LING/C SC 581:
Advanced Computational Linguistics

Lecture 17

Today's Topic

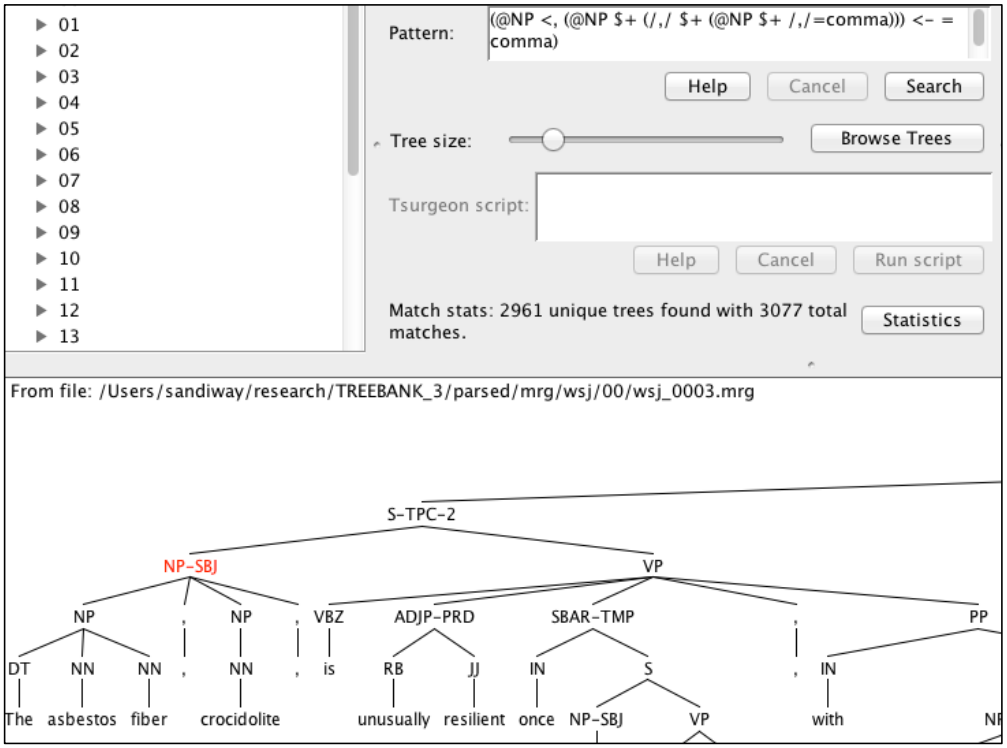
- tregex contd.
- Homework 8
  - due next Monday night, review next Tuesday

tregex

- Pattern:

```
(@NP <, (@NP $+ (/,/ $+ (@NP $+ /,/=comma))) <- =comma)
```

means: same node



Key:

`<,`	has first child
`$+`	immediate left sister of
`<-`	has last child

tregex

- Help

Recall regex grouping using parentheses:
e.g. (a+)(b+) defines groups 1 and 2

Variable Groups

If you write a node description using a regular expression, you can assign its matching groups to variable names. If more than one node has a group assigned to the same variable name, then matching will only occur when all such groups capture the same string. This is useful for enforcing coindexation constraints. The syntax is

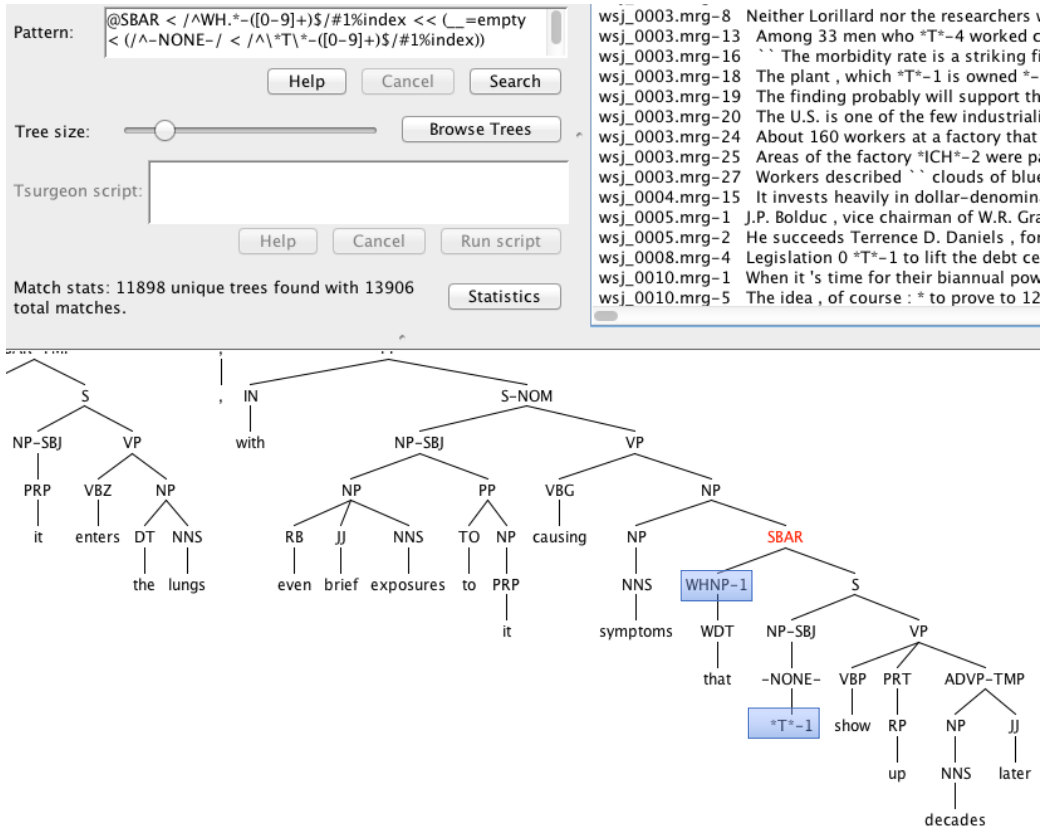
```
/ <regex-stuff> /#<group-number>%<variable-name>
```

For example, the pattern (designed for Penn Treebank trees)

```
@SBAR < /^WH.*-([0-9]+)$/#1%index << (__=empty < (/^-NONE-/ <
/^\*T\*-([0-9]+)$/#1%index))
```

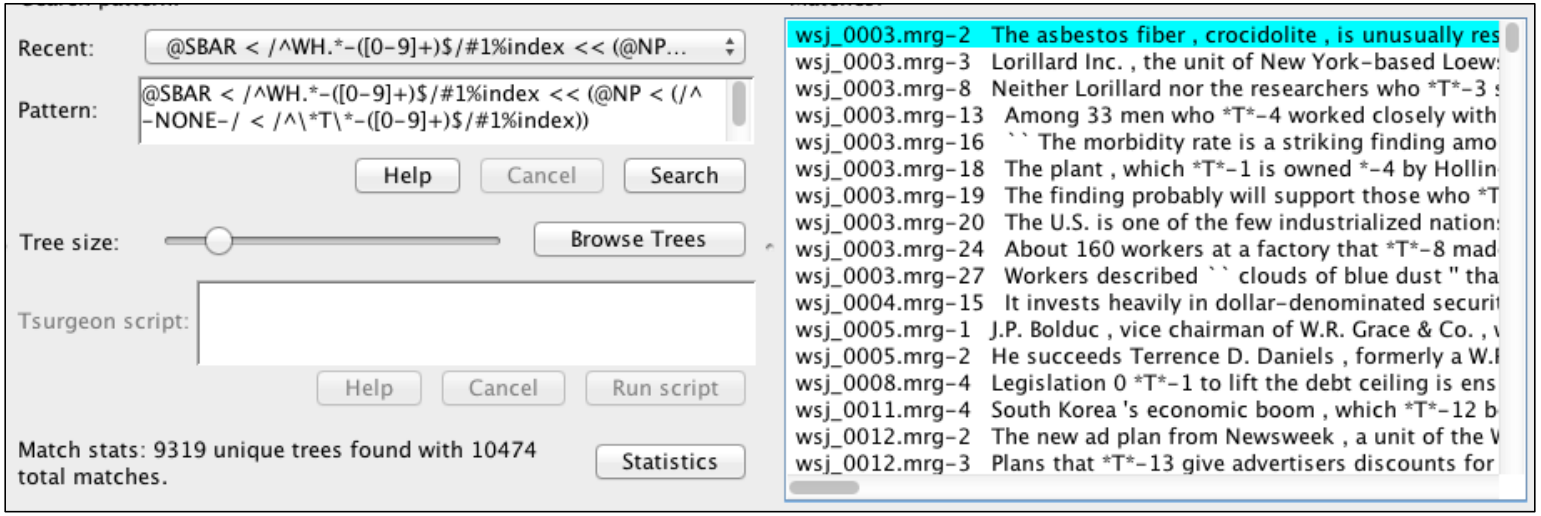
will match only such that the WH- node under the SBAR is coindexed with the trace node that gets the name `empty`.

tregex

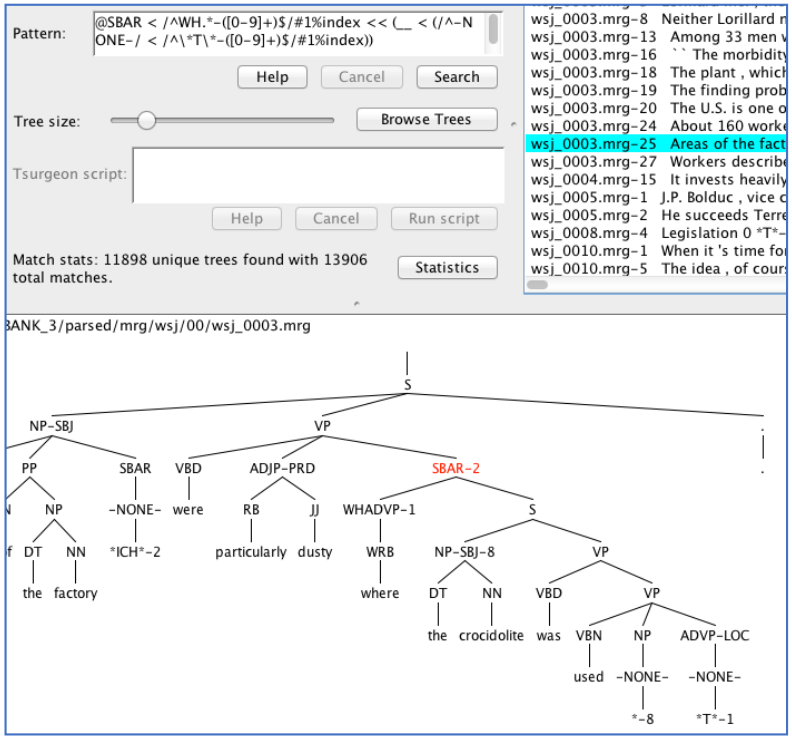

Pattern:
@SBAR < /^WH.*-([0-9]+)$/#1%index << (__=empty < (/^-NONE-/ < /^*T*-([0-9]+)$/#1%index))
Help
Cancel
Search
Tree size:
Browse Trees
Tsurgeon script:
Help
Cancel
Run script
Match stats: 11898 unique trees found with 13906 total matches.
Statistics
wsj_0003.mrg-8 Neither Lorillard nor the researchers
wsj_0003.mrg-13 Among 33 men who *T*-4 worked
wsj_0003.mrg-16 `` The morbidity rate is a striking
wsj_0003.mrg-18 The plant , which *T*-1 is owned *-
wsj_0003.mrg-19 The finding probably will support th
wsj_0003.mrg-20 The U.S. is one of the few industriali
wsj_0003.mrg-24 About 160 workers at a factory that
wsj_0003.mrg-25 Areas of the factory *ICH*-2 were
wsj_0003.mrg-27 Workers described `` clouds of blu
wsj_0004.mrg-15 It invests heavily in dollar-denomin
wsj_0005.mrg-1 J.P. Bolduc , vice chairman of W.R. Gra
wsj_0005.mrg-2 He succeeds Terrence D. Daniels , fo
wsj_0008.mrg-4 Legislation 0 *T*-1 to lift the debt ce
wsj_0010.mrg-1 When it 's time for their biannual pow
wsj_0010.mrg-5 The idea , of course : * to prove to 12
S
NP-SBJ
VP
PRP
it
VBZ
enters
NP
DT
the
NNS
lungs
IN
with
S-NOM
NP-SBJ
VP
NP
RB
even
JJ
brief
NNS
exposures
PP
TO
to
NP
PRP
it
VBG
causing
NP
NNS
symptoms
NP
SBAR
WHNP-1
WDT
that
S
NP-SBJ
-NONE-
T-1
VP
VBP
show
PRT
RP
up
ADVP-TMP
NP
NNS
decades
JJ
later

tregex

- Different results from:
  - `@SBAR < /^WH.*-([0-9]+)$/#1%index << (@NP < (/^-NONE-/ < /^\*T\*-([0-9]+)$/#1%index))`



tregex



Reason for difference

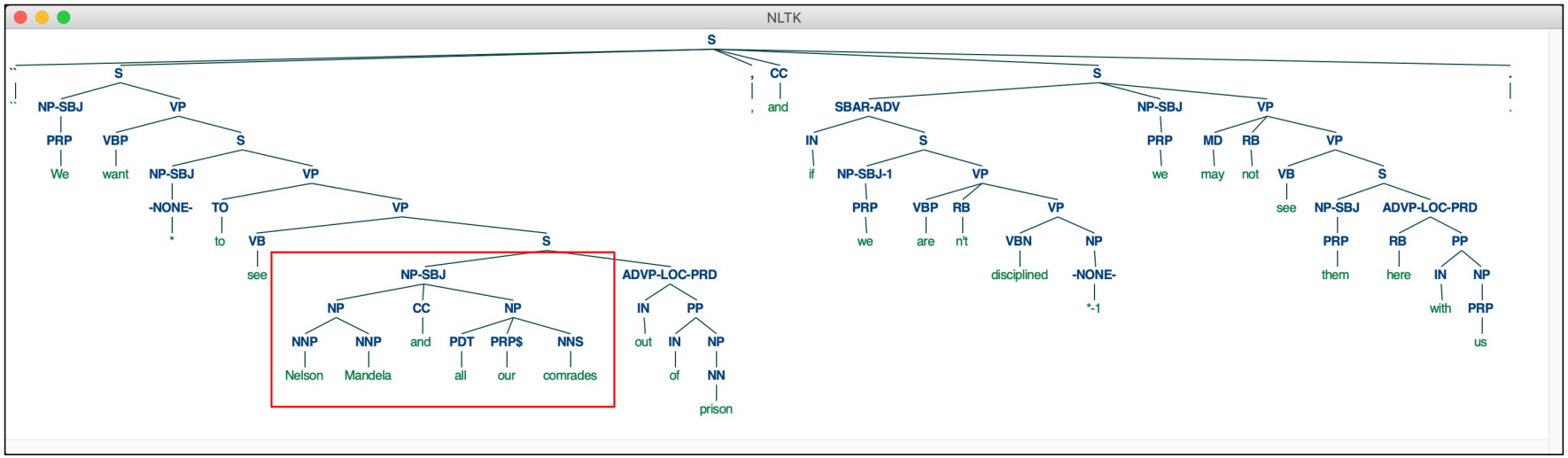
Example:

WHADVP also possible (not just WHNP)

Worked Example

From an earlier lecture, recall we can use the ptb package in Python …

```
from nltk.corpus import ptb
>>> ptb.parsed_sents(categories=['news'])[-1].draw()
```



Worked Example

- Stanza has a CoreNLP client:

```
>>> s = [w for w in ptb.sents(categories=['news'])[-1] if not w.startswith('*')]
>>> s2 = ' '.join(s)
>>> s2
"`` We want to see Nelson Mandela and all our comrades out of prison , and if we are n't disciplined we may not see them here with us ."
>>> with CoreNLPClient (annotators='tokenize,ssplit,pos,parse', output_format='text') as client:
...     ann = client.annotate(s2)
```

Worked Example

Constituency parse:

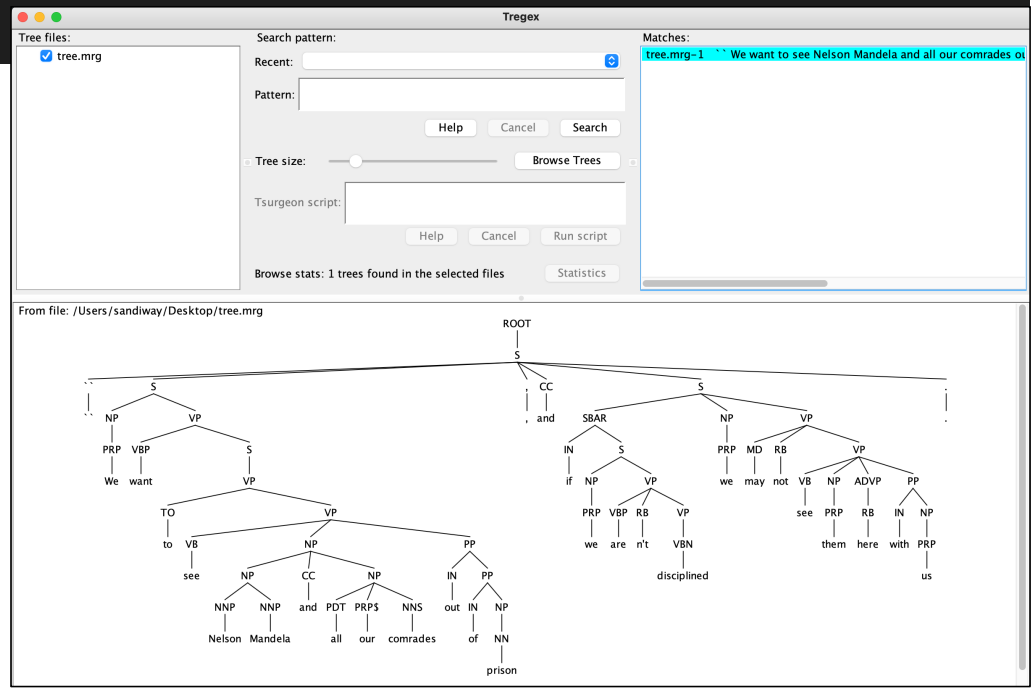
```
(ROOT
  (S (`` ``)
    (S
      (NP (PRP We))
      (VP (VBP want)
        (S
          (VP (TO to)
            (VP (VB see)
              (NP
                (NP (NNP Nelson) (NNP Mandela))
                (CC and)
                (NP (PDT all) (PRP$ our) (NNS comrades)))
              (PP (IN out)
                (PP (IN of)
                  (NP (NN prison)))))))))
    (, ,)
    (CC and)
    (S
      (SBAR (IN if)
        (S
          (NP (PRP we))
          (VP (VBP are) (RB n't)
            (VP (VBN disciplined)))))
      (NP (PRP we))
      (VP (MD may) (RB not)
        (VP (VB see)
          (NP (PRP them))
          (ADVP (RB here))
          (PP (IN with)
            (NP (PRP us))))))
    (. .)))
```

can put this parse in a text file, e.g. tree.mrg

Worked Example

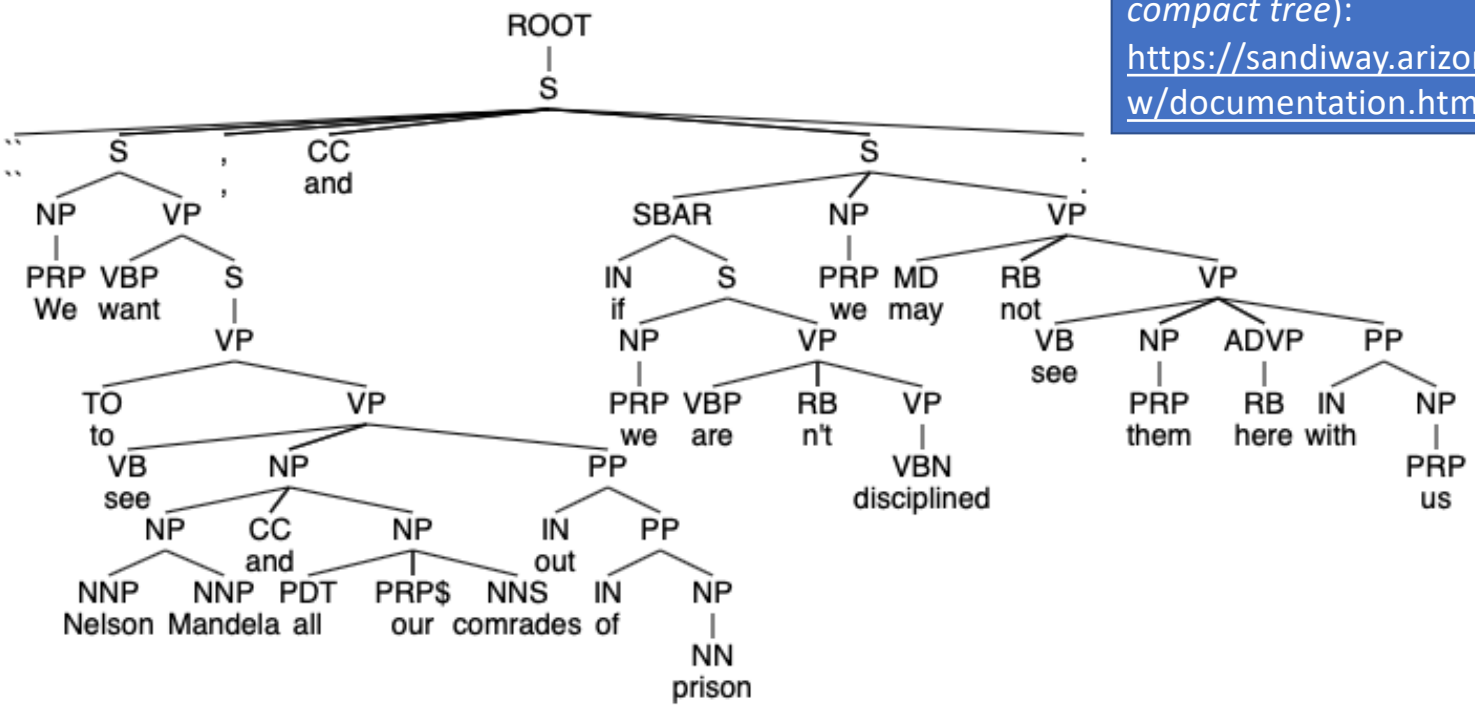
```
(base) ~$ cd courses/581/ling581-22/stanford-tregex-2020-11-17/
(base) stanford-tregex-2020-11-17$ ./run-tregex-gui.command
```

load tree.mrg into tregex

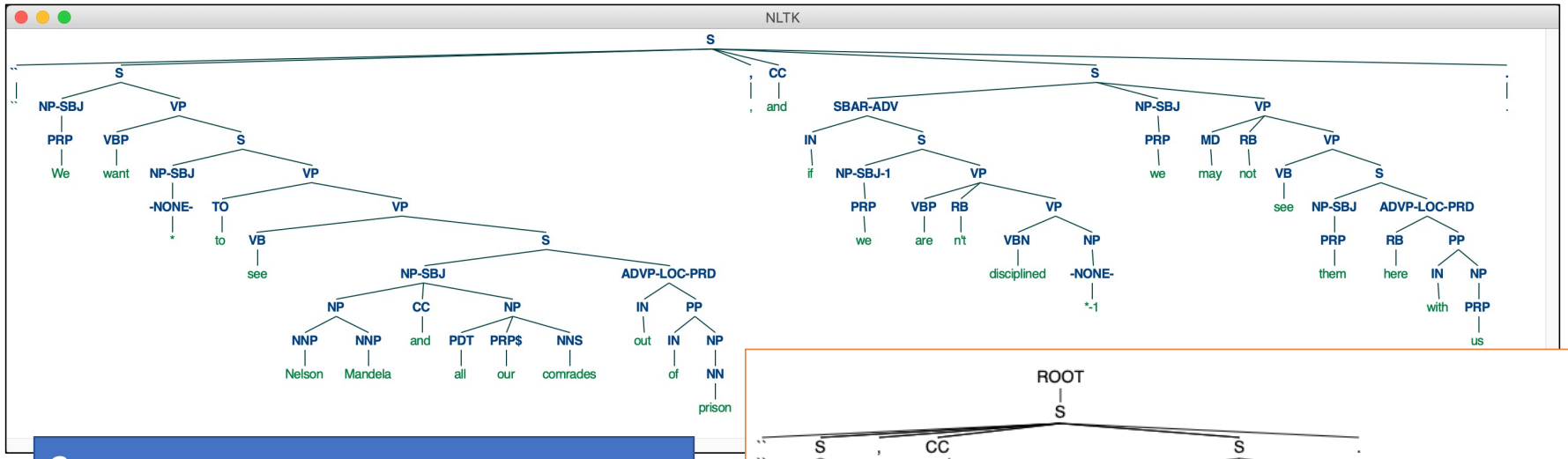


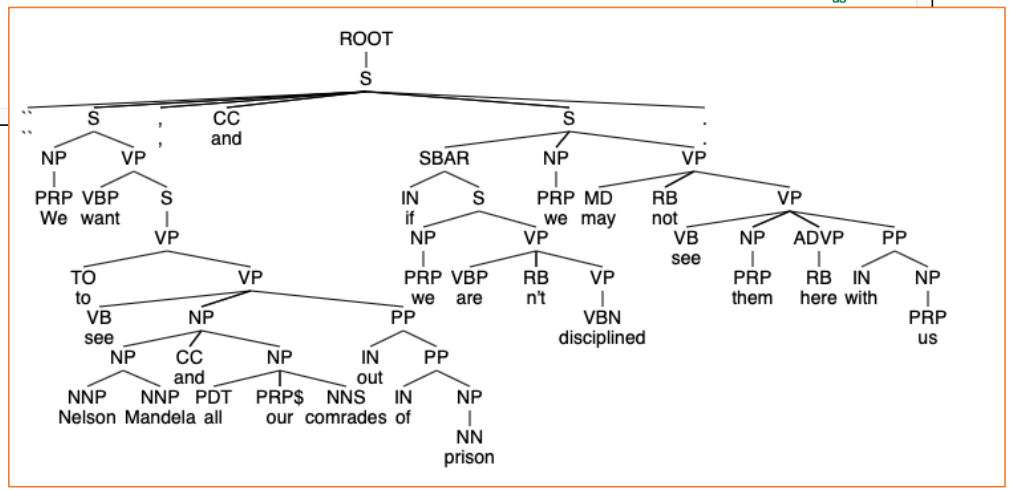
Worked Example

Or use my software (*for a more compact tree*): https://sandiway.arizona.edu/treedraw/documentation.html



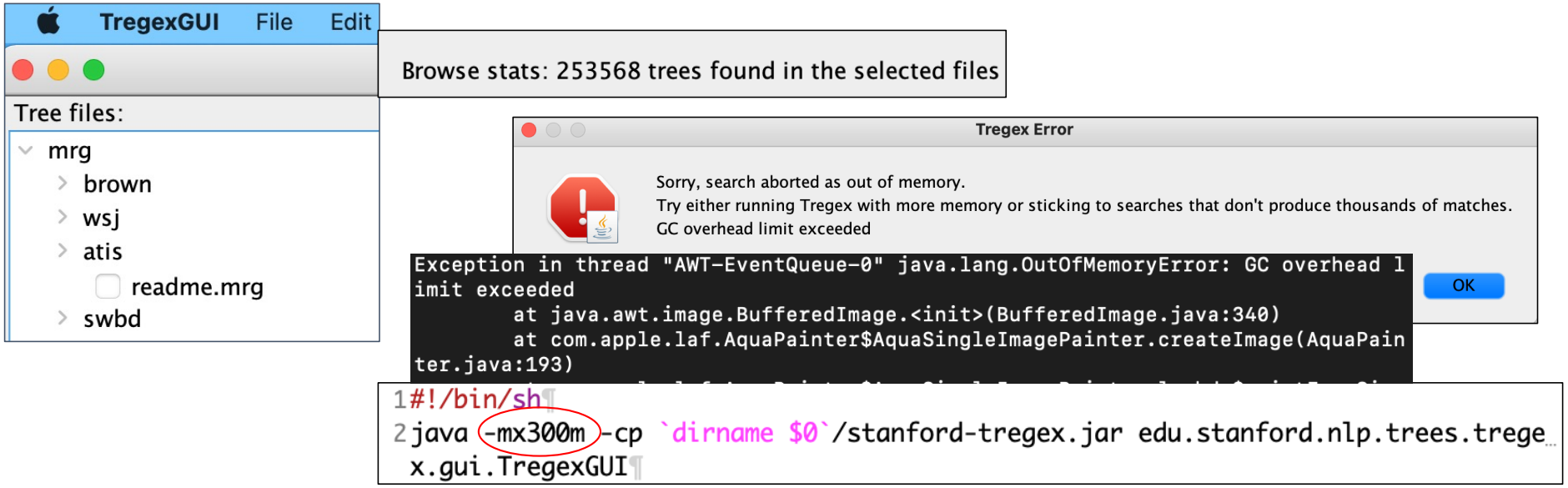
Worked Example





Compare:

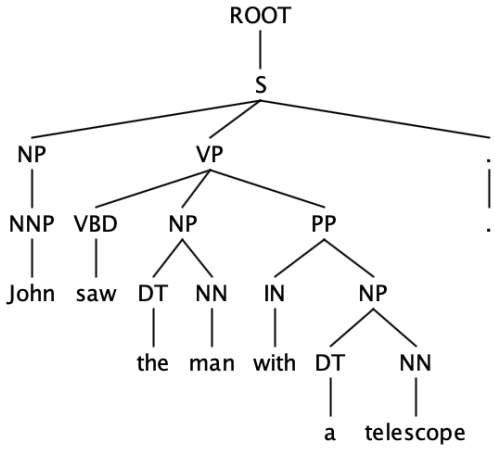
- placement of punctuation
- empty categories (-NONE-)
- syntactic sub-labels (-SBJ/LOC/PRD/-1)
  - "functional"

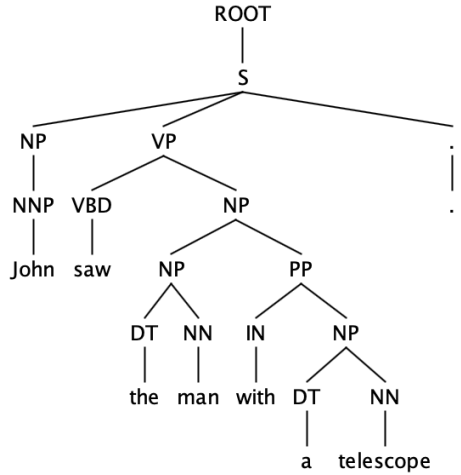

TregexGUI
File
Edit
Tree files:
mrg
brown
wsj
atis
readme.mrg
swbd
Browse stats: 253568 trees found in the selected files
Tregex Error
Sorry, search aborted as out of memory.
Try either running Tregex with more memory or sticking to searches that don't produce thousands of matches.
GC overhead limit exceeded
OK
Exception in thread "AWT-EventQueue-0" java.lang.OutOfMemoryError: GC overhead limit exceeded
at java.awt.image.BufferedImage.<init>(BufferedImage.java:340)
at com.apple.laf.AquaPainter$AquaSingleImagePainter.createImage(AquaPainter.java:193)
#!/bin/sh
java -mx300m -cp `dirname $0`/stanford-tregex.jar edu.stanford.nlp.trees.tregex.gui.TregexGUI

Homework 8: Question 1

Recall PP attachment ambiguity to NP vs. VP (Homework 7)?

- Check the TREEBANK_3 corpus
- Are there more cases of PP attachment to NPs or VPs?
- **CAUTION**: be aware of syntactic sub-labels
- Show your search and statistics





Homework 8: Question 2

- Find matches for PPs headed by preposition *with* immediately dominated by NPs or VPs.
- Which is more frequent?
- **CAUTION**: be aware of syntactic sub-labels
- Show your search and statistics
- Are the statistics compatible with the rank of the parses returned by the standalone Stanford parser?
- Look at both Q2 and Q3 from Homework 7.

Homework 8: Question 3

- Let's investigate prepositional stranding in English in TREEBANK_3.
- Some examples:
  - Which city did you come **from**?
  - How long do you think you will be gone **for**?
  - That chair was sat **in** by John
  - Here is the place I told you **about**
- Devise a general tregex expression to search for trees of the form =>
- Show your search and statistics
- **Note**: PP object won't always be an NP
- **Hint**: ___ (represents any node, see Help button)