LING/C SC 581:
Advanced Computational Linguistics

Lecture 23

Today's Topics

- tregex: searching, part 3
- Homework 10

Last Time

- Changed the memory size (`-mx`) to accommodate larger corpora (brown + wsj) and our searching (without hanging). Inside `run-tregex-gui.command` we made a change:
  - `java -mx300m`
  - `java -mx900m` (*wasn't big enough*)
  - `java -mx2000m` (*searches don't hang*)

tregex

Naming nodes

Nodes can be given names (a.k.a. handles) using '='. A named node will be stored in a map that maps names to nodes so that if a match is found, the node corresponding to the named node can be extracted from the map. For example `(NP < NNP=name)` will match an NP dominating an NNP and after a match is found, the map can be queried with the name to retreived the matched node using `TregexMatcher#getNode(Object o)` with (String) argument "name" (not "=name"). Note that you are not allowed to name a node that is under the scope of a negation operator (the semantics would be unclear, since you can't store a node that never gets matched to). Trying to do so will cause a `ParseException` to be thrown. Named nodes *can* be put within the scope of an optionality operator.

Named nodes that refer back to previous named nodes need not have a node description -- this is known as "backreferencing". In this case, the expression will match only when all instances of the same name get matched to the same tree node. For example: the pattern

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

matches only an NP dominating exactly the sequence `NP , NP ,` -- the mother NP cannot have any other daughters. Multiple backreferences are allowed. If the node w/ no node description does not refer to a previously named node, there will be no error, the expression simply will not match anything.

Another way to refer to previously named nodes is with the "link" symbol: '~'. A link is like a backreference, except that instead of having to be *equal to* the referred node, the current node only has to match the label of the referred to node. A link cannot have a node description, i.e. the '~' symbol must immediately follow a relation symbol.

Key:

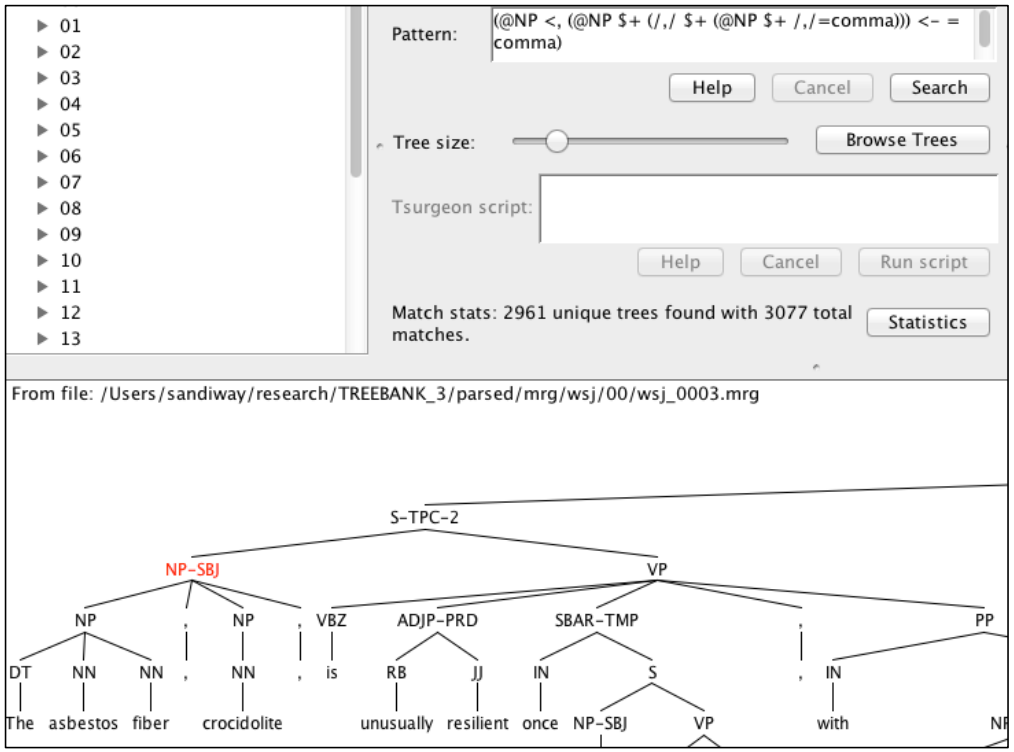
- `x <, y`
  1st child y
- `x <– y`
  last child y
- `x $+ y`
  x immediate left sister of y

tregex

- Pattern:

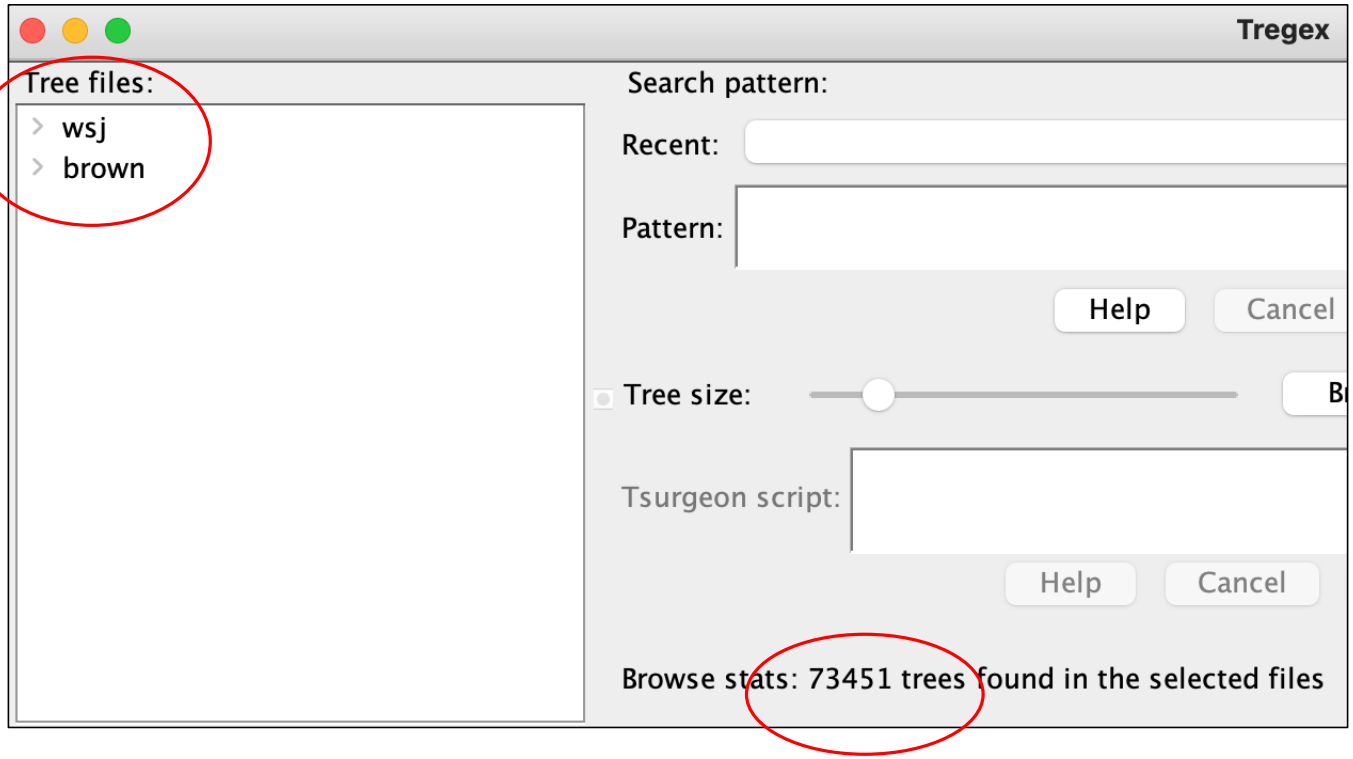
```
@NP <, (@NP $+ (/,/ $+ (@NP $+ /,/=comma))) <- =comma
```

must be same node

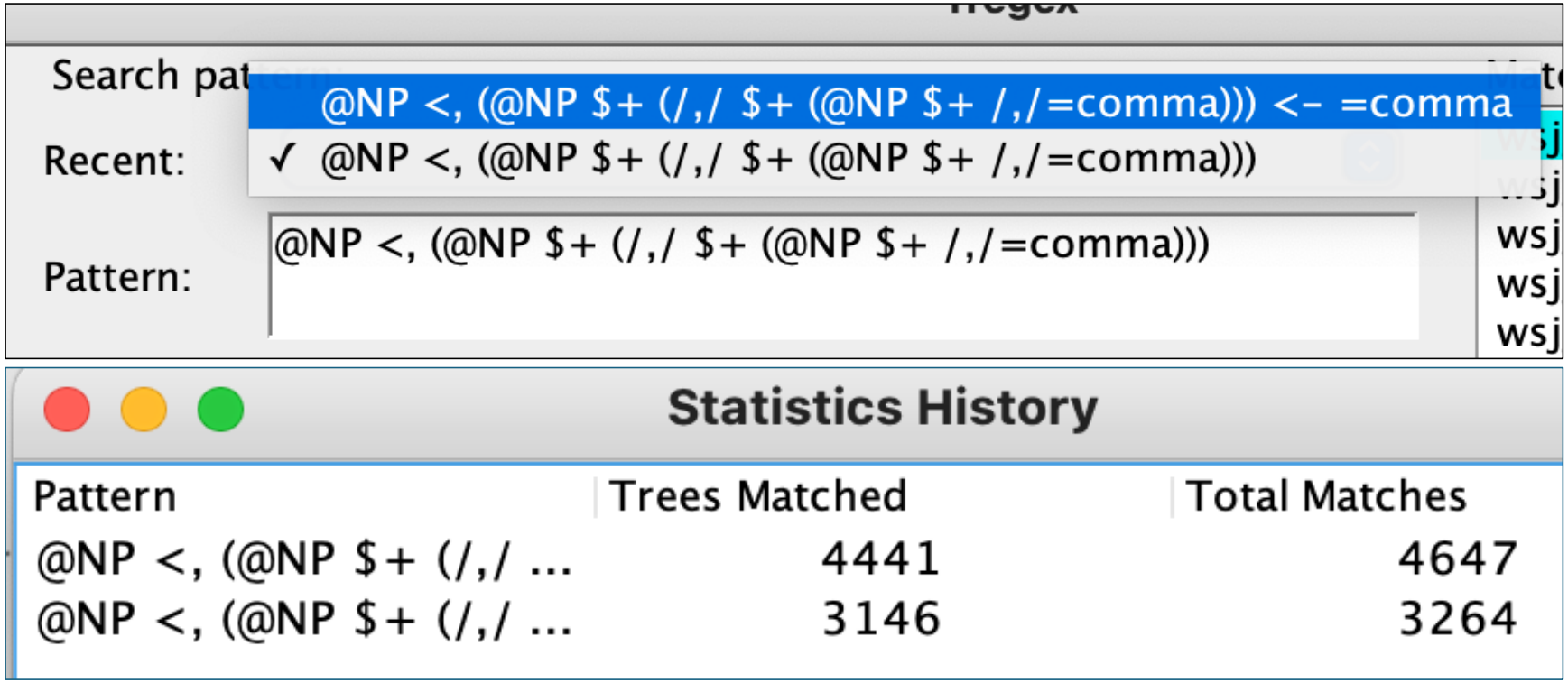


Key:
<, first child
$+ immediate left sister
<- last child

tregex


Tregex
Tree files:
wsj
brown
Search pattern:
Recent:
Pattern:
Help
Cancel
Tree size:
Tsurgeon script:
Help
Cancel
Browse stats: 73451 trees found in the selected files

tregex


Search pat
@NP <, (@NP $+ (/,/ $+ (@NP $+ /,/=comma))) <- =comma
Recent:
✓ @NP <, (@NP $+ (/,/ $+ (@NP $+ /,/=comma)))
Pattern:
@NP <, (@NP $+ (/,/ $+ (@NP $+ /,/=comma)))
wsj
wsj
wsj
Statistics History
Pattern
Trees Matched
Total Matches
@NP <, (@NP $+ (/,/ ...
4441
4647
@NP <, (@NP $+ (/,/ ...
3146
3264

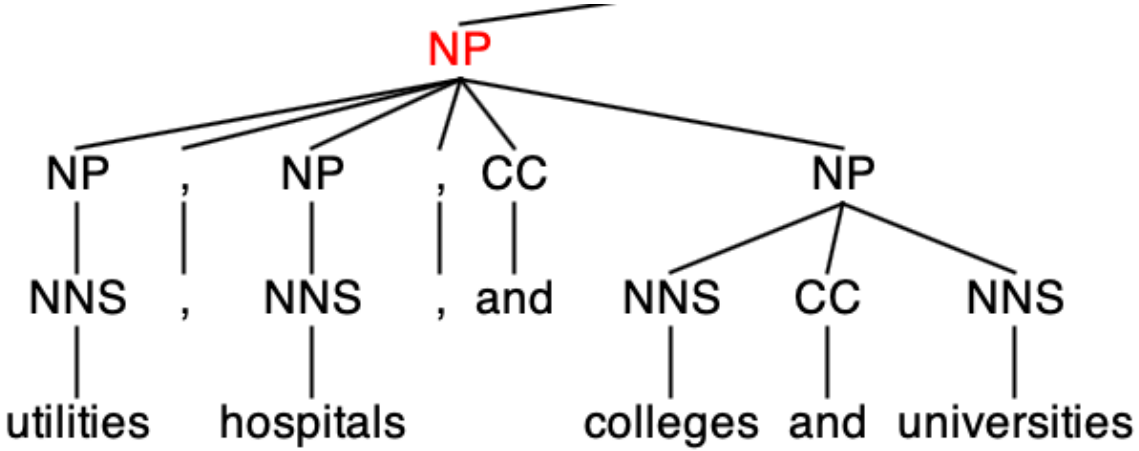
tregex

- Save search sentences to files `short` and `long` (form of query). Run diff.

```
(base) ~$ cd Desktop
(base) Desktop$ diff short long
2d1
< wsj_0353.mrg-2   The Financial Accounting Foundation voted 12-2 that FASB accounting rules supercede GASB rules in regard to utilities , hospitals , and colleges and universities owned * by the government .
11d9
< wsj_0341.mrg-3   Unocal said 0 the venture would enable it to recover more of its refining and marketing investment and prepare for expected growth in exploration , production , chemicals and other areas .
18d15
< wsj_0354.mrg-9   Uniroyal has 2,600 employees and facilities in the U.S. , Canada , Brazil , Italy and Taiwan .
```

tregex

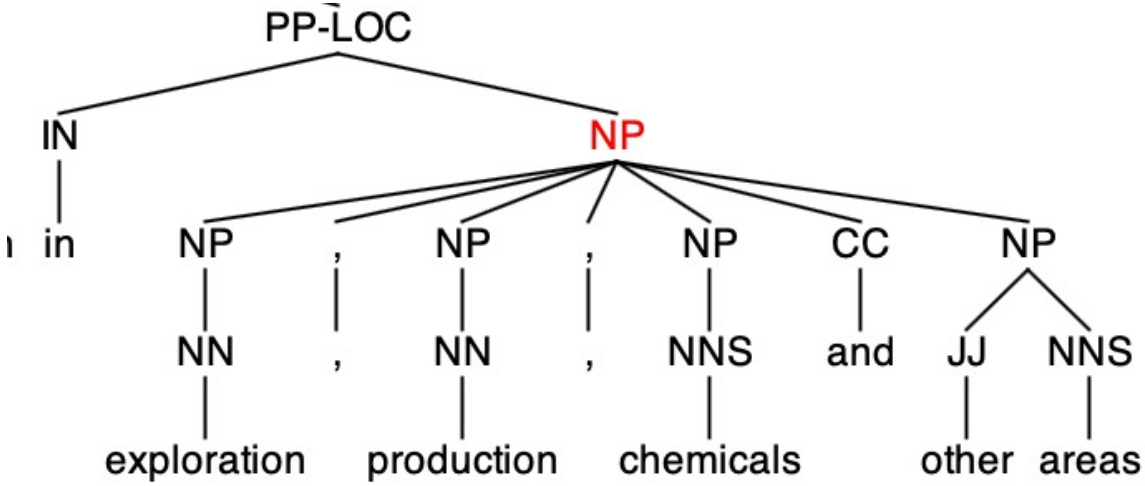
```
2d1
< wsj_0353.mrg-
2   The Financial
Accounting Foundation
voted 12-2 that FASB
accounting rules
supercede GASB rules
in regard to
utilities , hospitals
, and colleges and
universities owned *
by the government .
```



tregex

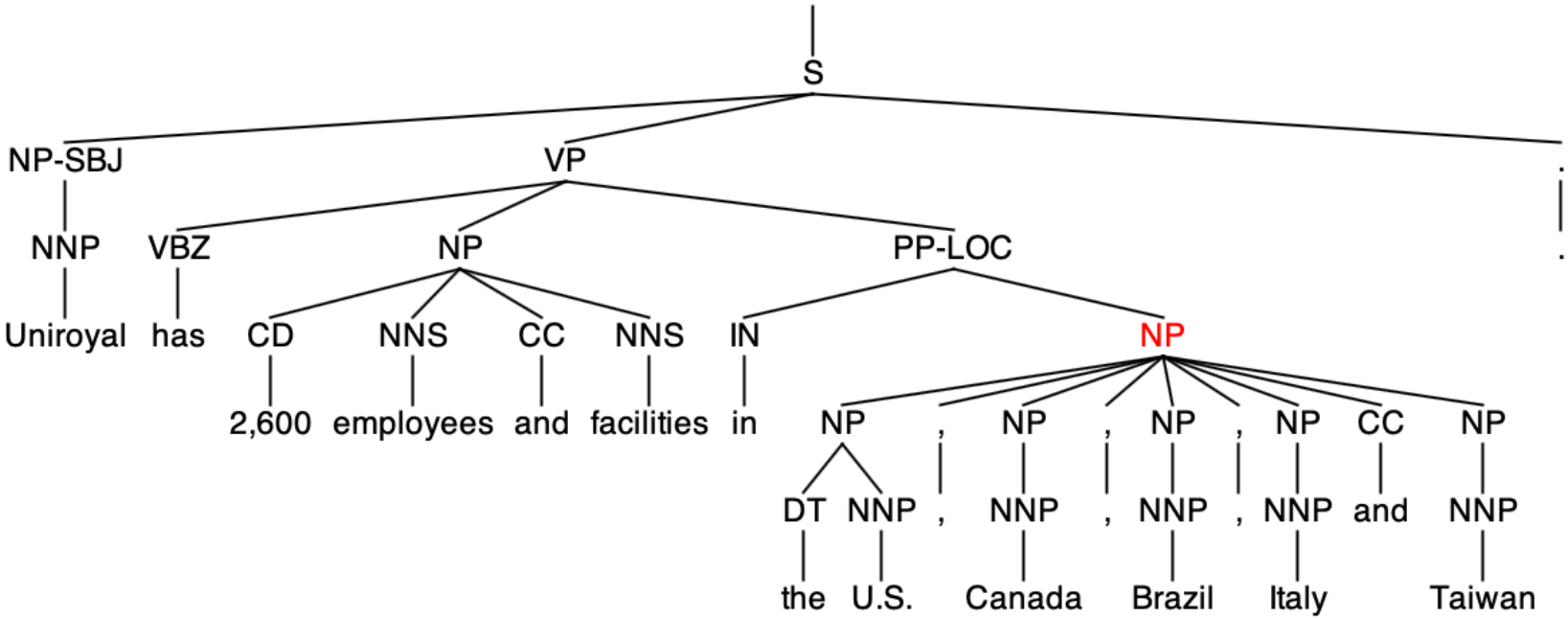
11d9

```
< wsj_0341.mrg-
3   Unocal said 0 the
venture would enable
it to recover more of
its refining and
marketing investment
and prepare for
expected growth in
exploration ,
production ,
chemicals and other
areas .
```



tregex

```
18d15
< wsj_0354.mrg-9   Uniroyal has 2,600 employees and facilities in the U.S. , Canada , Brazil , Italy and Taiwan .
```



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

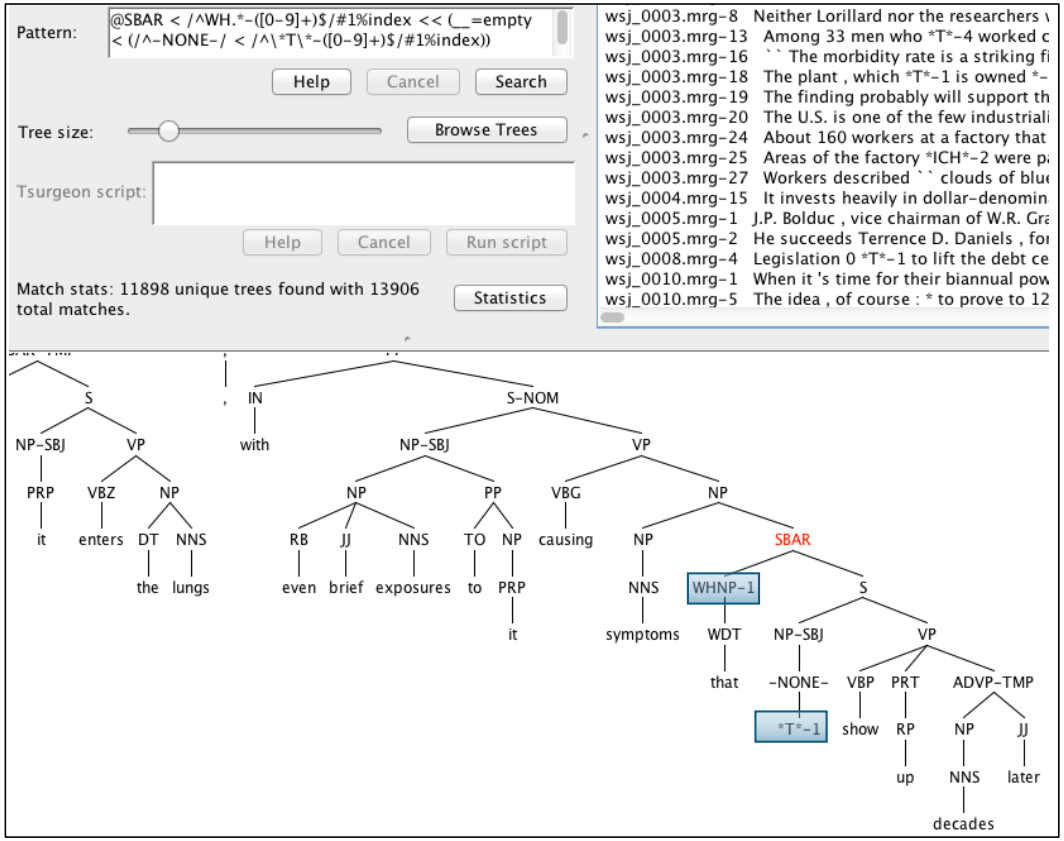
```
/ <regex-stuff> /#<group-number>%<variable-name>
```

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

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

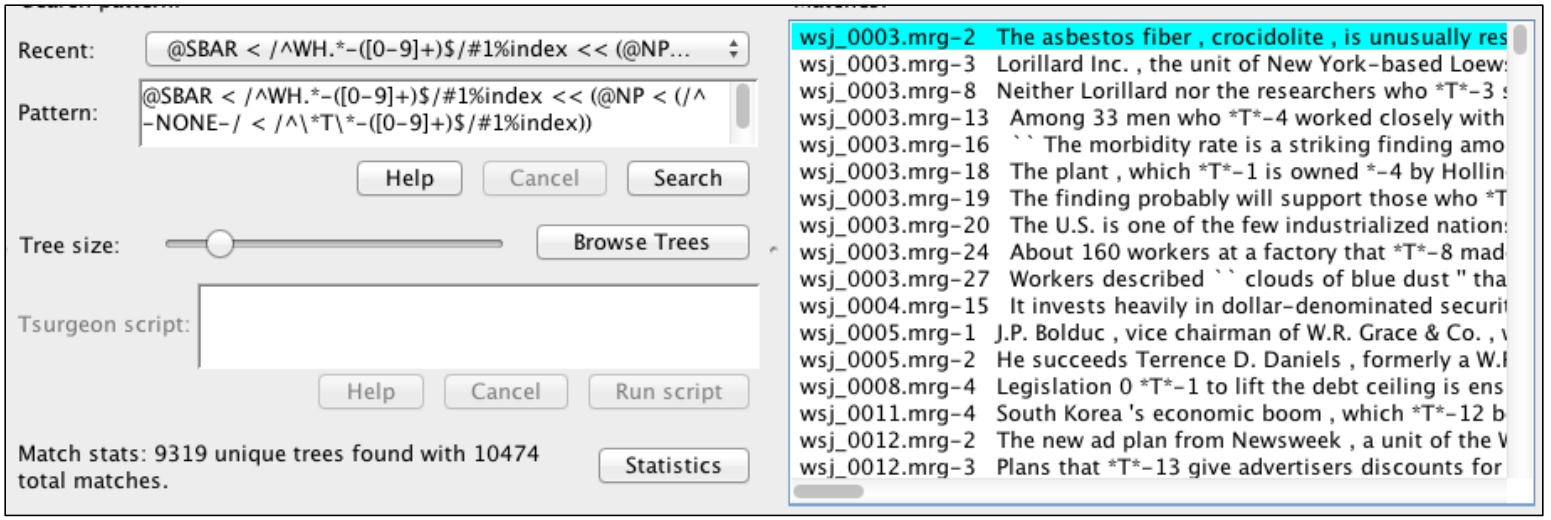
will match only such that the WH- node under the SBAR is coindexed with the trace node that gets the name `empty`.

tregex

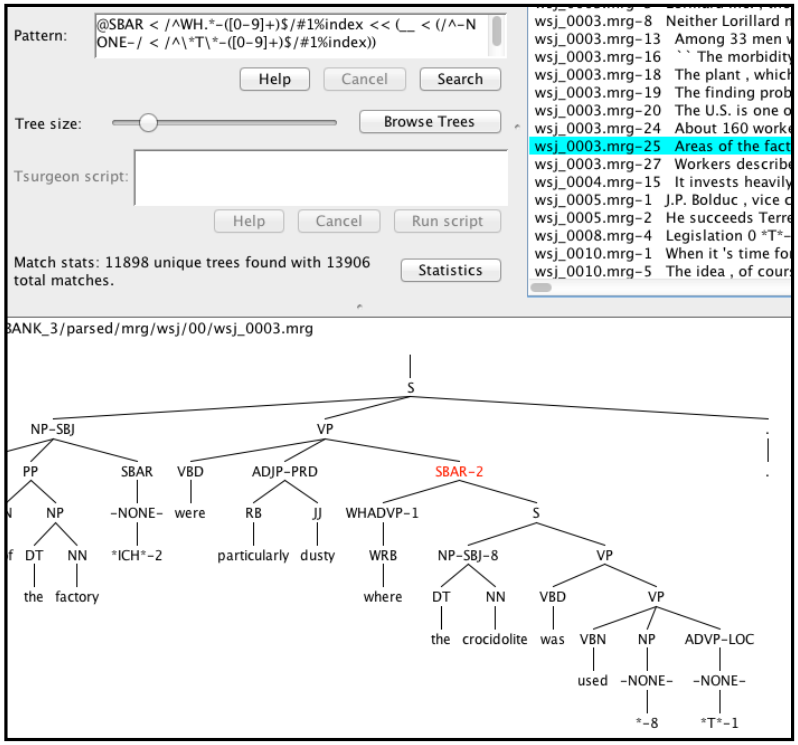

Pattern:
@SBAR < /^WH.*-([0-9]+)$/#1%index << (__=empty < (/^-NONE-/ < /^*T*-([0-9]+)$/#1%index))
Help
Cancel
Search
Tree size:
Browse Trees
Tsurgeon script:
Help
Cancel
Run script
Match stats: 11898 unique trees found with 13906 total matches.
Statistics
wsj_0003.mrg-8 Neither Lorillard nor the researchers
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 p
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
VBZ
NP
it
enters
DT
NNS
the
lungs
,
IN
with
S-NOM
NP-SBJ
VP
NP
PP
VBG
NP
RB
JJ
NNS
TO
NP
causing
NP
SBAR
even
brief
exposures
to
PRP
it
NNS
WHNP-1
S
symptoms
WDT
NP-SBJ
VP
that
-NONE-
VBP
PRT
ADVP-TMP
T-1
show
RP
NP
JJ
up
NNS
later
decades

tregex

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



tregex



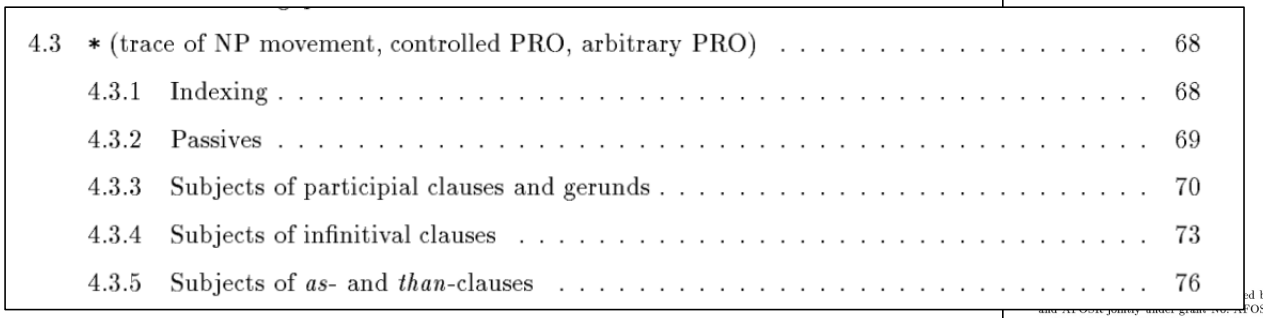
Reason for difference

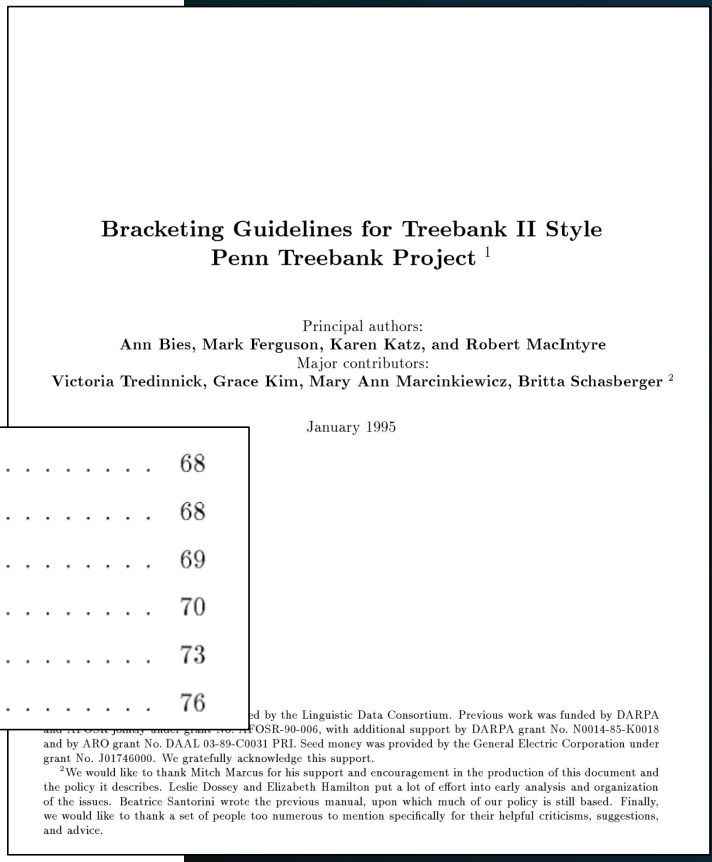
Example:

WHADVP also possible (not just WHNP)

Homework 10

- *Bracketing Guidelines for Treebank*
  - TREEBANK_3 > docs > prsguid1.pdf





Bracketing Guidelines for Treebank II Style
Penn Treebank Project [1]

Principal authors:
Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre
Major contributors:
Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, Britta Schasberger [2]

January 1995

ed by the Linguistic Data Consortium. Previous work was funded by DARPA and AFOSR jointly under grant No. AFOSR-90-006, with additional support by DARPA grant No. N0014-85-K0018 and by ARO grant No. DAAL 03-89-C0031 PRI. Seed money was provided by the General Electric Corporation under grant No. J01746000. We gratefully acknowledge this support.

[2]We would like to thank Mitch Marcus for his support and encouragement in the production of this document and the policy it describes. Leslie Dossey and Elizabeth Hamilton put a lot of effort into early analysis and organization of the issues. Beatrice Santorini wrote the previous manual, upon which much of our policy is still based. Finally, we would like to thank a set of people too numerous to mention specifically for their helpful criticisms, suggestions, and advice.

Homework 10

4.3.2 Passives

Object of verb. The trace (NP *) is put after the passive verb and coindexed with the constituent in subject position.

```
(S (NP-SBJ-1 John)
   (VP was
       (VP hit
           (NP *-1)
           (PP by
               (NP-LGS a ball)))))
```

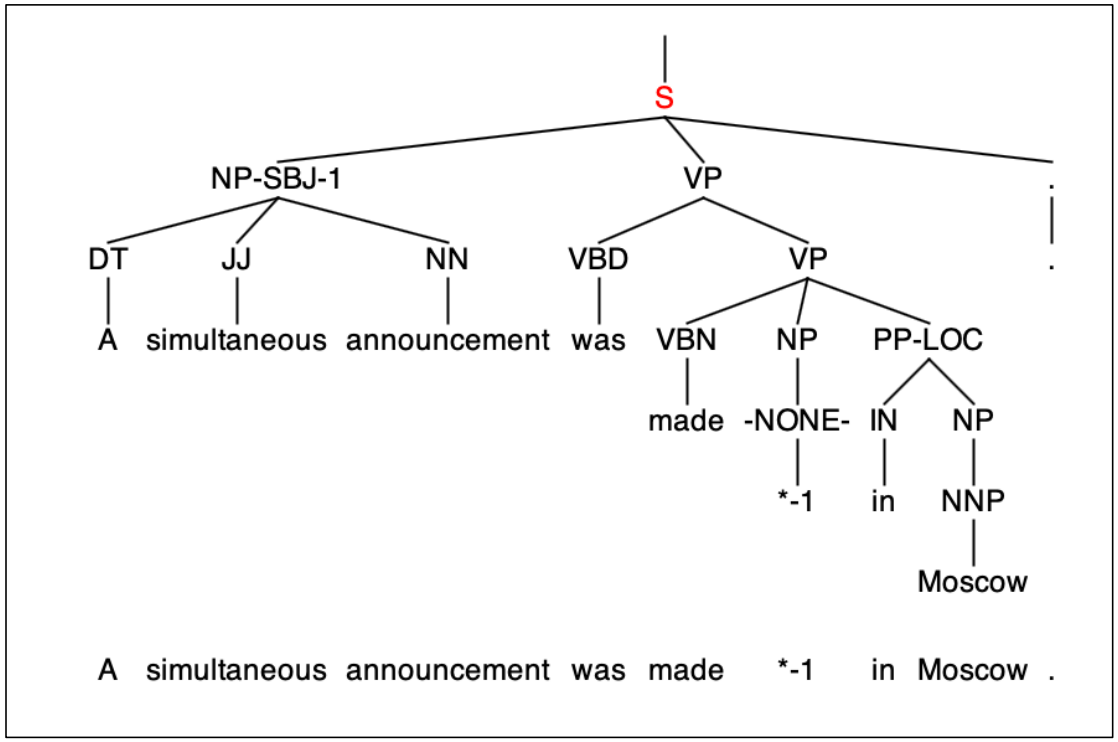
Note that the * may come before or after a PRT (particle). There is no policy governing this and either order is possible, though it is somewhat more likely for the PRT to come second:

```
(S (NP-SBJ-1 Arthur)
   (VP was
       (VP picked
           (NP *-1)
           (PRT up)
           (PP by
               (NP-LGS aliens)))))

(S (NP-SBJ-1 Arthur)
   (VP was
       (VP picked
           (PRT up)
           (NP *-1)
           (PP by
               (NP-LGS aliens)))))
```

- This is the PRD (not MRG) form:
  - no POS tags, just syntax labels
  - MRG = PRD + POS
- Past participle form of the verb has POS tag:
  - VBN
- Empty categories have POS tag:
  - -NONE-
- Let's write a search for passives!

Homework 10

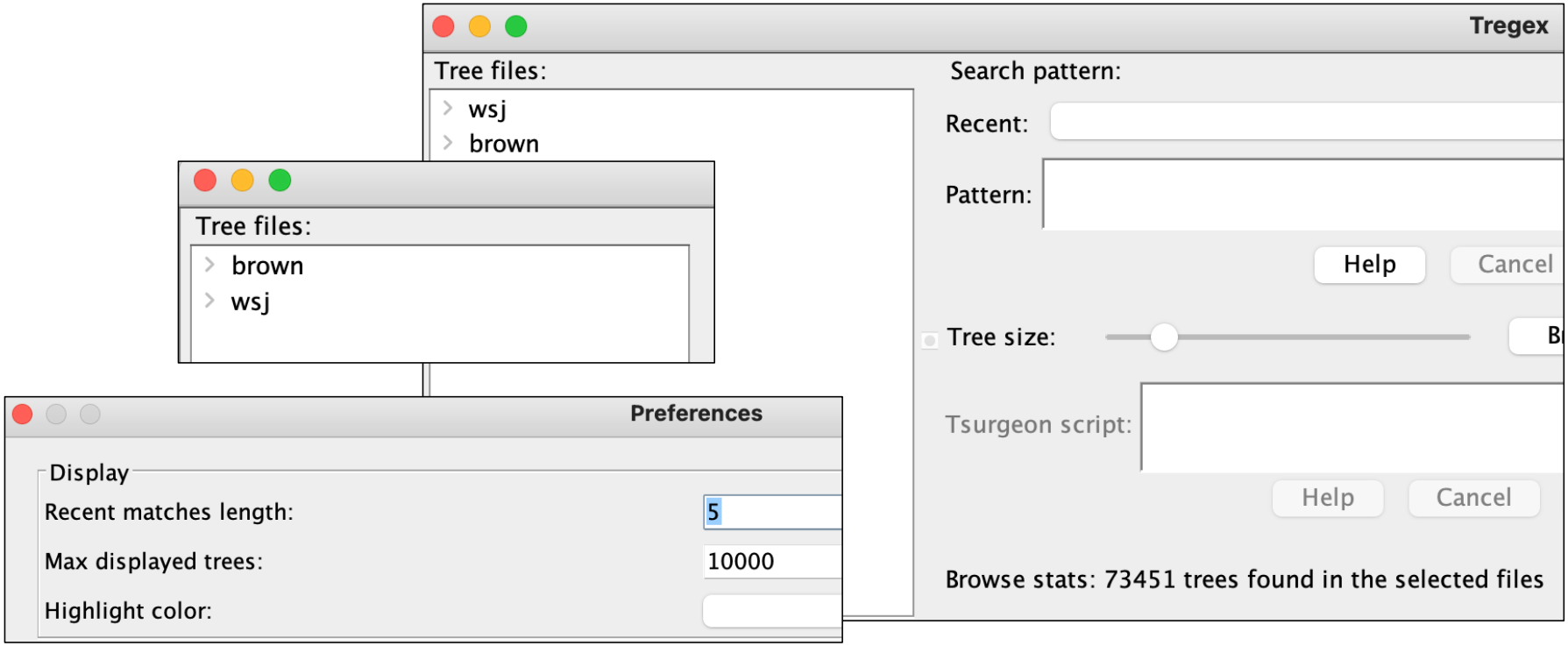


Build your search pattern a bit at a time and check your results!

1. Look for basic S (@S)
2. Immediately dominates a basic VP
3. VP immediately dominates another VP
4. This last VP must dominate both VBN and NP –NONE-

Homework 10

- Make sure you have brown and wsj loaded for the search.
- Max displayed trees: > 1000



Homework 10

- Other conditions to add to refine your search:
  1. basic S immediately dominates a subject NP with an *index*, a number.
  2. NP –NONE- immediately dominates ∗…–*index* (same index as in 1)
  3. **Note**: guide says ∗T∗–*index* (for trace of passive movement), but example in previous slide has no T, just ∗–*index*.
- Part 1: after stage 3 above, how many passives do you get? Report your search expression and statistics.
  4. We didn't specify the passive *be*. What are the forms of *be*?
  5. Add the restriction in 4. to the search.
- Part 2: after stage 5 above, how many passives do you get? Report your search expression and statistics.

Homework 10

- Part 3: can you give **examples** showing why there is a difference in matches between Part 1 and Part 2 (limited to passive *be*)?
  - i.e. what other verbs aside from *be* can be used to form a passive?
  - **Hint**: save matched sentences and use diff, then find the tree(s).
  - show tree or tree fragment screenshots

Homework 10

- Usual rules
- ONE PDF FILE ONLY!
- email: sandiway@arizona.edu
- subject: 581 Homework 10 *YOUR NAME*
- due date: next Monday midnight