Upcoming ERG 2020

Dan Flickinger

CSLI, Stanford University

DELPH-IN Summit 2020

13 July 2020

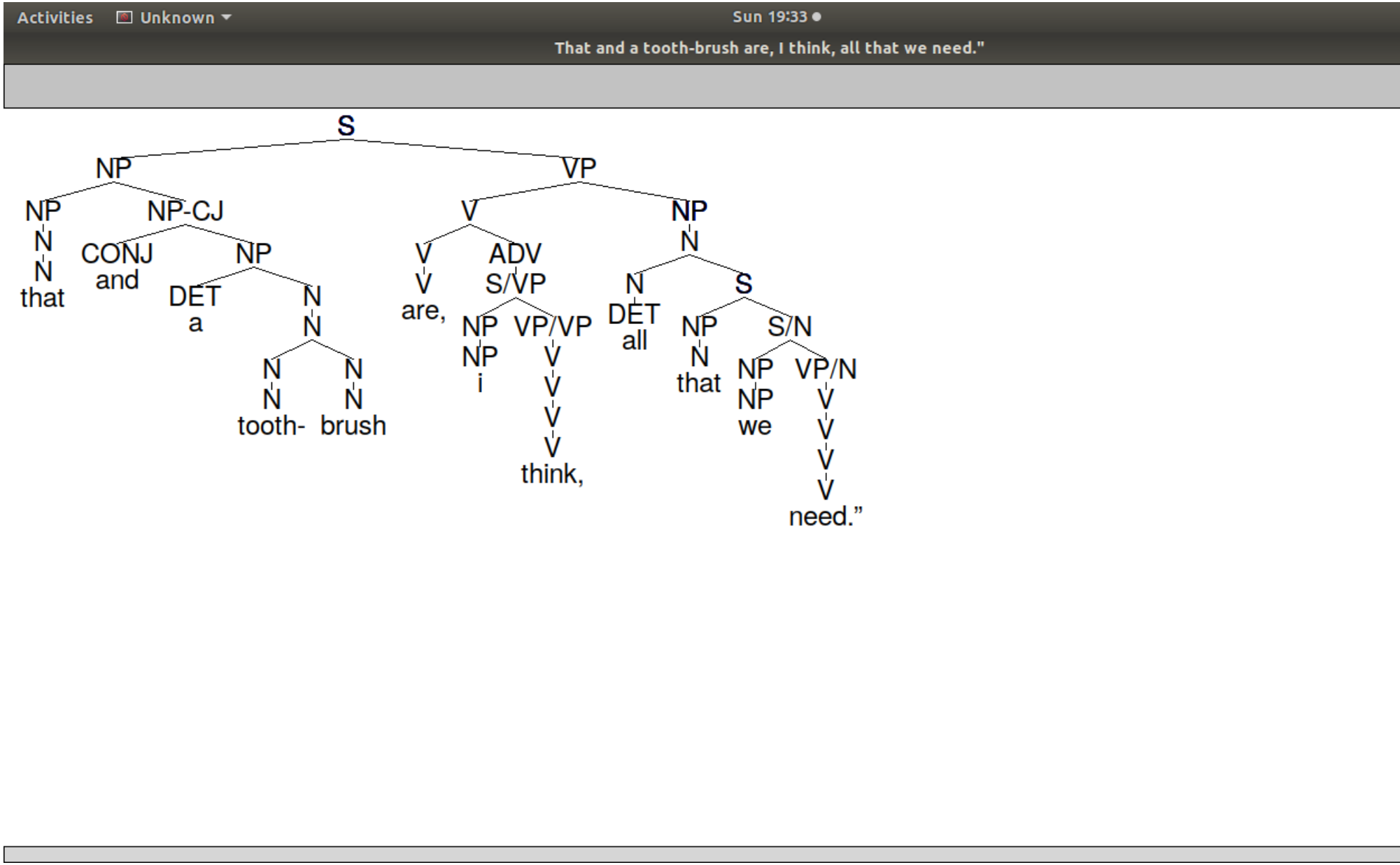
Overview of changes from 2018

- Punctuation marks now separate tokens
- Documentation strings now on all instances
- Completion of full Redwoods treebank with FFTB
- Separation of mal-rules from masking instances
- Expansion of mal-rules
- Better consistency in MRSs

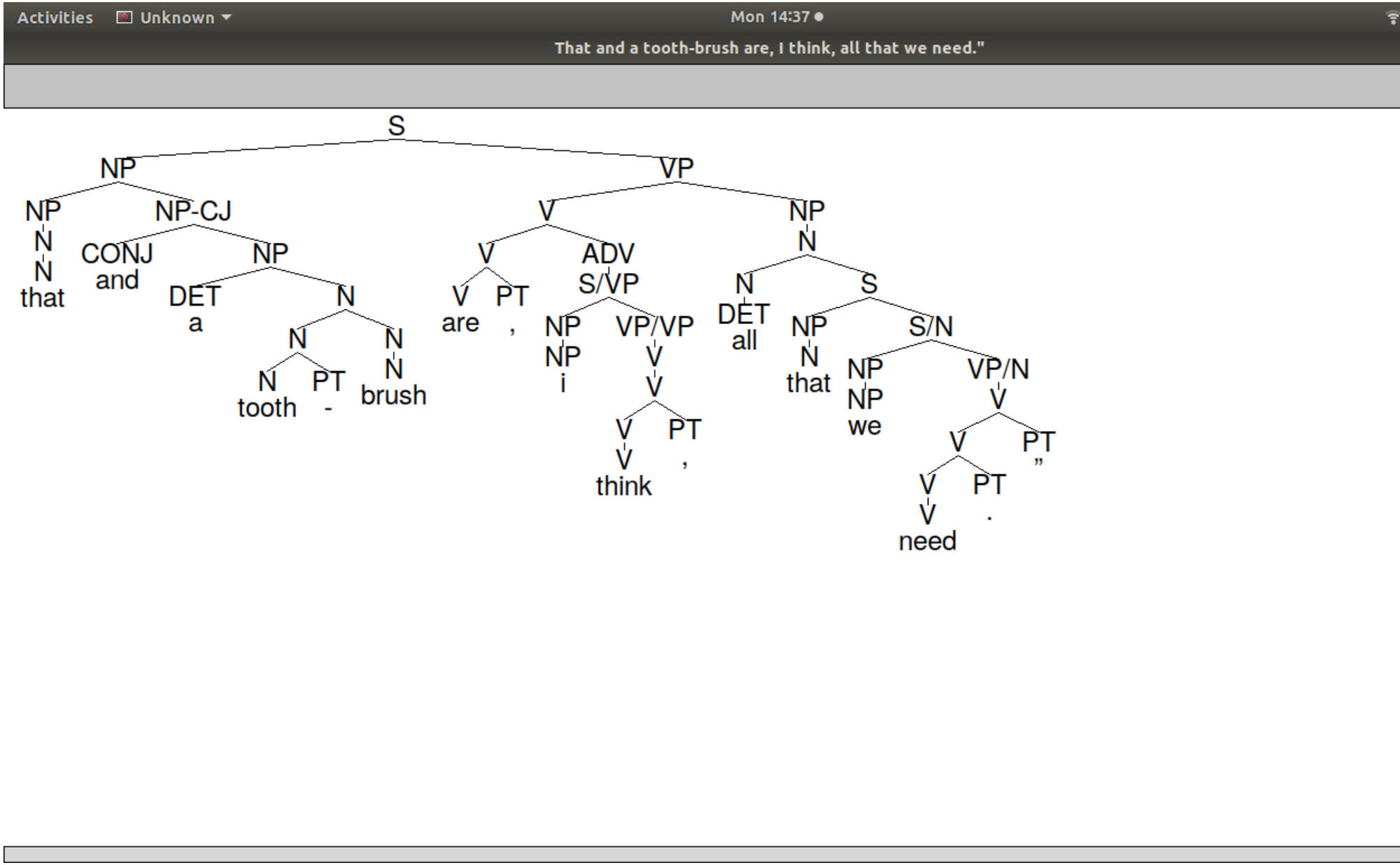
Punctuation marks now separate tokens

- In the past, treated punctuation marks as affixes
- Want better interoperability with standards in tokenization
- Now make (almost) all punctuation marks separate tokens
- Still attach low, but now with constructions, not lexical rules
- Thanks to Stephan and Woodley for assistance

Example of 2018 punctuation analysis



Example of 2020 punctuation analysis



Advantages of new tokenization

- In MRS EPs, from/to values exclude punctuation characters
- Fewer or no overlapping characterization issues
- All elements of hyphenated 'words' receive PoS tag
  snow-covered now tagged as noun followed by verb
- No longer need to ambiguate single apostrophes
- No longer need to have LKB mimic reattachment of punct marks
- (Thanks to Stephan for reminding us of this list)

Documentation strings

- Since 2018, moving to include doc strings within TDL definition
- Triple double quotes to begin and end doc string
- For 2018, used only for lexical leaf types
- Now generalized to all instances, including rules and roots
- Full support in ACE and LKB-FOS; patch for LKB; PET underway
- Thanks to Francis, Rebecca, Woodley, and John for assistance

Example of doc string

```
vp_sbrd-pre-lx_c := subconj_prdp_v_init_lex_rule &
"""
Predicative subordinate phrase from lexical VP: pre-head modifier

<ex>[Smiling,] Kim arose.
"""
  [ SYNSEM.PUNCT.RPUNCT comma_punct,
    RNAME scvl ].
```

Completion of Redwoods update with FFTB

- 2018 included only 5 of 22 sections of WSJ
- 2020 will include all 22 WSJ sections that were in 1214, now FFTB
- Should enable training of better model for parse selection

  An additional 30,000 sentences: 700,000 words
- Perhaps also include treebanked section 23 for testing

Separation of mal-rules from masking

- Recent releases have included variant for grammar-checking
- Additions to the standard grammar were of two types
  - Mal-rules and instances licensing ungrammatical signs
  - Masking instances that blocked or altered existing signs
- The masking was a patch for lack of informed parsing model
- Now clean separation of the two sets of changes
- Enables treebanking of student data using only mal-rules
- Allows customized parse selection model for grammar-checking

Expansion of mal-rules

- 2018: 200 error code groups
- 2020: 350 groups, aimed at Mandarin learners of English
- Most student data currently from single-sentence compositions
- Now beginning to gather short essays
- Also working with Singapore-based learners of English

Better consistency in MRS outputs

- Aiming at improved compliance with (revised) semantic algebra
- Making use of error diagnosis tools already available for treebank
- Fixing bugs as they are reported