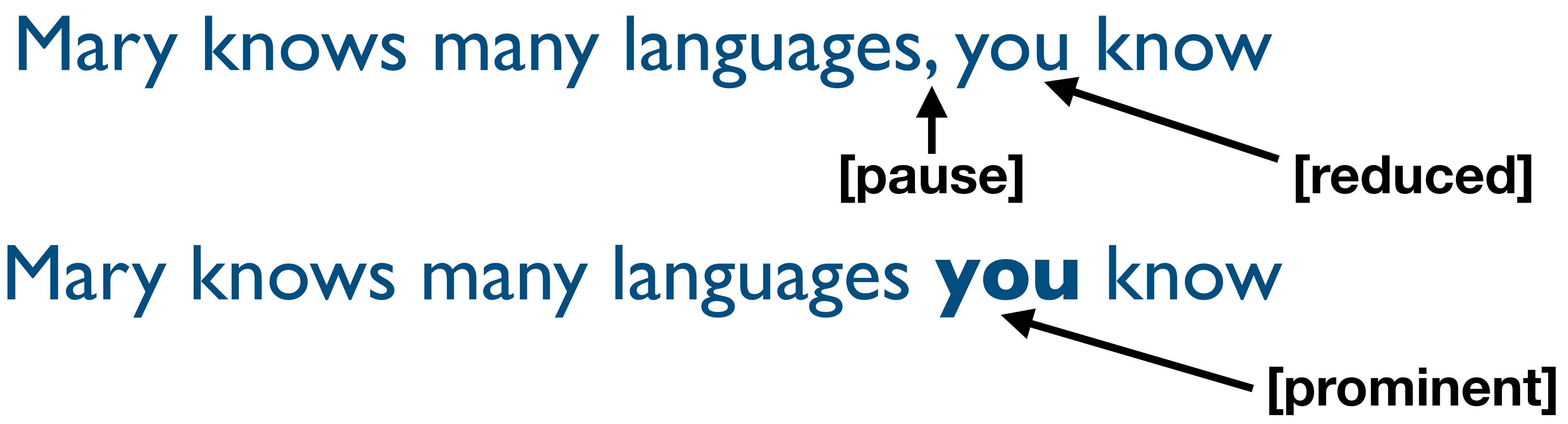
Mary knows many languages **you** know

[prominent]

Prosody and Parsing

- Prosody
  - Symbolic level: phrase boundaries (constituents) and prominence (stress, pitch accent)
  - Acoustic cues: pauses, word/syllable lengthening, pitch (f0) contour, energy, voice quality

- Prosodic information in the acoustic signal can help parsing
  - Prosodic cues signal disfluencies (interruption points)
  - Prosodic boundaries align with constituent boundaries (Grosjean et al., 1979)
  - Boundary and prominence help resolve ambiguities (Price et al., 1991)



Mary knows many languages, you know

[pause] [reduced]

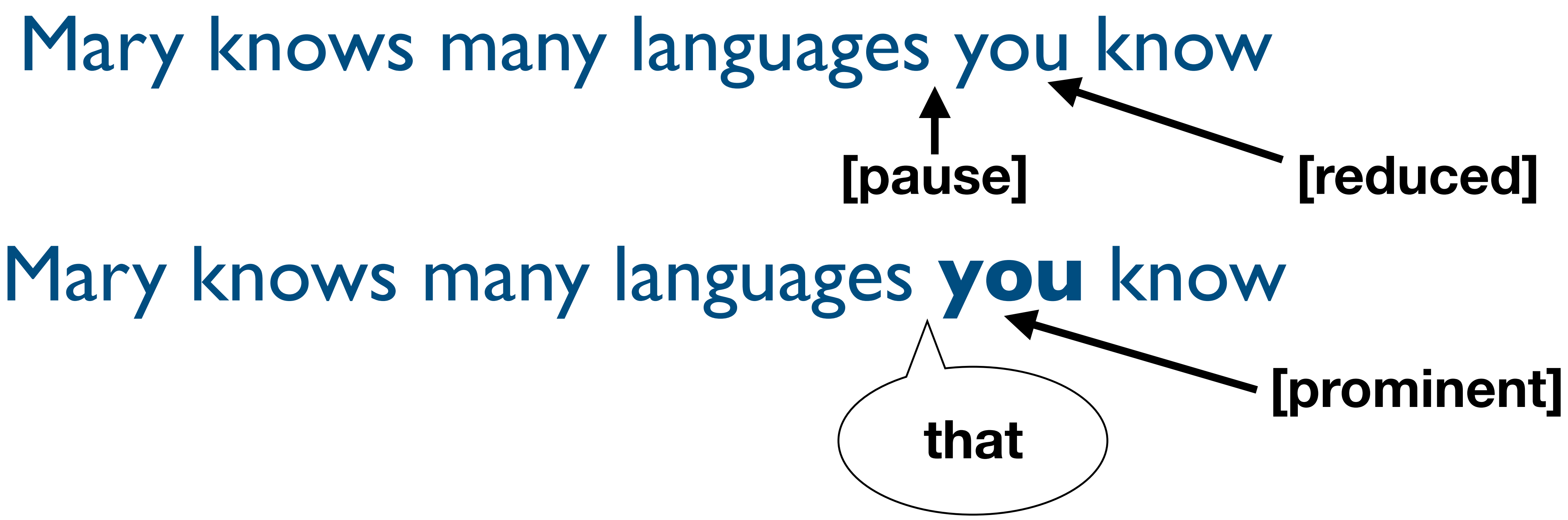
Mary knows many languages **you** know

[prominent]

Prosody and Parsing

- Prosody
  - Symbolic level: phrase boundaries (constituents) and prominence (stress, pitch accent)
  - Acoustic cues: pauses, word/syllable lengthening, pitch (f0) contour, energy, voice quality

- Prosodic information in the acoustic signal can help parsing
  - Prosodic cues signal disfluencies (interruption points)
  - Prosodic boundaries align with constituent boundaries (Grosjean et al., 1979)
  - Boundary and prominence help resolve ambiguities (Price et al., 1991)



Mary knows many languages you know

[pause]

[reduced]

Mary knows many languages **you** know

that

[prominent]

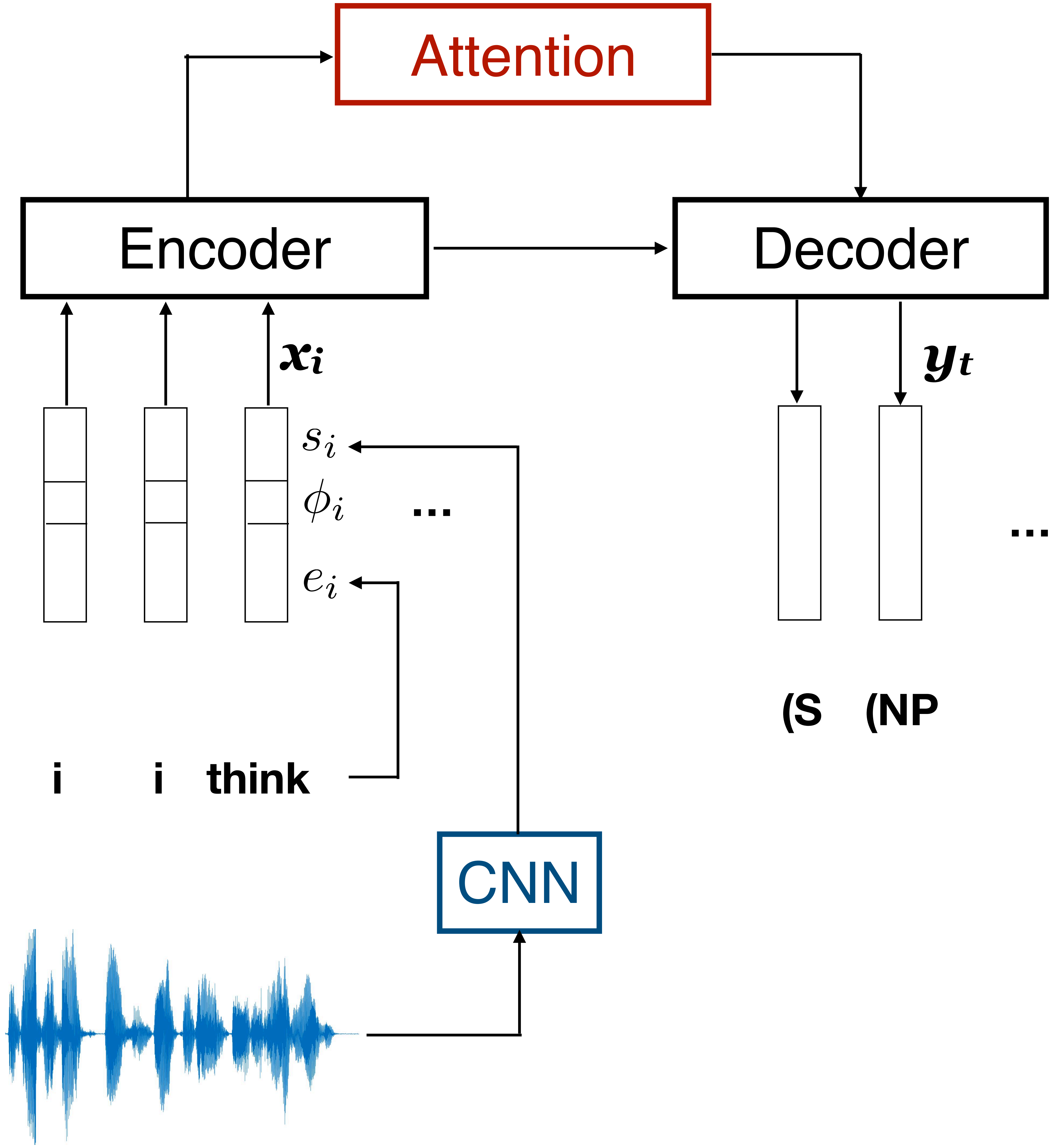
Using Prosody

- Prior work:
  - Most gains were obtained in unknown sentence boundary setting (Kahn and Ostendorf, 2012)
  - Need expensive human annotations (Kahn et al., 2005; Hale et al., 2006; Dreyer and Shafran, 2007)
  - Direct use of acoustic cues and sentence-internal prosody seemed to hurt parsing (Gregory et al., 2004)

- Our contributions:
  - Framework for integrating acoustic-prosodic features without prosodic labels
  - Gains in using sentence-internal prosody: disfluent sentences, reduced attachment errors
  - Assessment of transcription error effects on utility of prosody

Task and Model Overview

- Encoder-decoder with attention (Vinyals et al., 2015)
  - Input: word-level features
    $\boldsymbol{x_i} = [e_i, (s_i, \phi_i)]$
    - e_i : word embeddings
    - ϕ_i: pause and duration features
    - s_i: f0/E features
  - Output: linearized parse symbols $\boldsymbol{y_t}$
- Location-aware attention (Chorowski et al., 2015)
- CNN-learned pitch/energy features $\boldsymbol{s_i}$



Attention Mechanism

- Standard attention (global/content-only):

$$c_t = \sum_{i=1}^{T_s} \alpha_{i,t} h_i$$

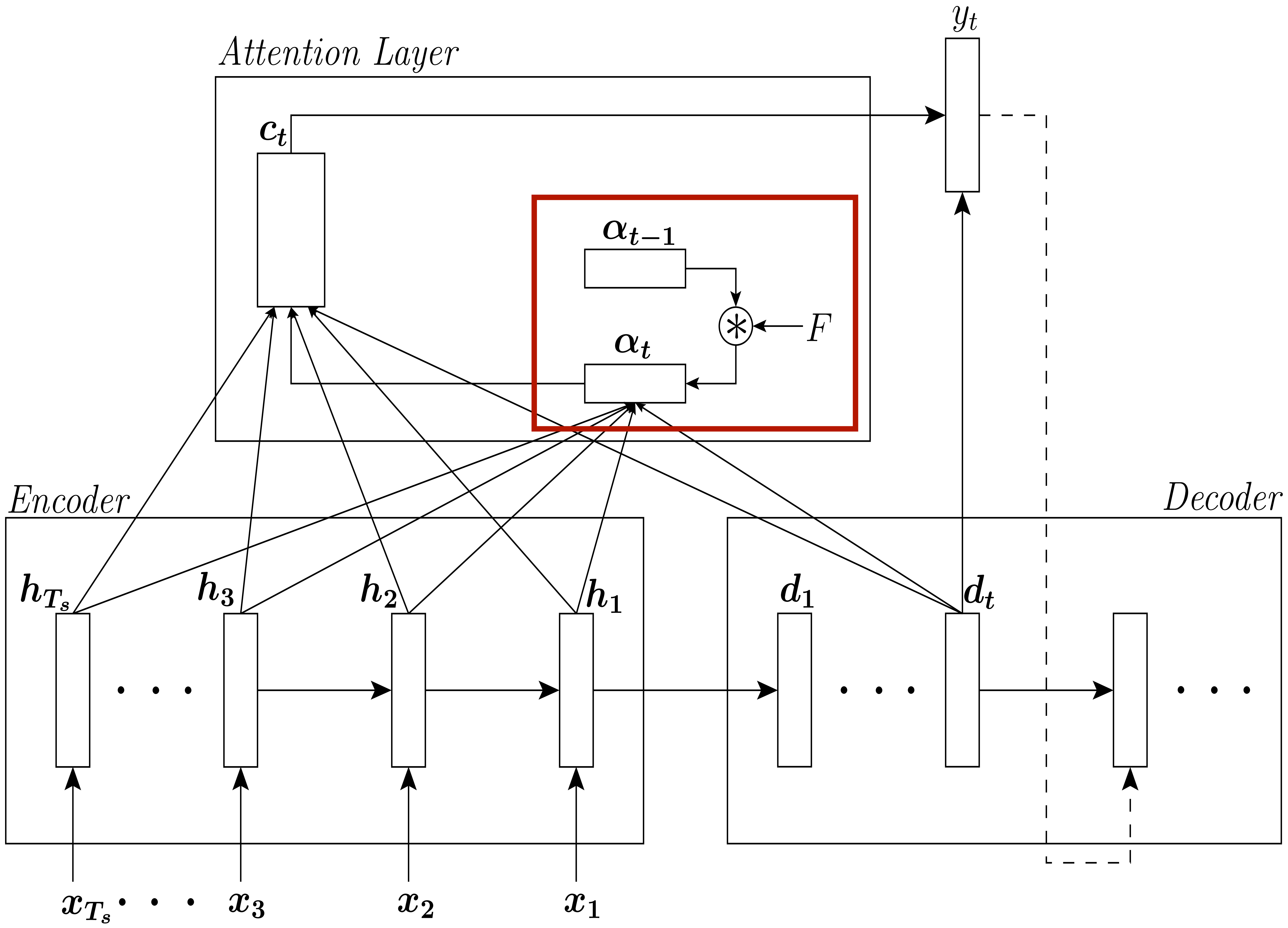
$$\alpha_t = \mathrm{softmax}(u_t)$$

$$u_{i,t} = f(h_i, d_t)$$

- Convolutional attention (content+location):

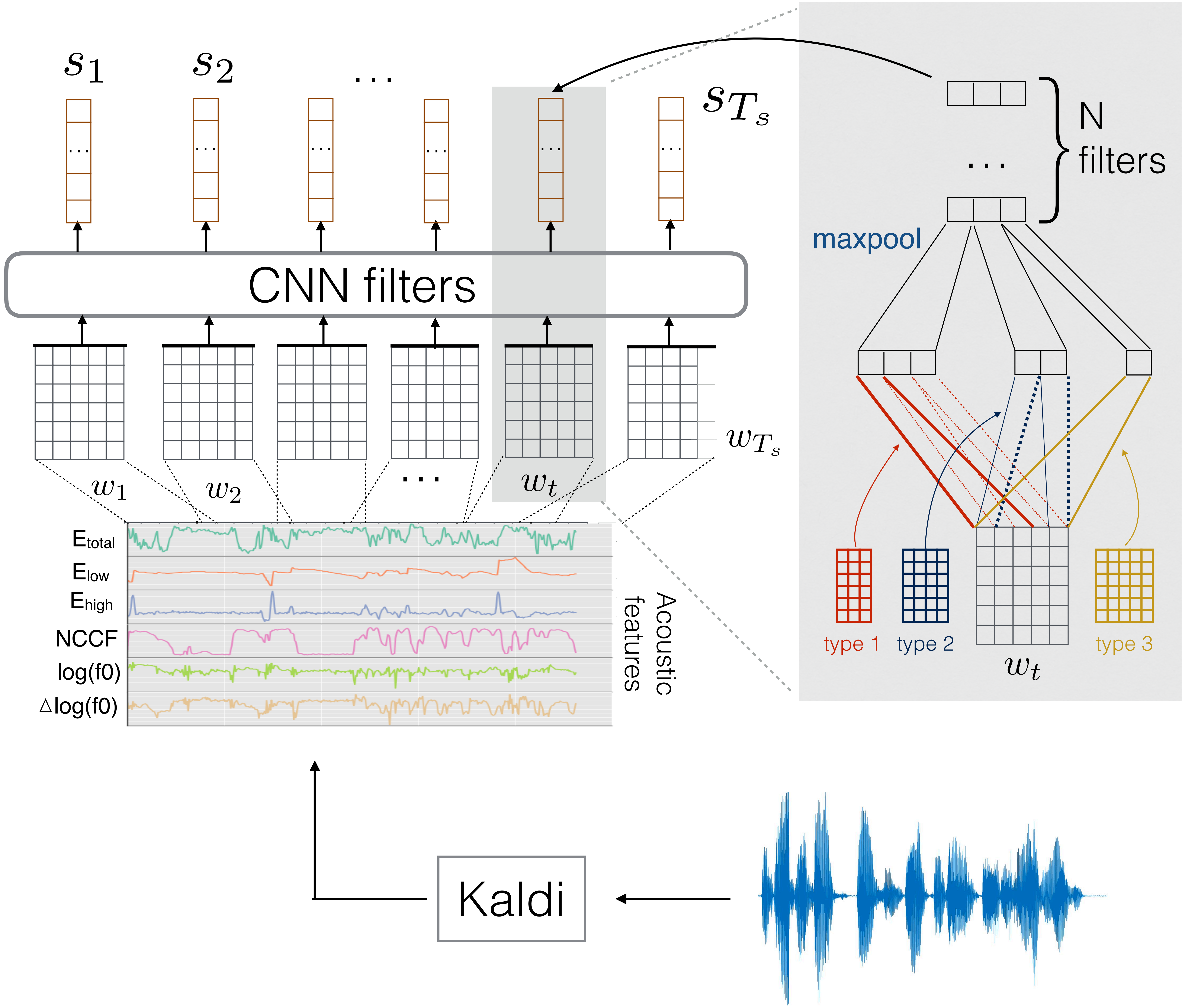
  (Chorowski et al., 2015)

$$u_{i,t} = f(h_i, d_t, F * \alpha_{t-1})$$



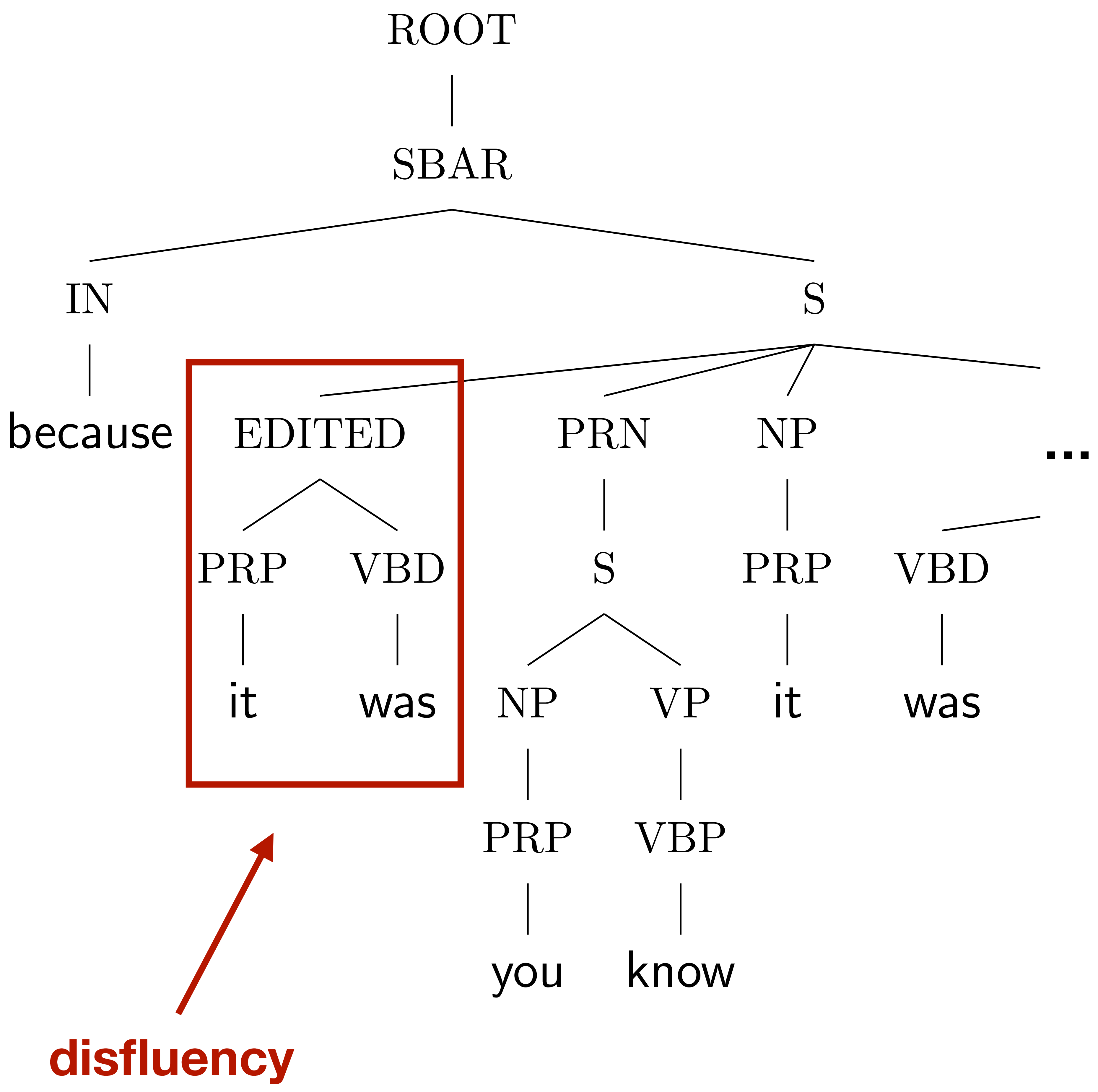
CNN-learned Acoustic-Prosodic Features

- Pause (p)
  - Before and after
  - Bin and embed
- Word duration (d)
- Pitch and energy contours (f0/E)
  - Learned via CNN
  - Frame-level filters capturing sub-word, word, word boundary context

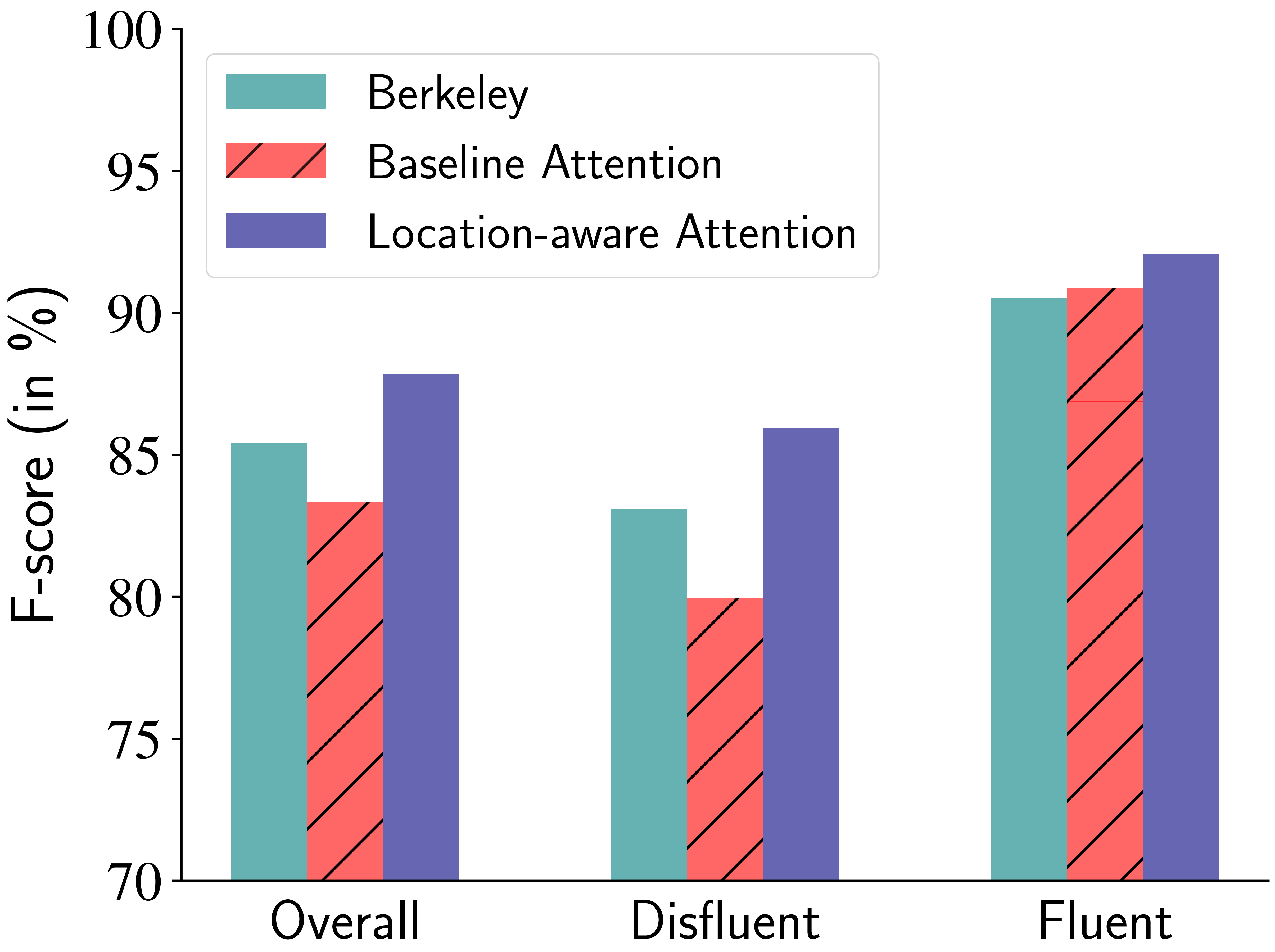


Data and Metrics

- Data
  - Switchboard NXT (Calhoun et al., 2010)
  - 642 telephone conversations
  - 100K sentences, 14K vocabulary
- Metrics
  - Parseval F1 (label and span)
  - Disfluency F1 (detection)

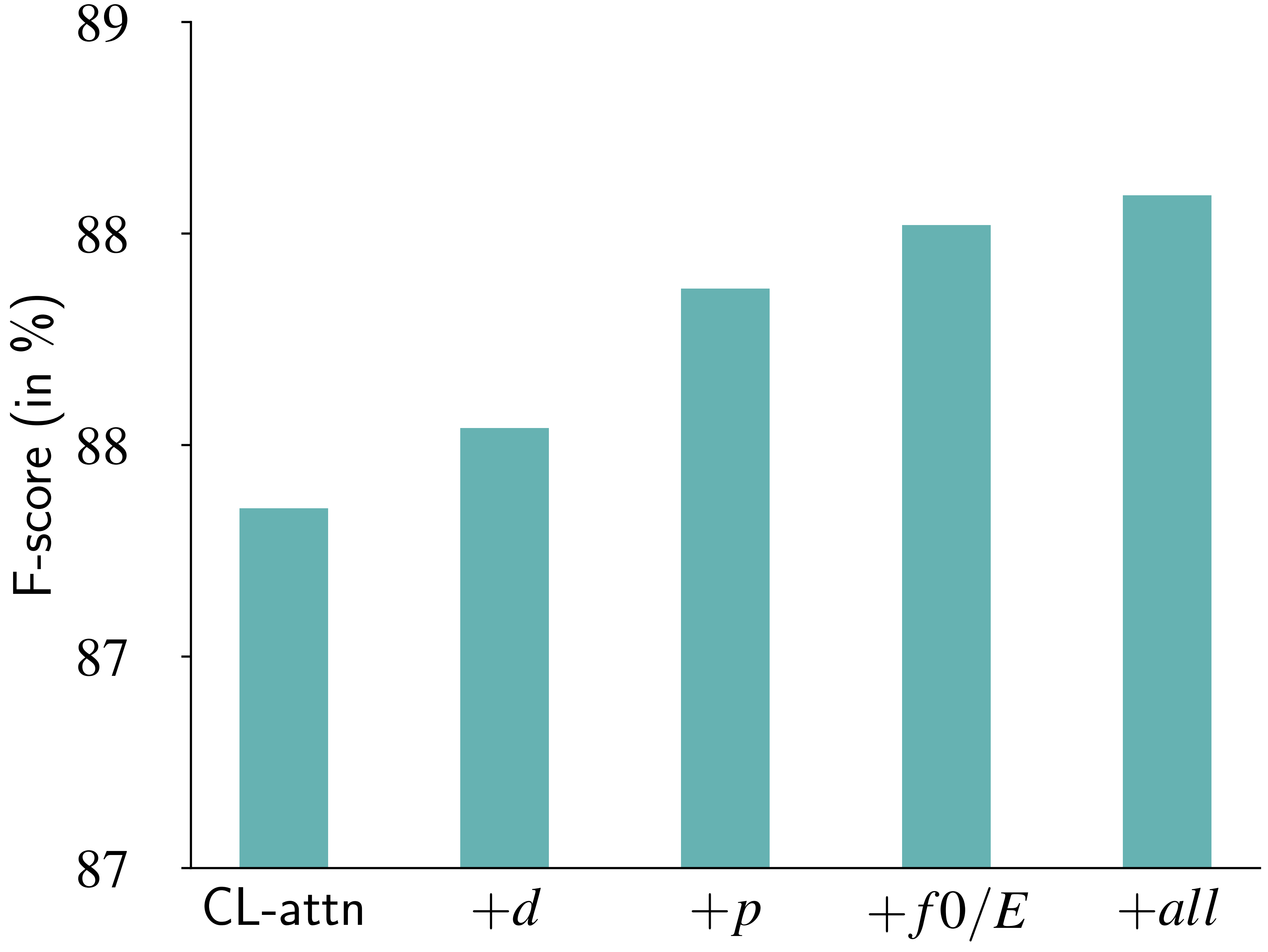


Results: Text-Only & Baselines (dev set)

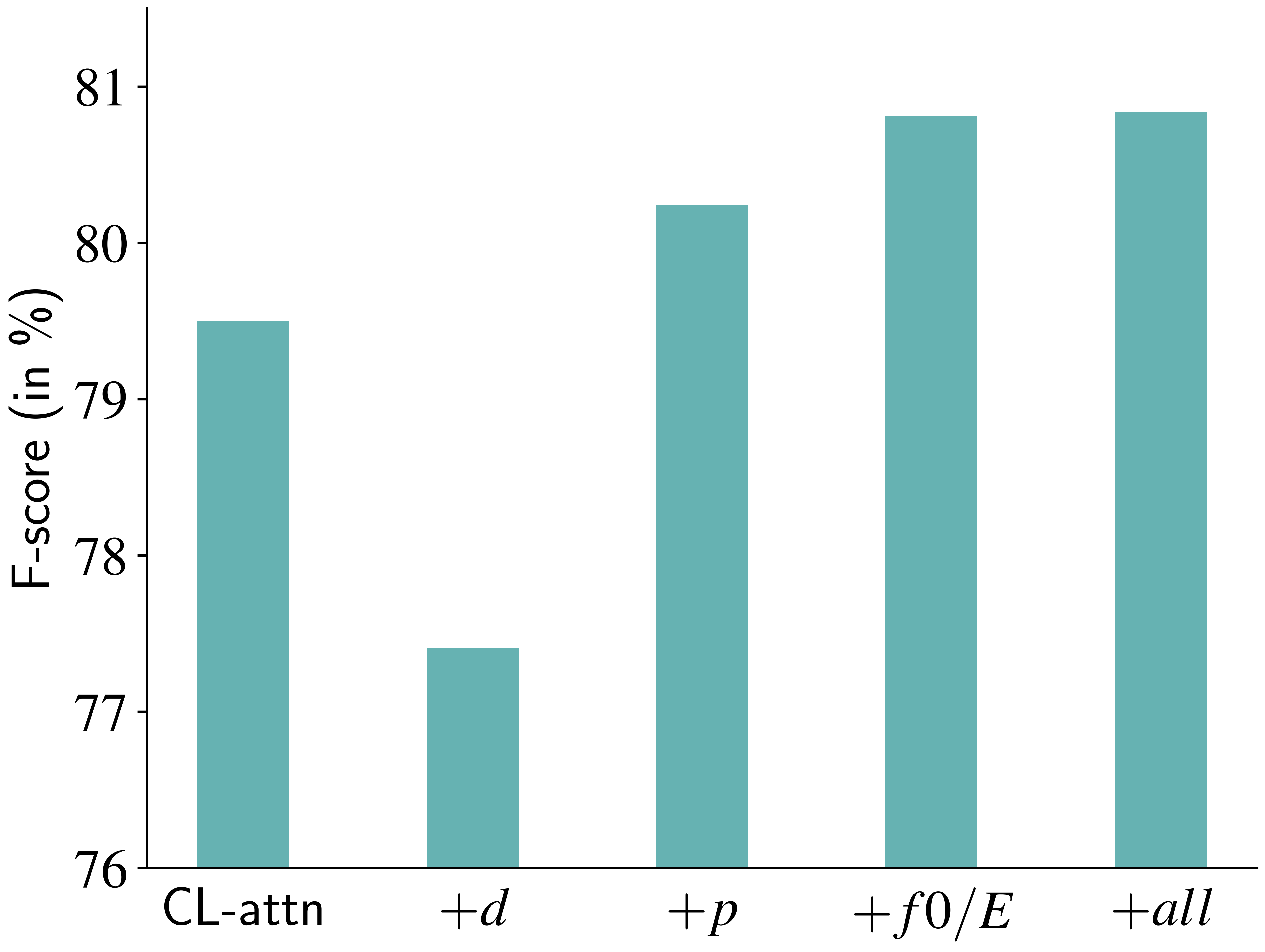


- Location-aware attention (CL-attn) overcomes problems of baseline in handling disfluencies
- Use CL-attn for the rest of the experiments

Results: Text + Prosody (dev set)



Parse F1 Results



Disfluency F1 Results

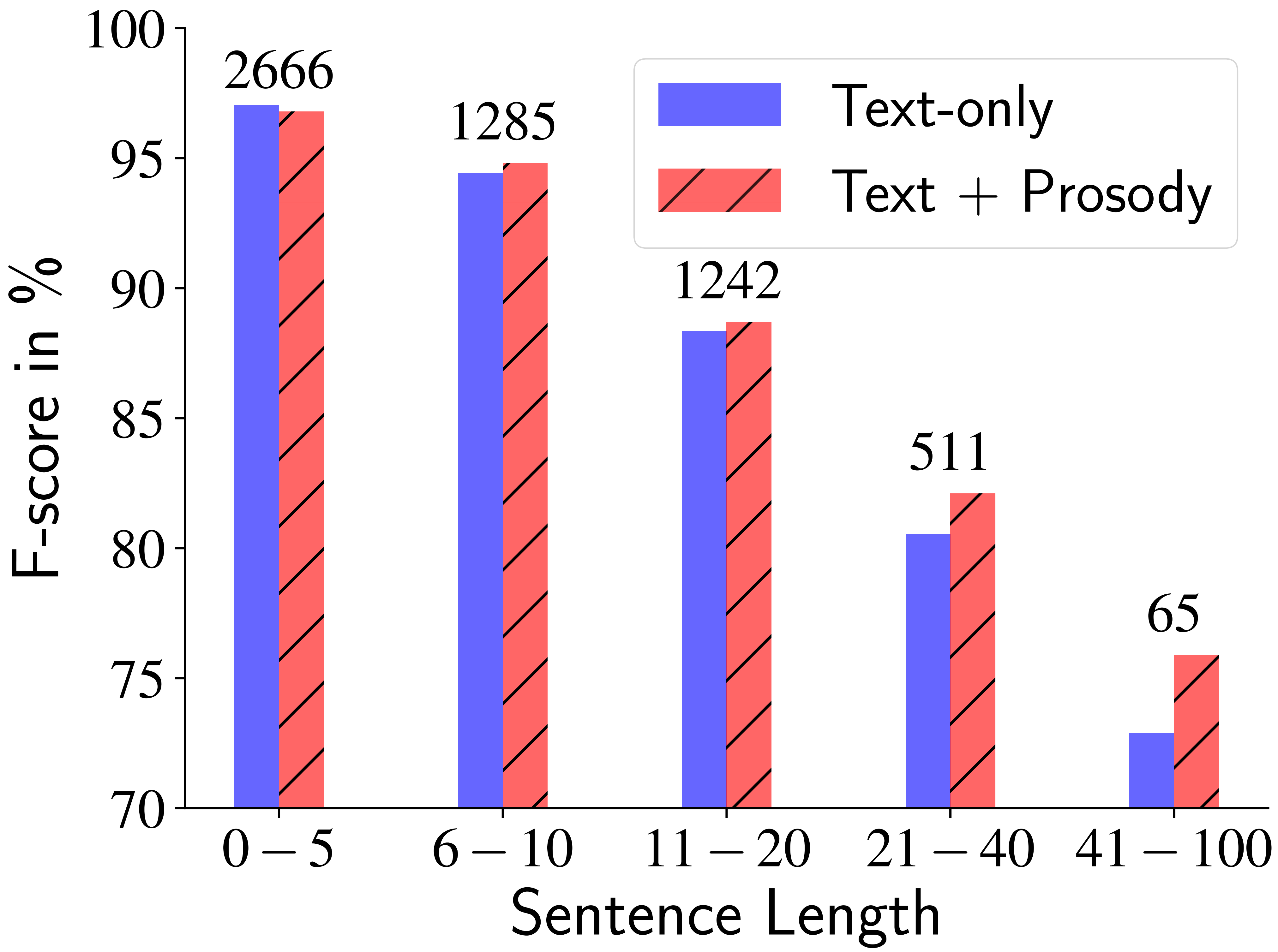
- Adding acoustic-prosodic features helps
- Pause and f0/E contribute most of the gain

Comparison with Previous Work (test set)

Model	Text-only	Text+Prosody	Rel. (1-F) reduction
Kahn et al., 2005	86.4	86.6	+1.5%
Hale et al., 2006	71.2	71.1	-0.3%
CL-attn	88.0	88.5	+4.2%

- Slightly different training data and experiment settings → compare relative performance
- We are gaining more over text-only baselines
- Results (text vs. text + prosody) are statistically significant (p-value < 0.02)

Analysis: Sentence Types

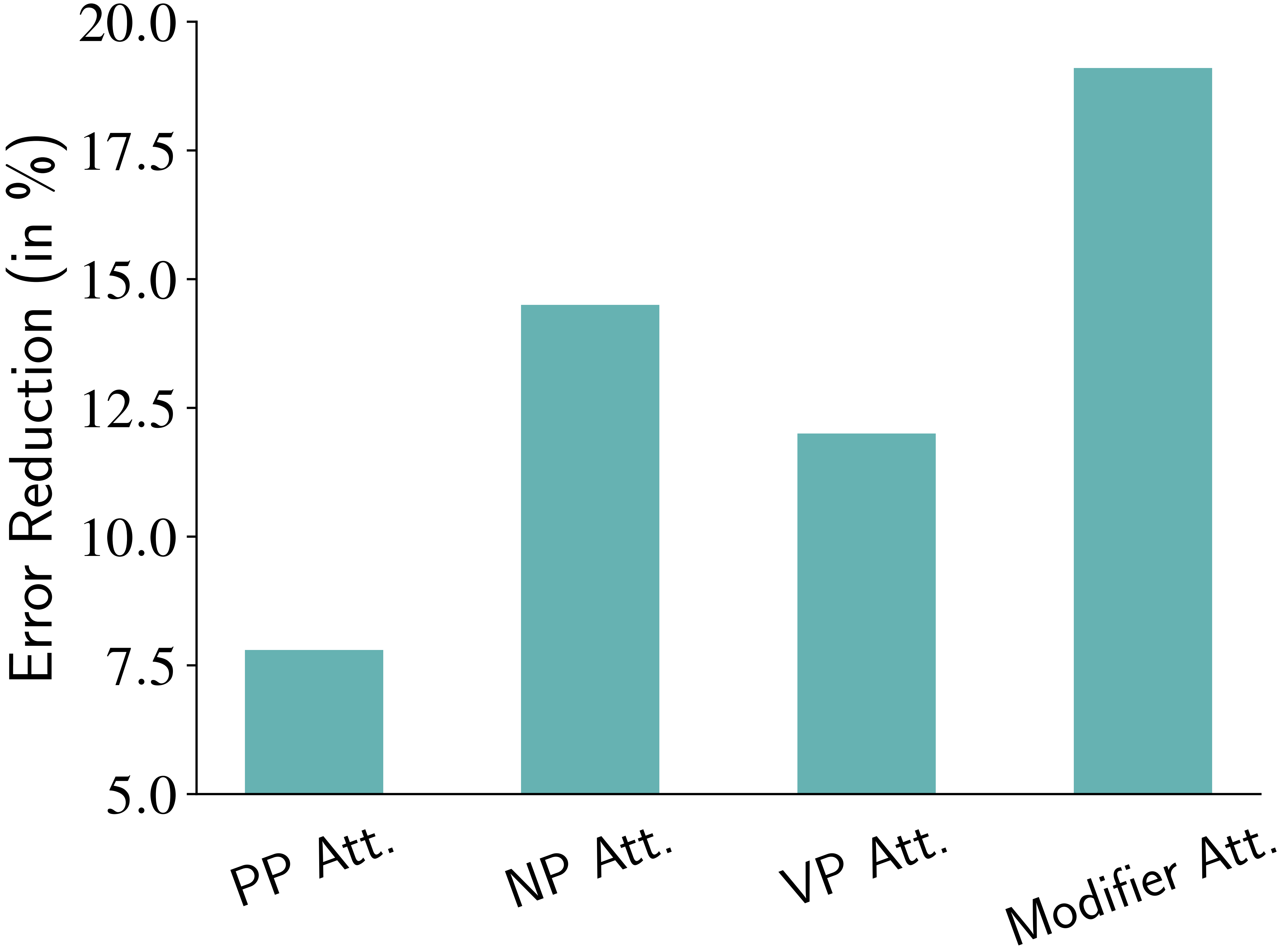


Model	Fluent	Disfluent
Text-only	92.07	85.90
Text + Prosody	92.03	87.02

Prosody helps in longer sentences

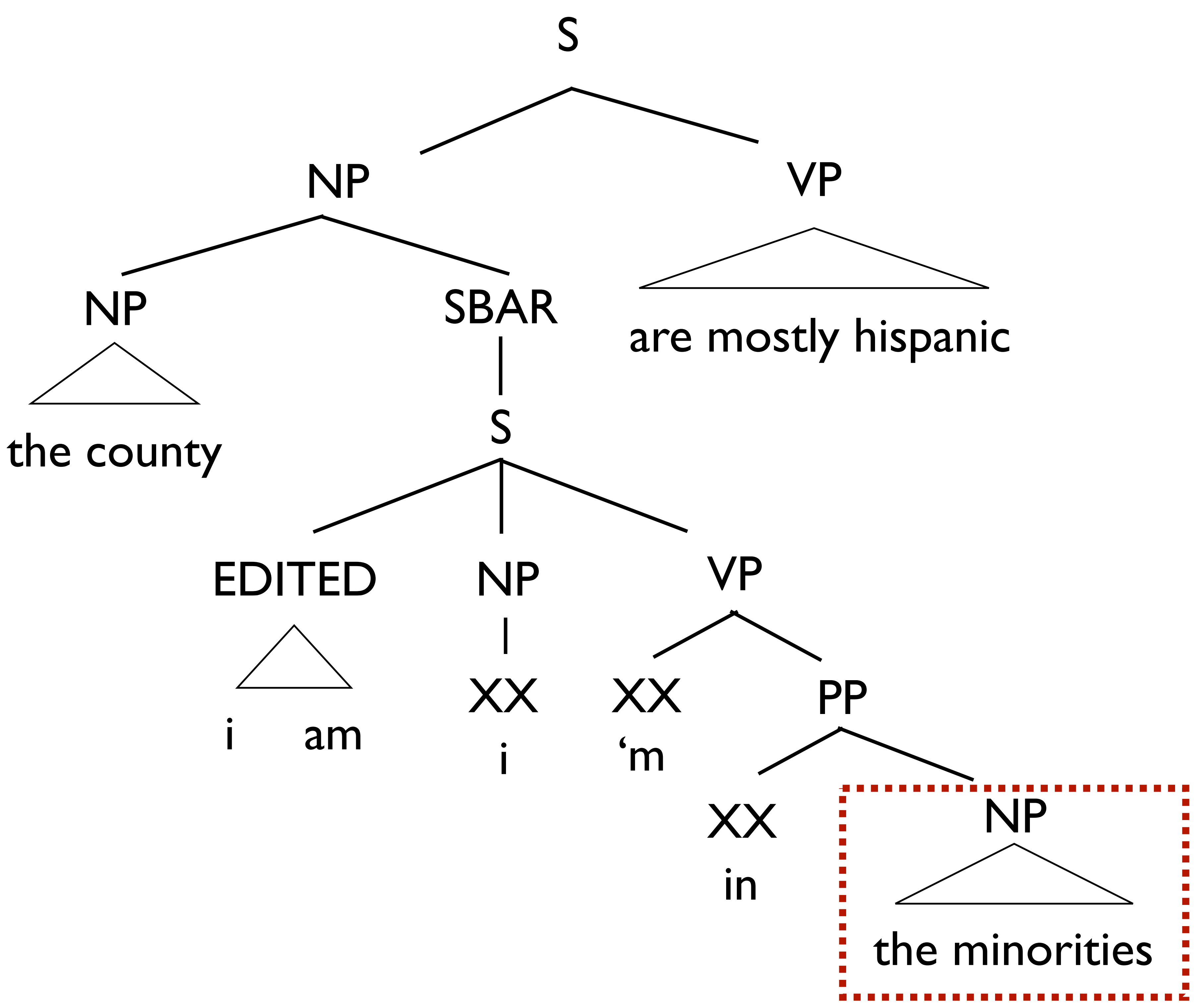
Prosody helps in disfluent sentences

Analysis: Parse Error Types

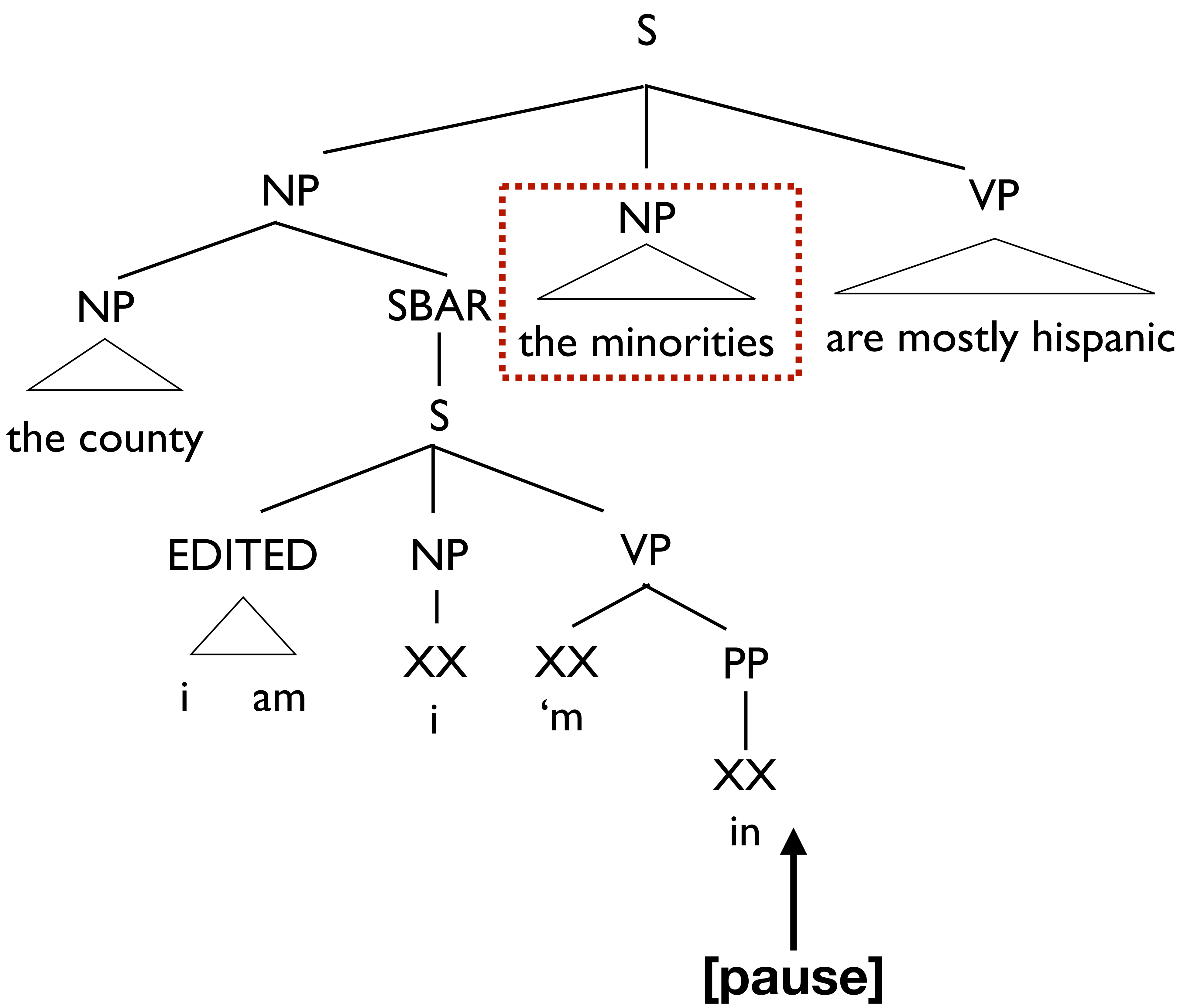


- Error classifications from Berkeley Parser Analyzer (Kummerfeld et al., 2012)
- Prosody helps most in reducing attachment errors

Analysis: Parse Error Example



Text-only Parser



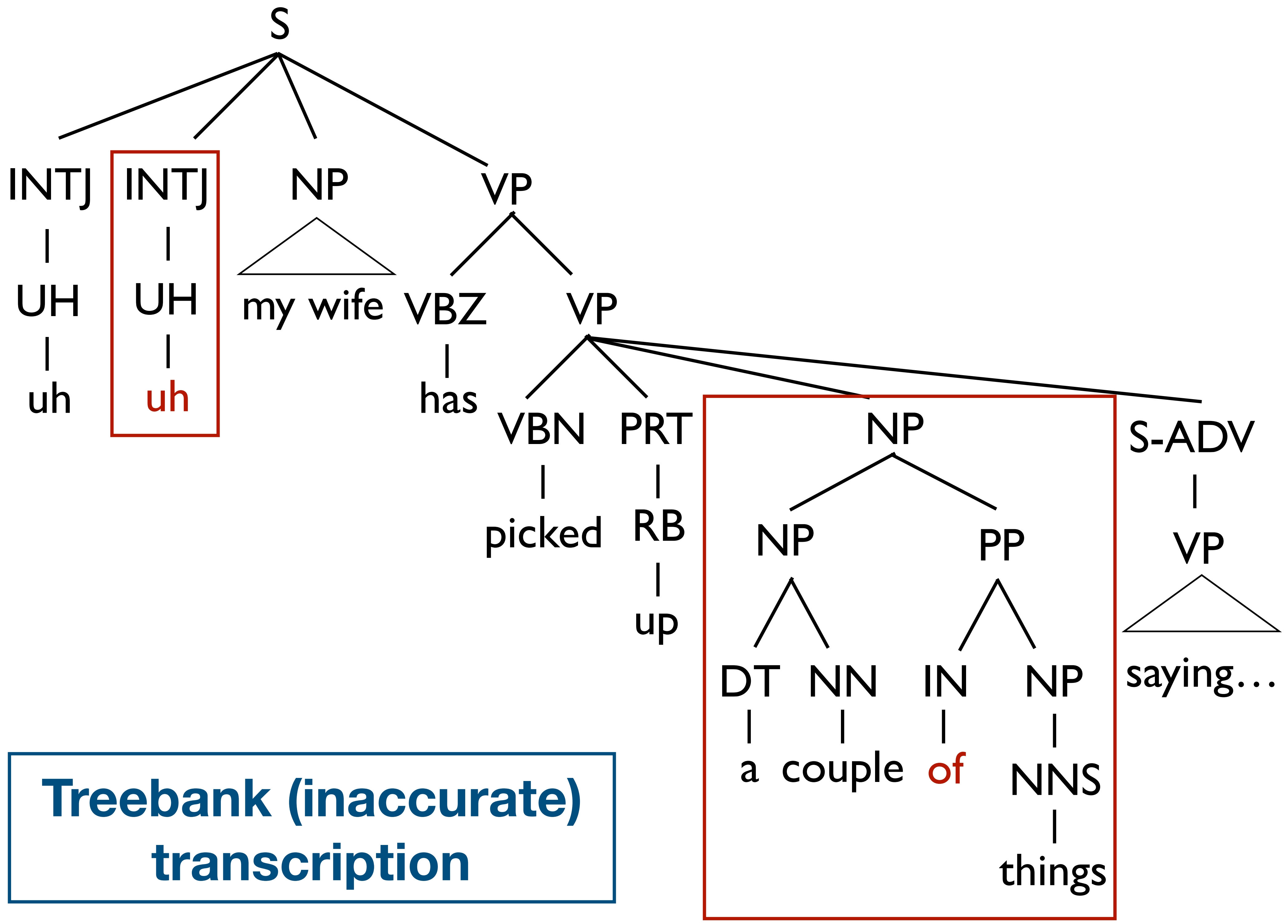
Text+Prosody Parser

Prosody (pause) helped avoid attachment error

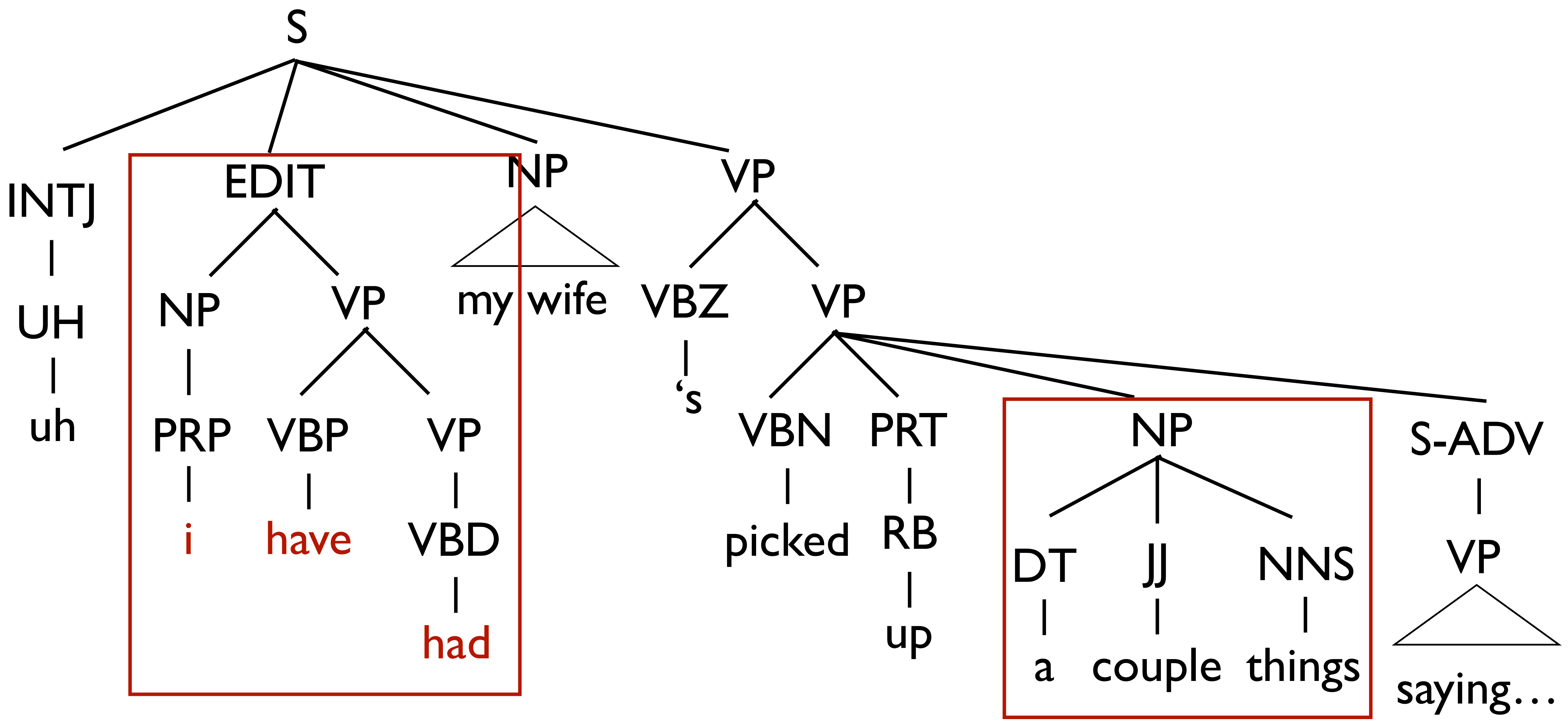
Analysis: Transcription Error Effects

- Prosody seems to hurt in fluent sentences, what is going on?
- Compare parser performance on sentences with and without transcription errors
- Errors result in inconsistent prosody features

# Fluent sentences	Prosody helped	Prosody “hurt”
with errors	57	82
no errors	270	269



Treebank (inaccurate) transcription



MS-State (accurate) transcription

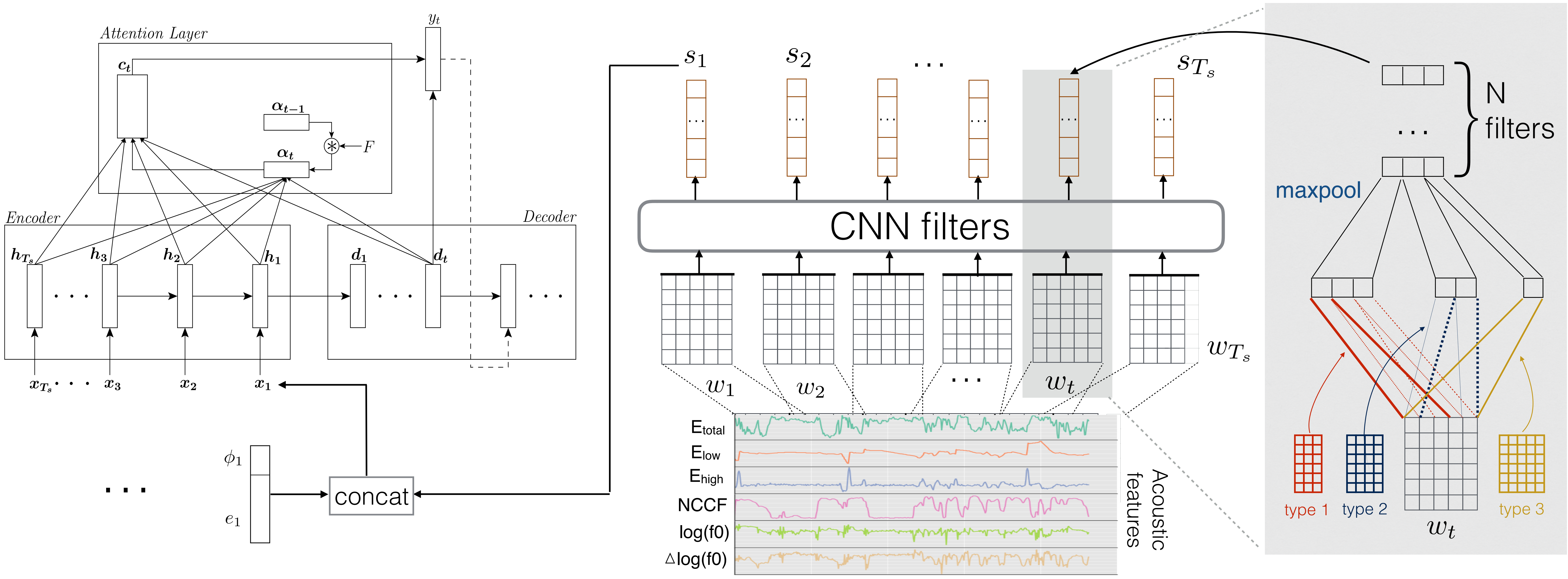
Conclusion

- Contributions:
  - Framework for automatically integrating acoustic-prosodic features, which previously was a challenge
  - For sentence-internal structure, prosody helps:
    - in disfluent and long sentences
    - in reducing attachment errors
  - Gain from prosody has been underestimated due to transcription errors
- Future:
  - Extend to other parsing frameworks (dependency) and systems (transition-based)
  - Assess impact with unknown sentence boundaries and ASR errors
  - Transfer parses to accurate transcripts

Thank you!

Backup Slides

Full model details



$$c_t = \sum_{i=1}^{T_s} \alpha_{ti} h_i$$

$$\alpha_t = \mathrm{softmax}(u_t)$$

$$u_{it} = v^\top \tanh(W_1 h_i + W_2 d_t + b_a)$$

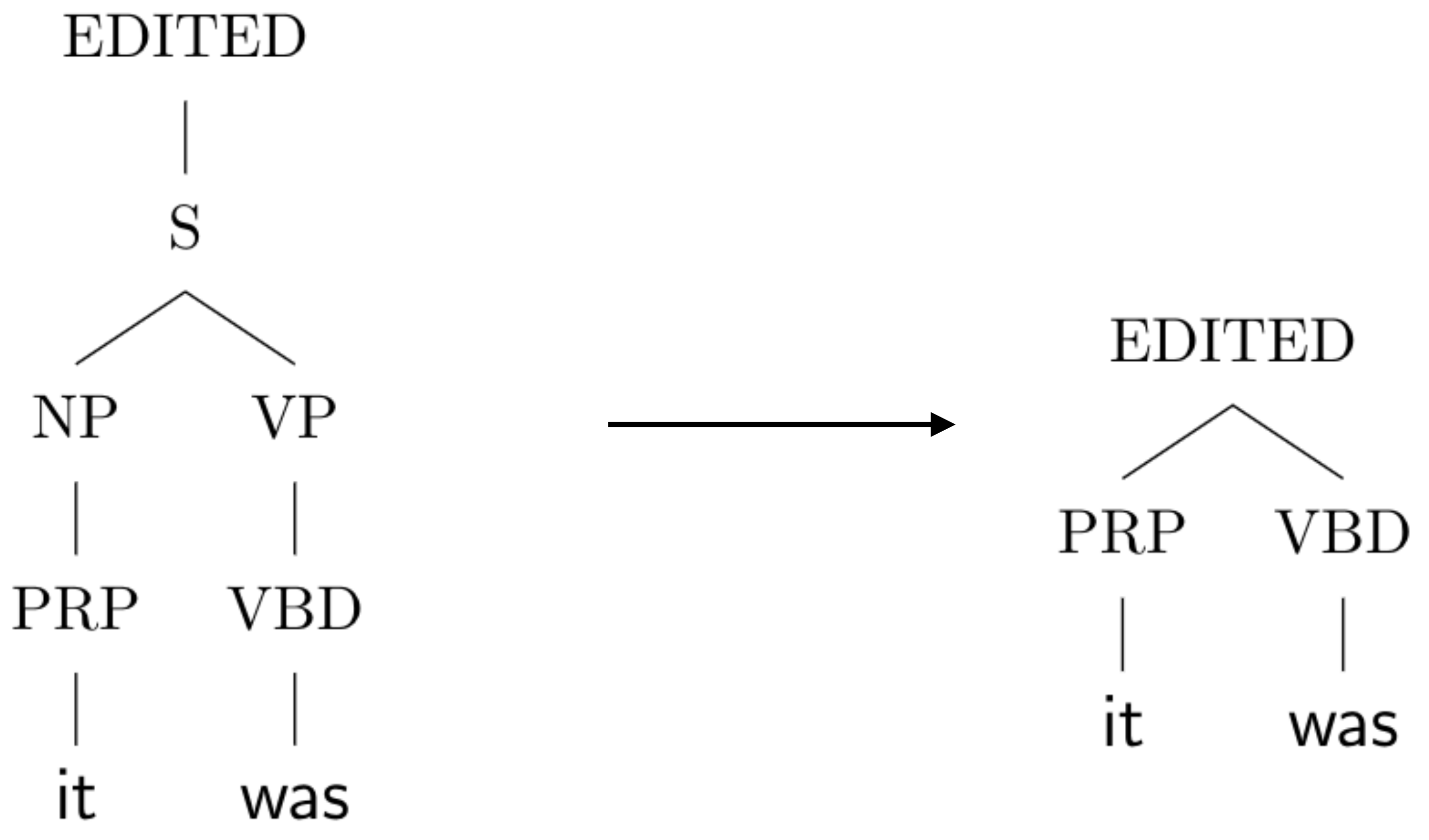
$$u_{it} = v^\top \tanh(W_1 h_i + W_2 d_t + W_f f_{ti} + b_a)$$

Data and Metrics (details)

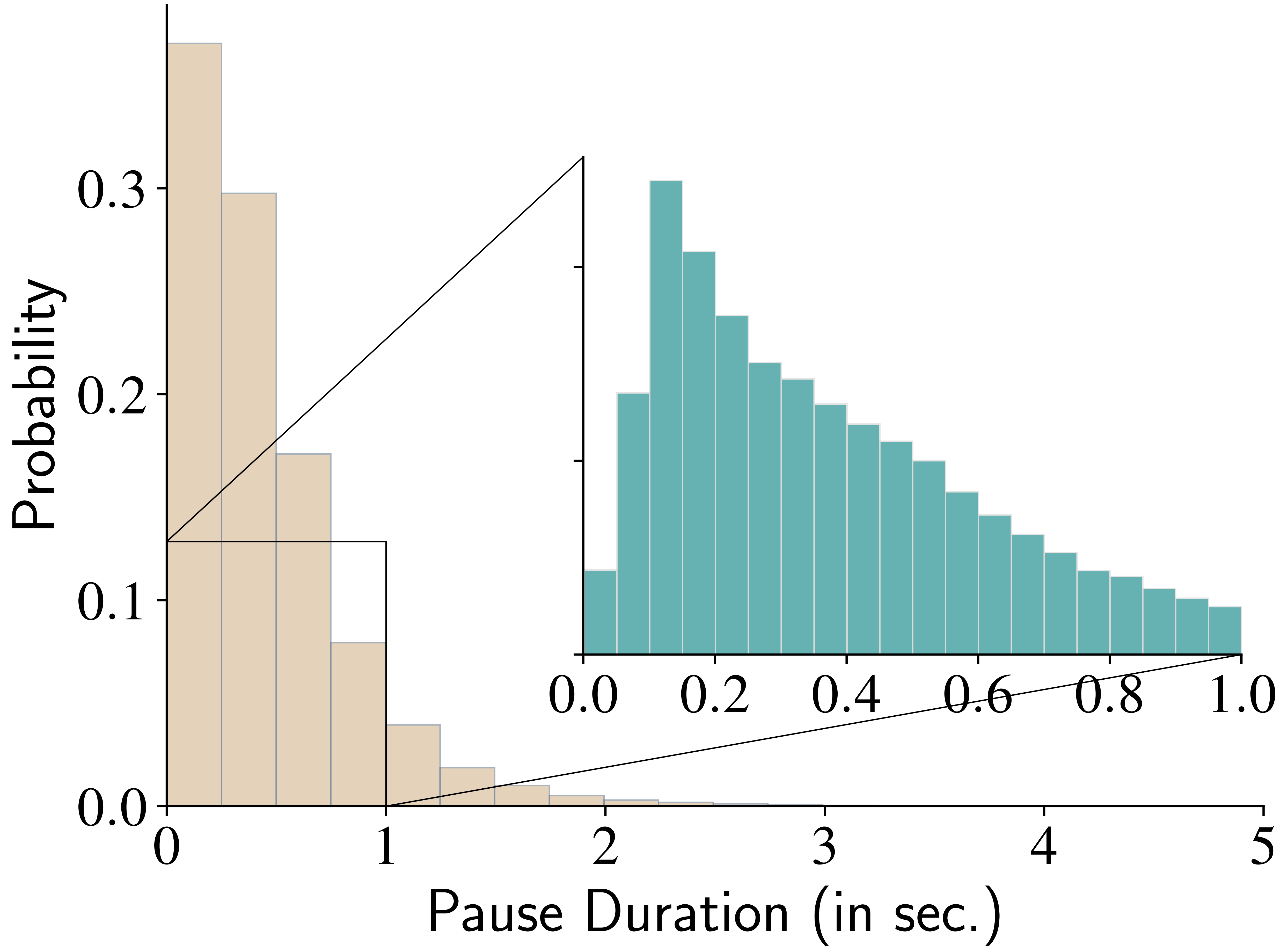
- Data
  - Switchboard NXT (Calhoun et al., 2010)
  - 642 conversations
  - Train/Dev/Test splits follow previous work (e.g. Charniak and Johnson, 2001)
  - Vocabulary: 14k

- Metrics
  - Standard Parseval F1
  - Flattened EDIT Parseval F1

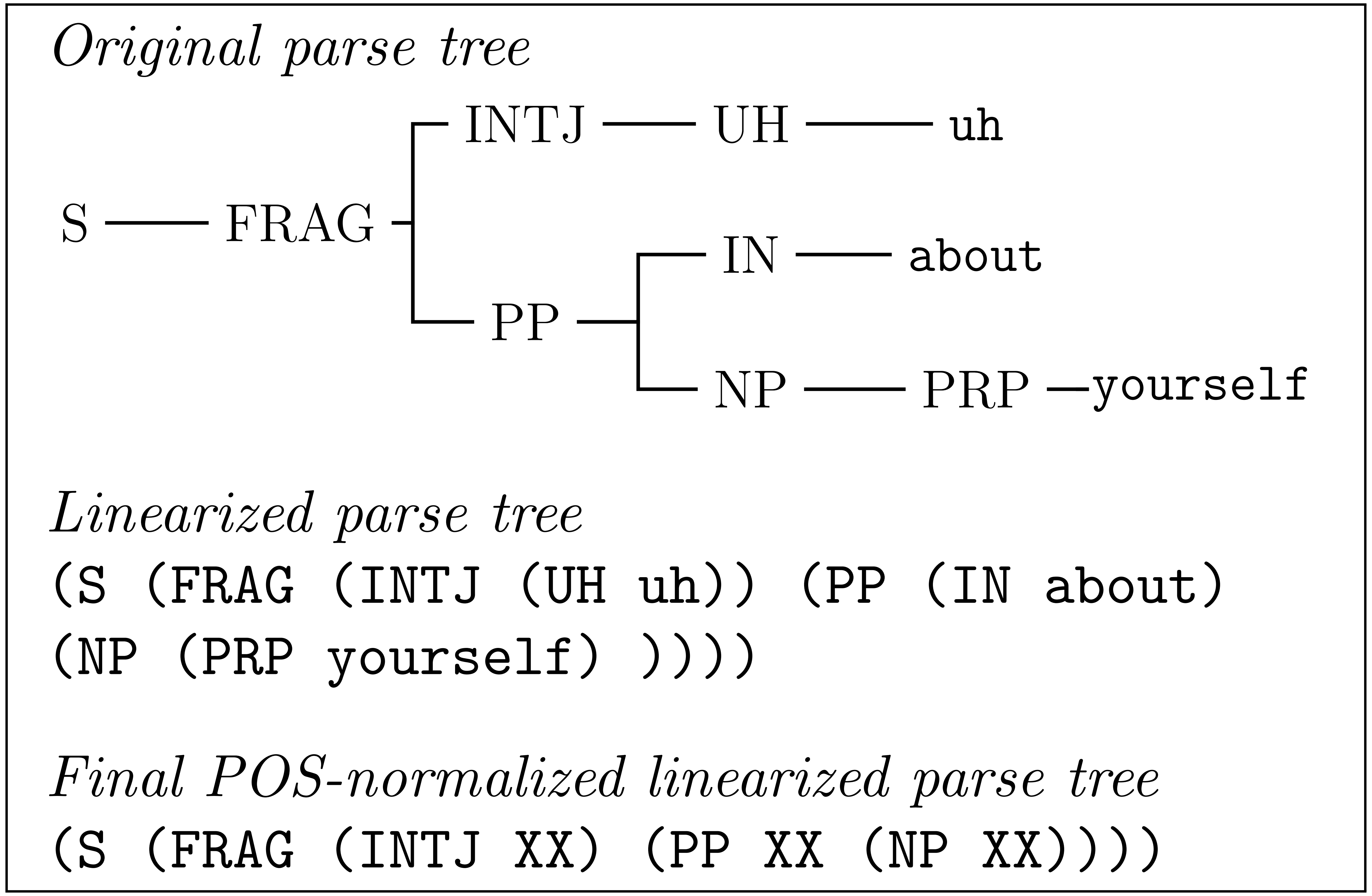
Split	# sentences	# tokens
Train	97,113	729,252
Dev	5,769	50,445
Test	5,901	48,625



Pause duration distribution



Preprocessing



Original parse tree

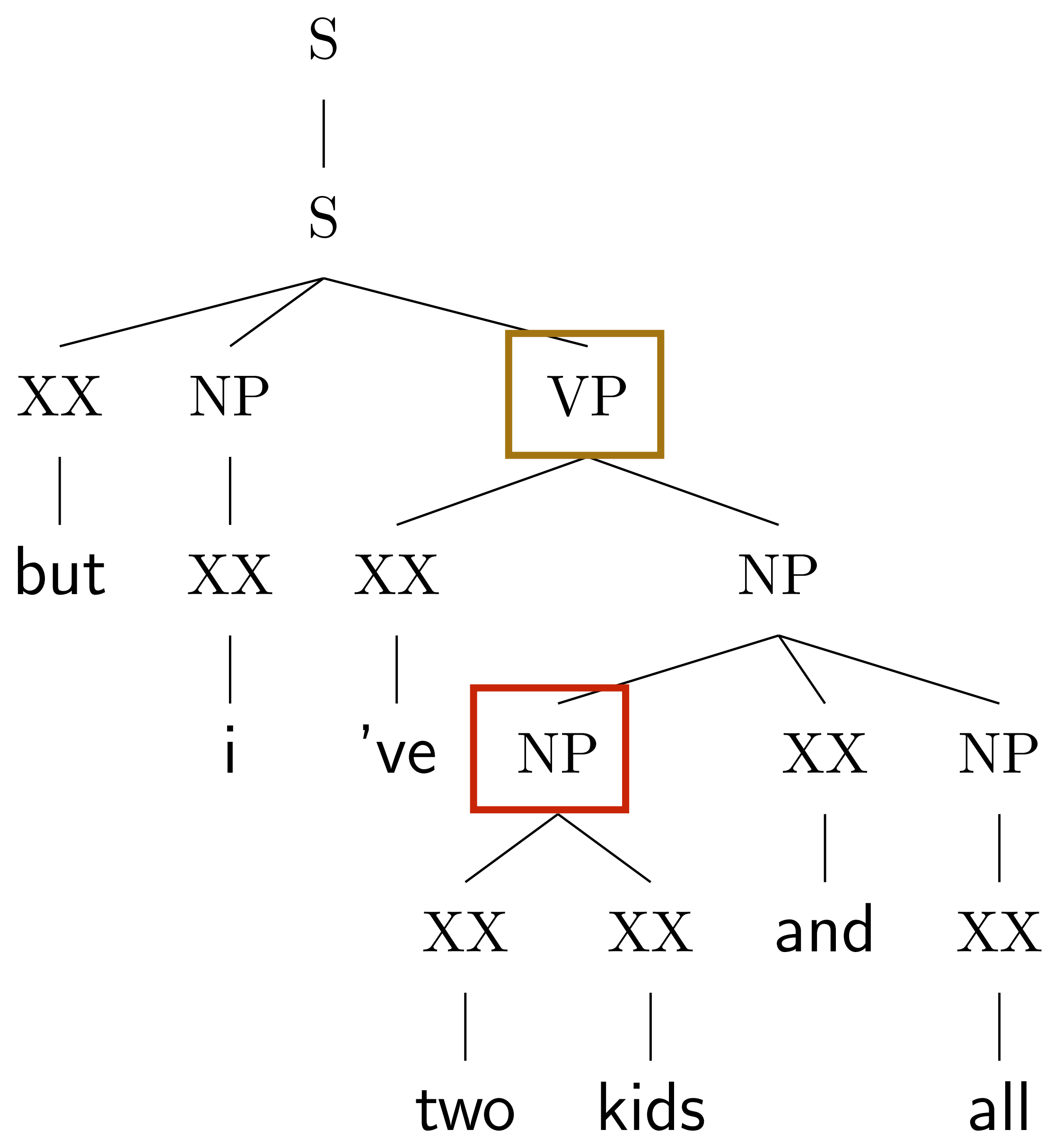
Linearized parse tree

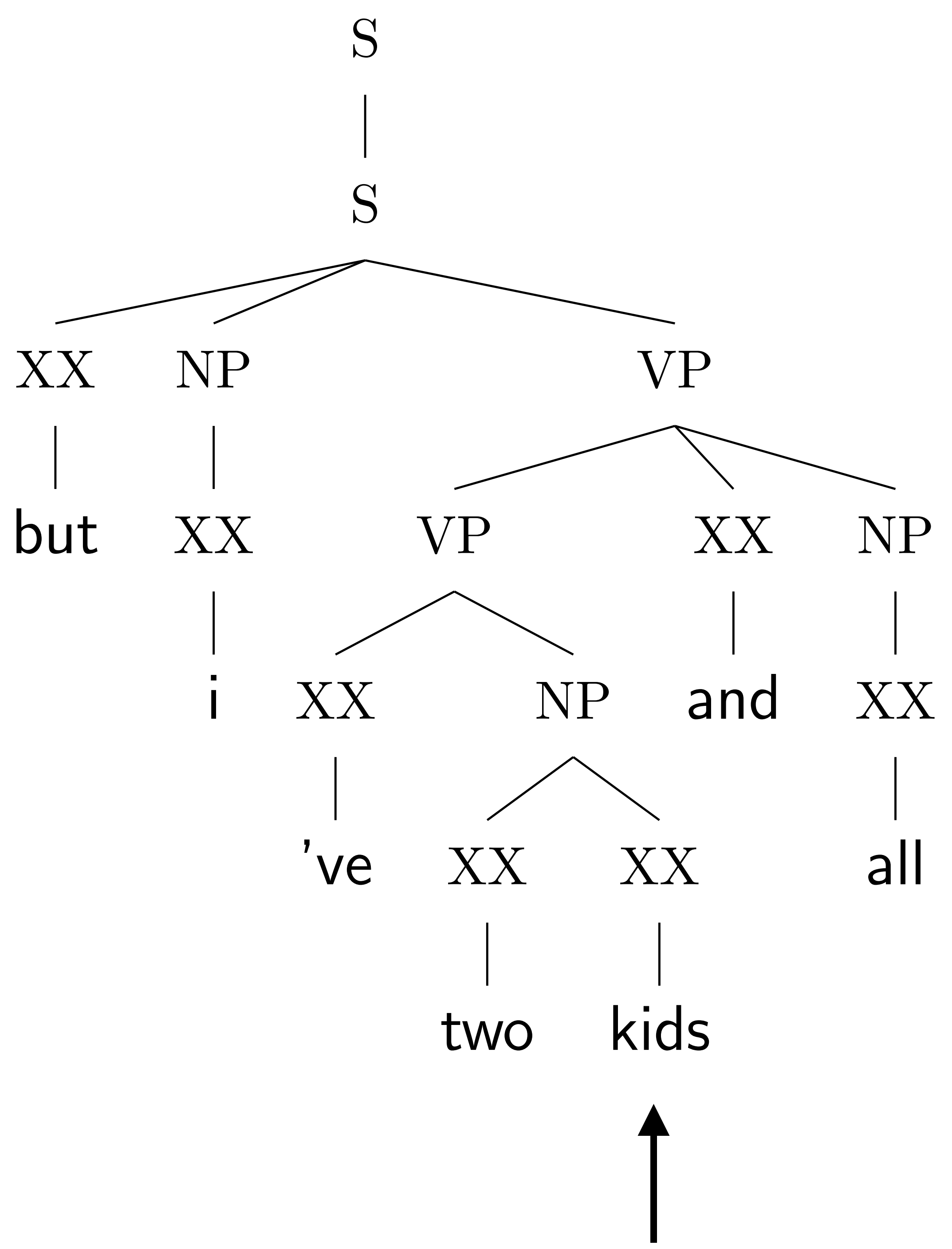
```
(S (FRAG (INTJ (UH uh)) (PP (IN about)
(NP (PRP yourself) ))))
```

Final POS-normalized linearized parse tree

```
(S (FRAG (INTJ XX) (PP XX (NP XX))))
```

Another NP attachment error example

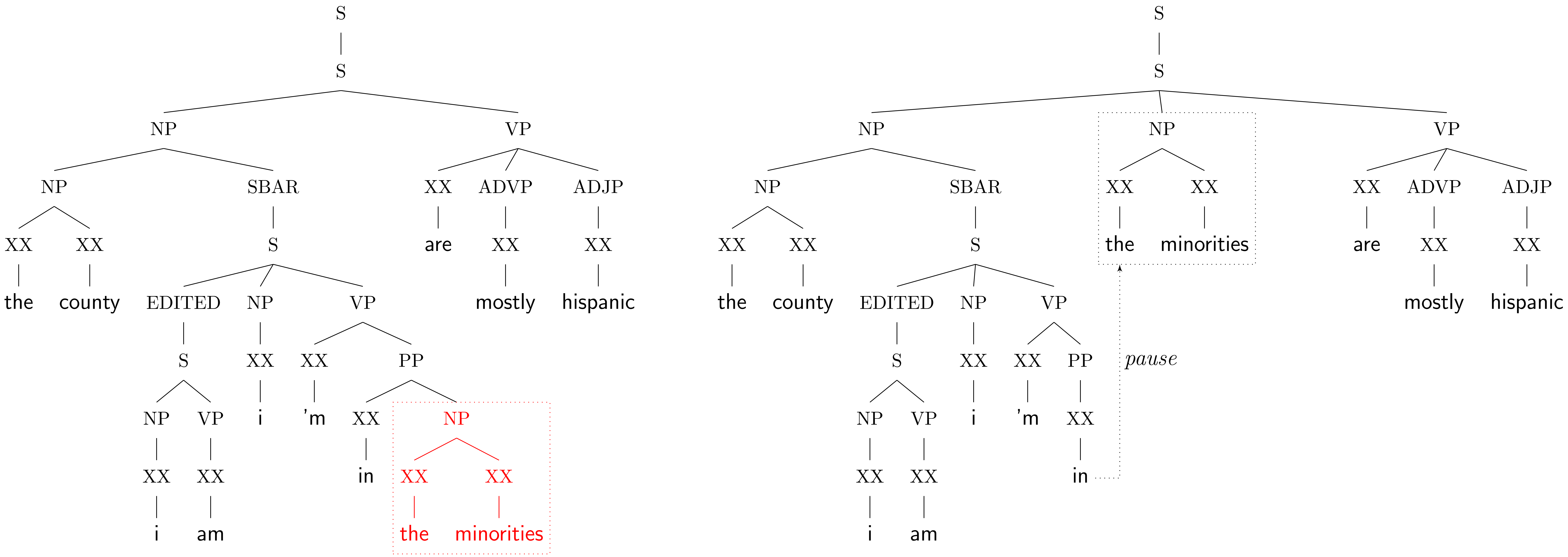




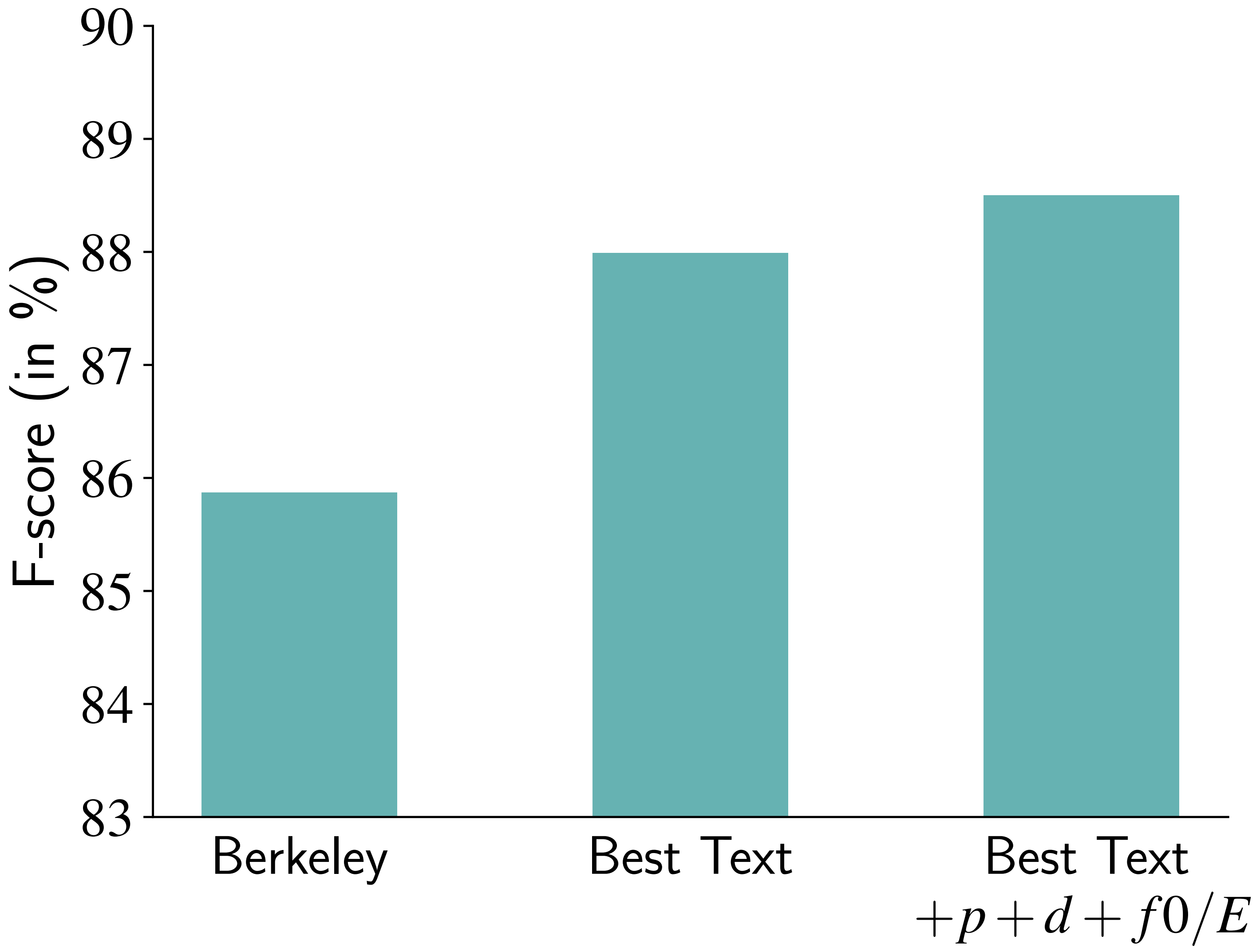
Text-only

Text + prosody

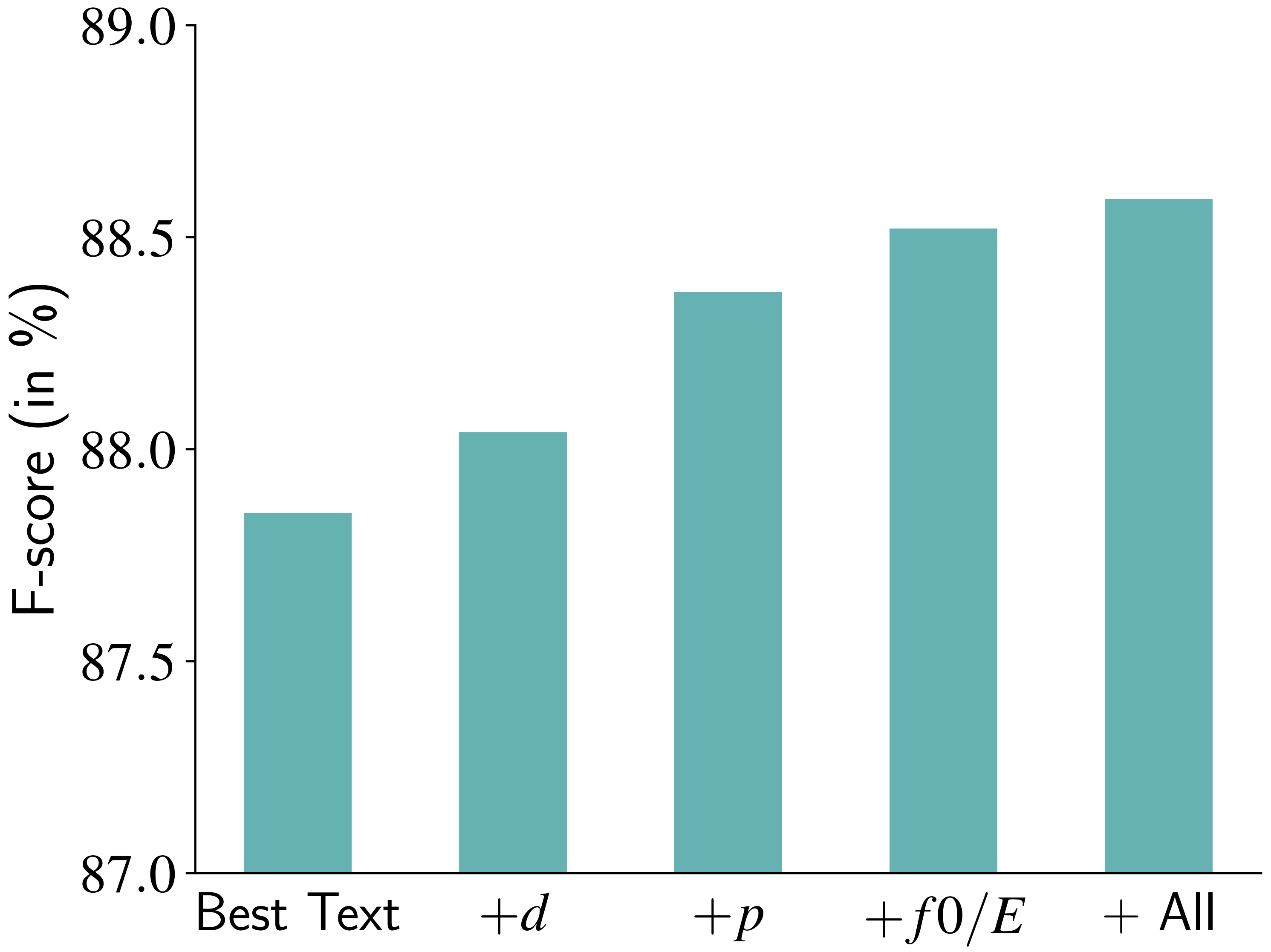
Original fig from paper



Results in different format



Test Set



Ablations

More Transcription Error Examples

- Parse structure changes:

  and because <uh> like if your spouse died <all of a sudden you be> all alone it 'd be nice to go someplace with people similar to you to have friends

- Disfluent → Fluent:

  uh uh <i have had> my wife 's picked up a couple of things saying uh boy if we could refinish that 'd be a beautiful piece of furniture

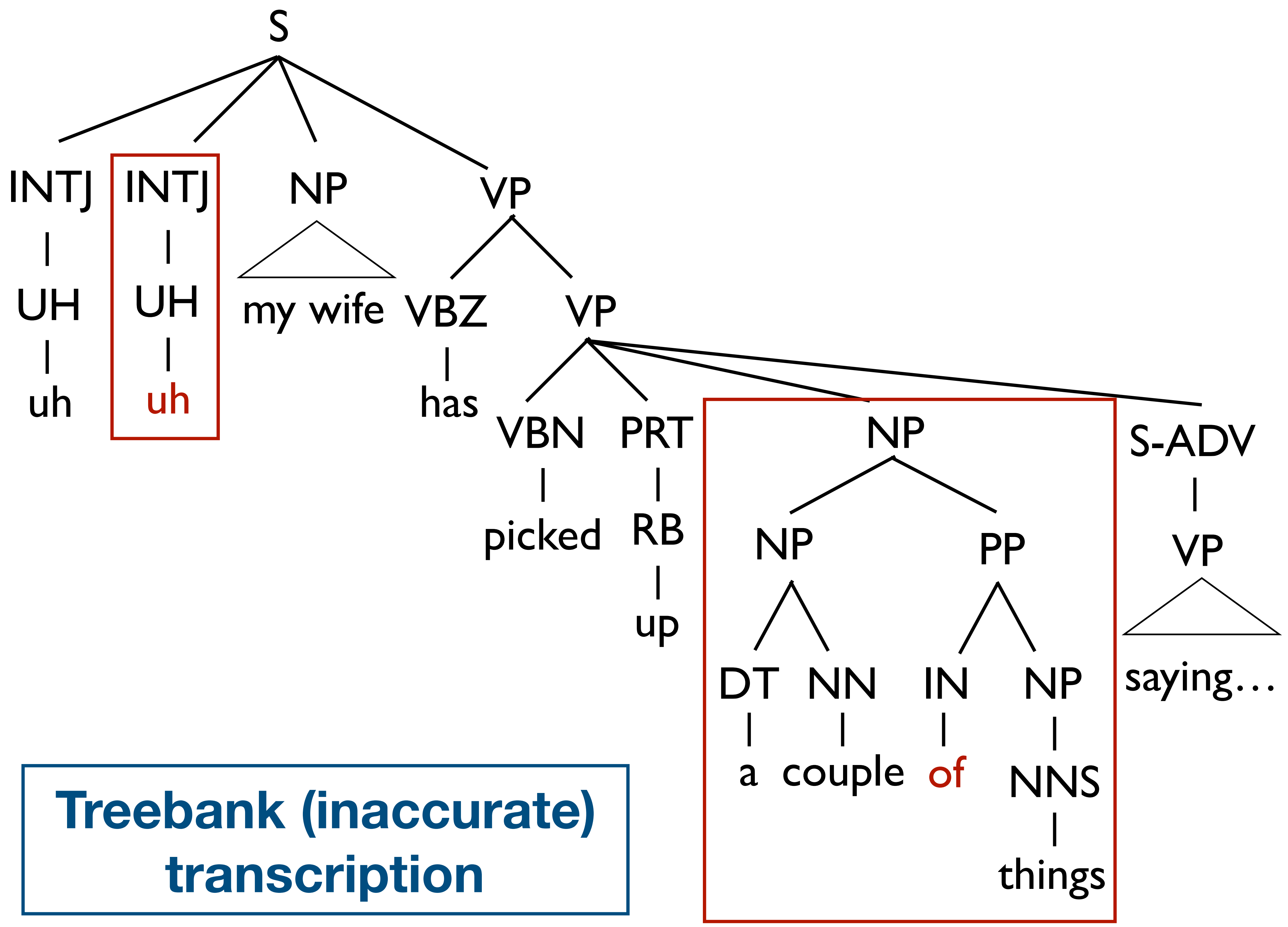
- Gains using prosody obscured by transcription errors
- Effect is statistically significant (p-value < 0.05)

Analysis: Transcription Error Effects

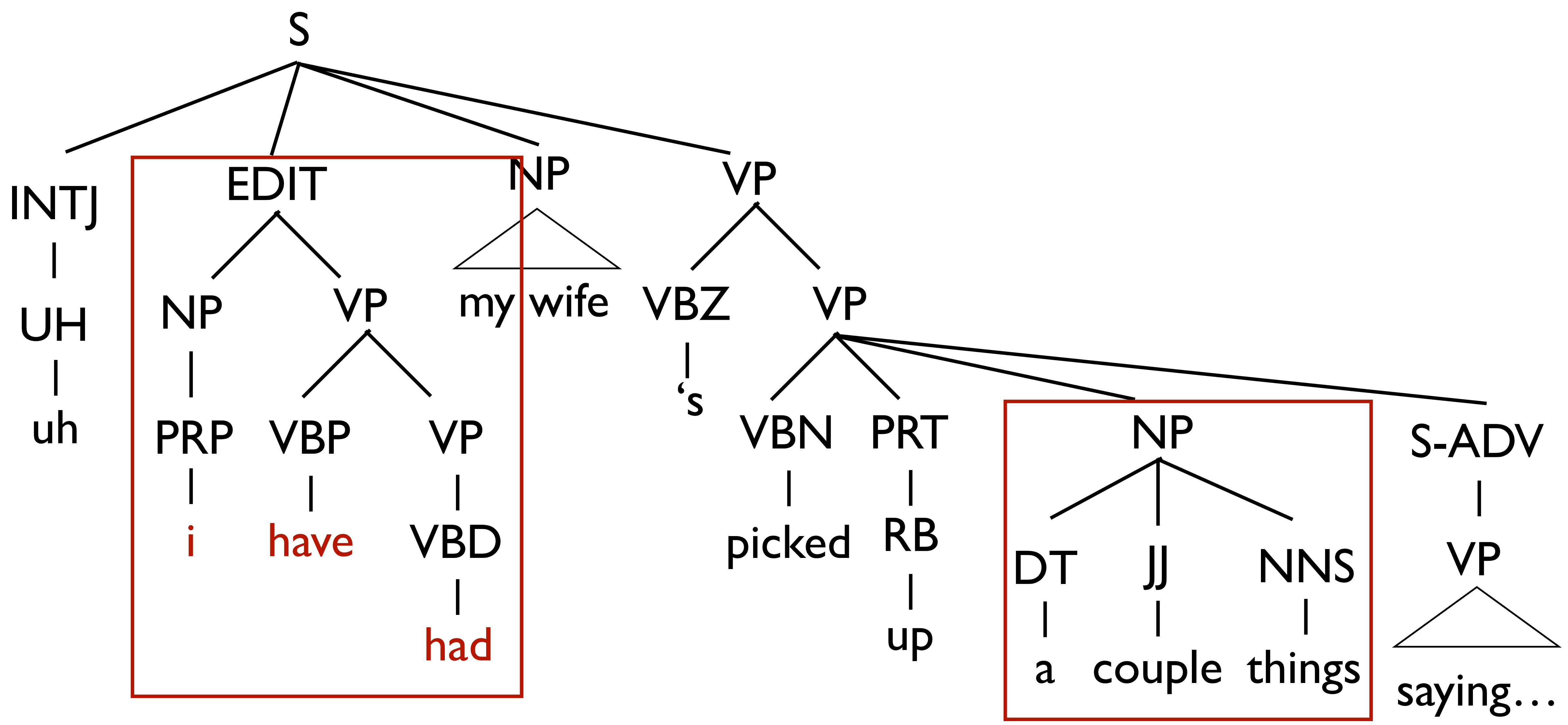
# Fluent Sentences	Prosody helped	Prosody "hurt"
with errors	57	82
no errors	270	269

→Parse structure changes

uh uh <i have had> my wife 's picked up a couple of things saying uh boy if we could refinish that 'd be a beautiful piece of furniture



Treebank (inaccurate) transcription



MS-State (accurate) transcription