

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 retrieve 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:

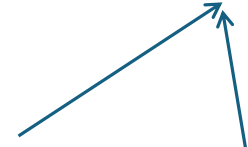
- X < , y
1st child y
- X <- y
last child y
- X \$+ y
x immediate
left sister of y

tregex

- Pattern:

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

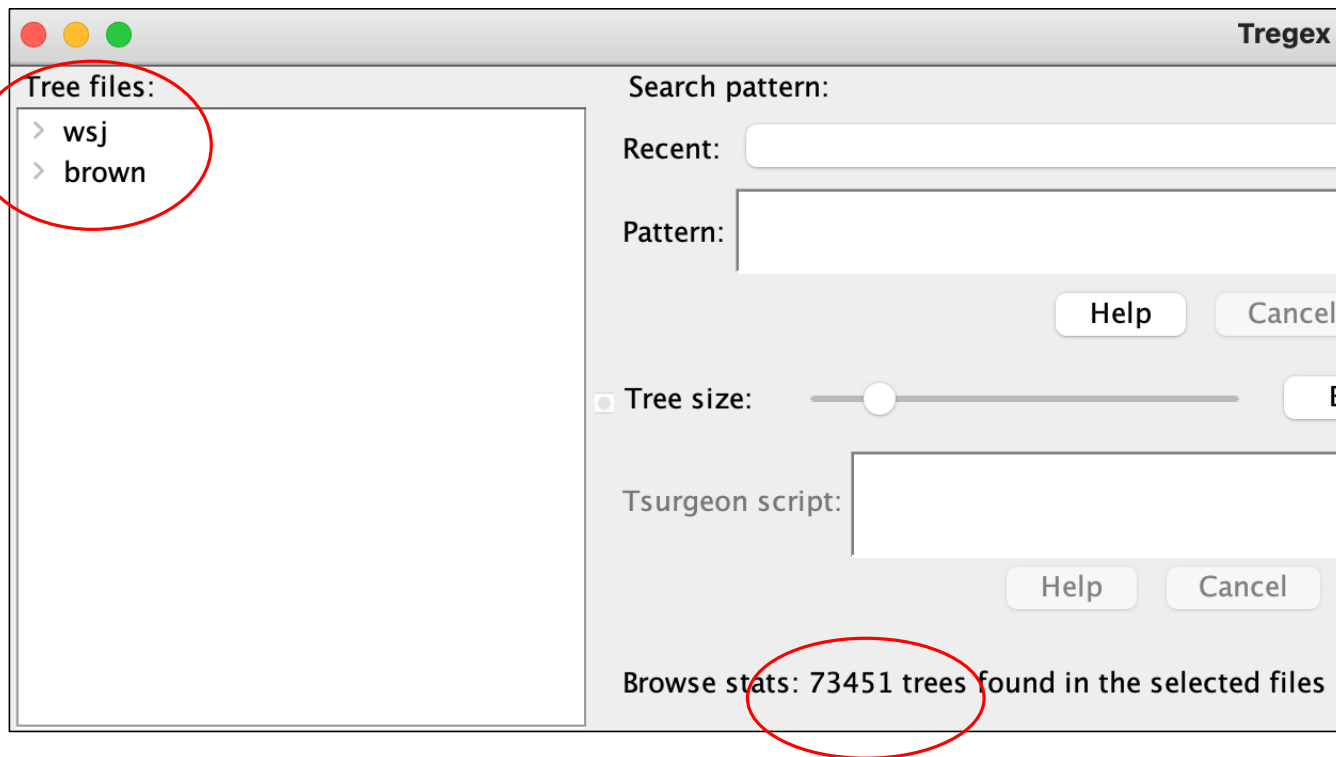
must be same node



The screenshot shows the tregex interface. The pattern field contains: `@NP <, (@NP $+ (/,/ $+ (@NP $+ /,/=comma))) <- =comma`. The tree size slider is set to a low value. The Tsurgeon script field is empty. The match statistics show 2961 unique trees found with 3077 total matches. Below the interface, a parse tree is displayed for the sentence "The asbestos fiber crocidolite is unusually resilient once with". The root node is S-TPC-2, which branches into NP-SBJ and VP. The NP-SBJ node branches into NP (DT: The, NN: asbestos, NN: fiber, ,: ,) and NP (NN: crocidolite, ,: ,). The VP node branches into VBZ: is, ADJP-PRD (RB: unusually, JJ: resilient), SBAR-TMP (IN: once, S: NP-SBJ, VP), ,: , and PP (IN: with, NP:).

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

tregex



tregex

The screenshot shows the tregex application interface. The main window has a search pattern field containing the regular expression: `@NP <, (@NP $+ (/,/ $+ (@NP $+ /,/=comma))) <- =comma`. Below this is a 'Recent:' list with a checked entry: `✓ @NP <, (@NP $+ (/,/ $+ (@NP $+ /,/=comma)))`. The 'Pattern:' field also contains the same regular expression. To the right, a list of text fragments is visible, including 't', 'sj', 'wsj', 'wsj', and 'wsj'.

Below the main window is a 'Statistics History' window. It features a table with the following data:

Pattern	Trees Matched	Total Matches
@NP <, (@NP \$+ (/,/ ...	4441	4647
@NP <, (@NP \$+ (/,/ ...	3146	3264

t regex

- 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 .
```

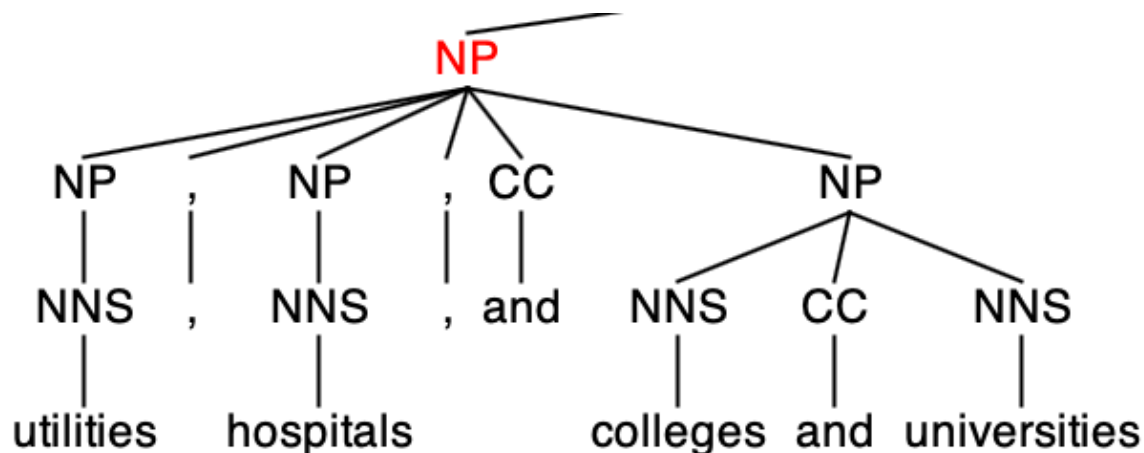
```
18d15
```

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


tregex

2d1

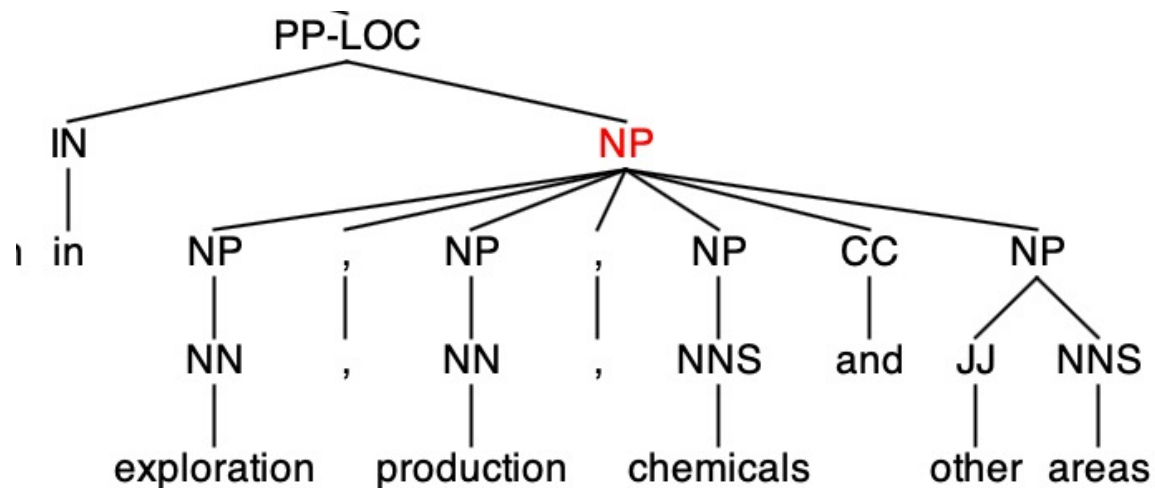
```
< wsj_0353.mrg-  
2 The Financial  
Accounting Foundation  
voted 12-2 that FASB  
accounting rules  
supercede GASB rules  
in regard to  
utilities , hospitals  
, and colleges and  
universities owned *  
by the government .
```



tregex

11d9

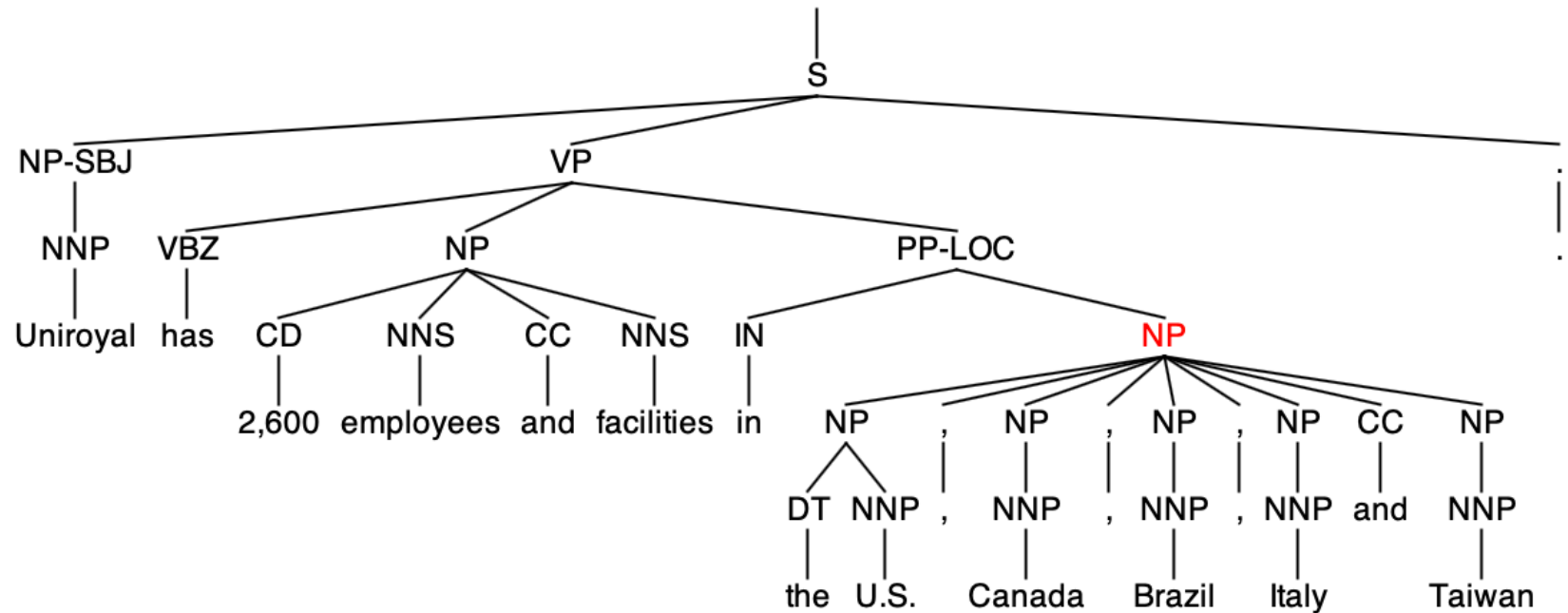
```
< wsj_0341.mrg-  
3 Unocal said 0 the  
venture would enable  
it to recover more of  
its refining and  
marketing investment  
and prepare for  
expected growth in  
exploration ,  
production ,  
chemicals and other  
areas .
```



tregex

18d15

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



t regex

- 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))

Tree size: Browse Trees

Tsurgeon script: Run script

Match stats: 11898 unique trees found with 13906 total matches. Statistics

wsj_0003.mrg-8 Neither Lorillard nor the researchers v
wsj_0003.mrg-13 Among 33 men who *T*-4 worked c
wsj_0003.mrg-16 `` The morbidity rate is a striking fi
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 industrial
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 blue
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 , for
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

```
graph TD
    S1[S] --- NP_SBJ1[NP-SBJ]
    S1 --- VP1[VP]
    NP_SBJ1 --- PRP[PRP]
    PRP --- it[it]
    VP1 --- VBZ[VBZ]
    VBZ --- enters[enters]
    VP1 --- NP2[NP]
    NP2 --- DT[DT]
    DT --- the[the]
    NP2 --- NNS[NNS]
    NNS --- lungs[lungs]

    IN[IN] --- with[with]
    S_NOM[S-NOM] --- NP_SBJ2[NP-SBJ]
    S_NOM --- VP2[VP]
    NP_SBJ2 --- NP3[NP]
    NP3 --- RB[RB]
    RB --- even[even]
    NP3 --- JJ[JJ]
    JJ --- brief[brief]
    NP3 --- NNS2[NNS]
    NNS2 --- exposures[exposures]
    NP_SBJ2 --- PP[PP]
    PP --- TO[TO]
    TO --- to[to]
    PP --- NP4[NP]
    NP4 --- PRP2[PRP]
    PRP2 --- it2[it]
    VP2 --- VBG[VBG]
    VBG --- causing[causing]
    VP2 --- NP5[NP]
    NP5 --- NP6[NP]
    NP6 --- NNS3[NNS]
    NNS3 --- symptoms[symptoms]
    NP5 --- SBAR[SBAR]
    SBAR --- WHNP1[WHNP-1]
    WHNP1 --- WDT[WDT]
    WDT --- that[that]
    SBAR --- NP_SBJ3[NP-SBJ]
    NP_SBJ3 --- NONE[-NONE-]
    NONE --- T1[*T*-1]
    SBAR --- VP3[VP]
    VP3 --- VBP[VBP]
    VBP --- show[show]
    VP3 --- PRT[PRT]
    PRT --- RP[RP]
    RP --- up[up]
    VP3 --- ADVP_TMP[ADVP-TMP]
    ADVP_TMP --- NP7[NP]
    NP7 --- NNS4[NNS]
    NNS4 --- later[later]
    ADVP_TMP --- JJ[JJ]
    JJ --- decades[decades]
```

tregex

- Different results from:

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

The screenshot displays the tregex application window. On the left, the 'Pattern' field contains the regular expression: `@SBAR < /^WH.*-([0-9]+)$/#1%index << (@NP < (/^-NONE-/ < /^*T*-([0-9]+)$/#1%index))`. Below the pattern field are buttons for 'Help', 'Cancel', and 'Search'. A 'Tree size' slider is set to a low value, with a 'Browse Trees' button next to it. A 'Tsurgeon script' field is empty, with 'Help', 'Cancel', and 'Run script' buttons below it. At the bottom left, the 'Match stats' section reports: '9319 unique trees found with 10474 total matches.' and a 'Statistics' button. On the right, a list of search results is shown, with the first entry highlighted in blue: 'wsj_0003.mrg-2 The asbestos fiber , crocidolite , is unusually res'. Other visible entries include 'wsj_0003.mrg-3 Lorillard Inc. , the unit of New York-based Loew:', 'wsj_0003.mrg-8 Neither Lorillard nor the researchers who *T*-3 s', 'wsj_0003.mrg-13 Among 33 men who *T*-4 worked closely with', 'wsj_0003.mrg-16 `` The morbidity rate is a striking finding amo', 'wsj_0003.mrg-18 The plant , which *T*-1 is owned *-4 by Hollin', 'wsj_0003.mrg-19 The finding probably will support those who *T', 'wsj_0003.mrg-20 The U.S. is one of the few industrialized nation:', 'wsj_0003.mrg-24 About 160 workers at a factory that *T*-8 mad', 'wsj_0003.mrg-27 Workers described `` clouds of blue dust '' tha', 'wsj_0004.mrg-15 It invests heavily in dollar-denominated securit', 'wsj_0005.mrg-1 J.P. Bolduc , vice chairman of W.R. Grace & Co. , v', 'wsj_0005.mrg-2 He succeeds Terrence D. Daniels , formerly a W.I', 'wsj_0008.mrg-4 Legislation 0 *T*-1 to lift the debt ceiling is ens', 'wsj_0011.mrg-4 South Korea 's economic boom , which *T*-12 b', 'wsj_0012.mrg-2 The new ad plan from Newsweek , a unit of the V', and 'wsj_0012.mrg-3 Plans that *T*-13 give advertisers discounts for'.

tregex

The screenshot shows the tregex interface with the following components:

- Pattern:** `@SBAR < /^WH.*-([0-9]+)$/#1%index << _ < (</^N ONE- / < /^*T*-([0-9]+)$/#1%index)`
- Buttons:** Help, Cancel, Search, Browse Trees, Run script, Statistics.
- Match stats:** 11898 unique trees found with 13906 total matches.
- Match list:** A list of matches with file names and snippets. The match `wsj_0003.mrg-25` is highlighted in blue, corresponding to the parse tree below.
- Parse Tree:** A hierarchical tree diagram for the sentence "In the factory, particularly dusty, where the crocidolite was used, areas of the fact...". The root node is `S`. It branches into `NP-SBJ` (prepositional phrase "In the factory"), `VP` (verb phrase "were particularly dusty"), and `SBAR-2` (relative clause "where the crocidolite was used..."). The `SBAR-2` node further branches into `WHADVP-1` ("where") and another `S` node. This second `S` node branches into `NP-SBJ-8` ("the crocidolite") and `VP` ("was used..."). The `VP` under `NP-SBJ-8` branches into `VBN` ("used"), `NP` ("areas of the fact..."), and `ADVP-LOC` ("...").

Reason for difference

Example:

WHADVP also possible (not just WHNP)

Homework 10

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

4.3	*	(trace of NP movement, controlled PRO, arbitrary PRO)	68
4.3.1		Indexing	68
4.3.2		Passives	69
4.3.3		Subjects of participial clauses and gerunds	70
4.3.4		Subjects of infinitival clauses	73
4.3.5		Subjects of <i>as-</i> and <i>than-</i> clauses	76

**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 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.

²We 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))))))
```

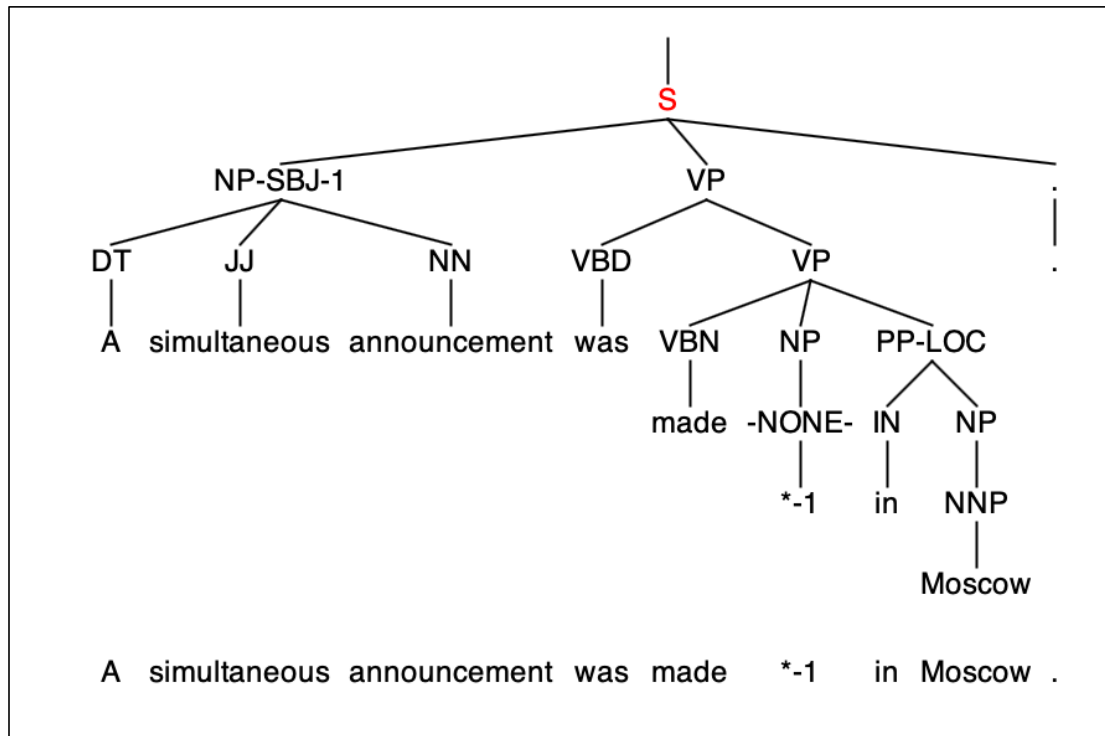
Note that the * may come before or after a PRT (particle). There is no policy governing this and either order is possible, though it is somewhat more likely for the PRT to come second:

```
(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

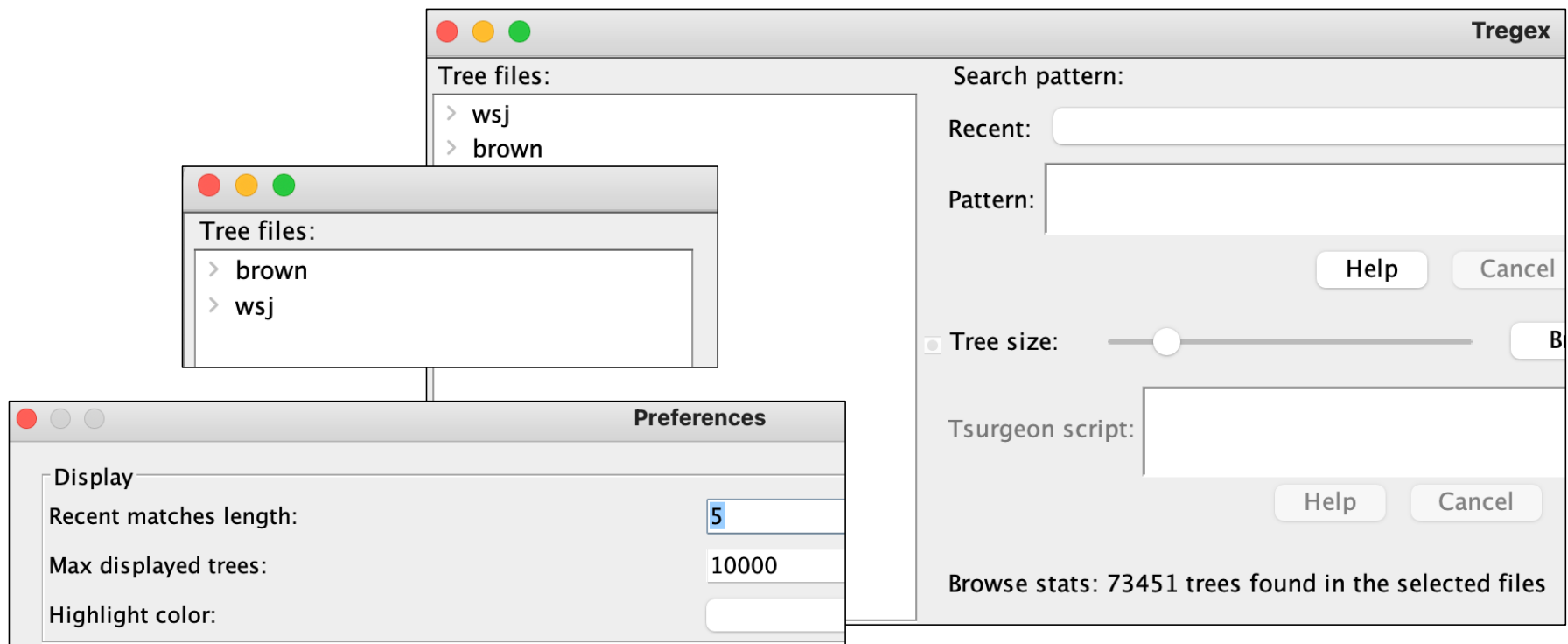


Build your search pattern a bit at a time and check your results!

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