Using *tgrep* (based on an earlier handout by Jeanette & Tantiana)

Local tutorial: http://www.stanford.edu/dept/linguistics/corpora/cas-tut-tgrep.html

What is tgrep?
tgrep is grep for trees. In other words, tgrep is designed to conduct searches on syntactically parsed corpora. You specify a pattern using node names and the relationships between the nodes and then tgrep matches that pattern against the corpus of tree structures.

How to use tgrep:
To use tgrep, you need to be logged into any one of the Stanford computers except cardinal. For more information about using tgrep and setting up your account, please see the "How to use tgrep" link at the bottom of the page.

Stanford currently has three corpora that are ready to be searched using tgrep :

- Brown — a 1 million word balanced corpus
- Switchboard — 1.4 million words of telephone conversations
- Wall Street Journal — corpus of Wall Street Journal articles (10 million words)

In fact, any corpus can be made "tgrep-able" – ask Roger Levy to set this up for you.

Some simple tgrep commands and options
commands:

A < B	A immediately dominates B
A < X B	B is the Xth child of A (i.e. the first child is A < 1 B)
A <-X B	B is the Xth to last child of A (i.e. the last child is A <- 1 B)
A <- B	the last child of A is B (same as A <-1 B)
A << B	A dominates B
A >> B	A is dominated by B
A > B	A is immediately dominated by B
A <<, B	B is a leftmost descendant of A
A >>' B	B is a rightmost descendant of A
A . B	A immediately precedes B
A .. B	A precedes B
A $ B	A and B are sisters (note that A $ A is FALSE)
A $. B	A and B are sisters and A immediately precedes B
A $.. B	A and B are sisters and A precedes B

You can make any of these negative by adding **!** in front of the relation. For example:
A **!**< B A does not immediately dominate B

tgrep patterns are composed of a node followed by the relationships which that node participates in. For example:

S < NP << S = S_i < NP and S_i << S

will match an S node which immediately dominates an NP **and** which dominates some other S node.

S < (NP << S) = S < NP_j and NP_j << S

will match an S node which immediately dominated an NP node which in turn dominates some S node.

options (for more options look at the tgrep man page):

-a to match on all the patterns (usually tgrep will just find the first match in a sentence and move on. This option tells it to look for all occurances in all sentences)

-w this returns the whole sentence (tgrep will usually only show you the node that you searched on but this option guarantees seeing the whole sentence)

-n puts the search string on one line (the parsed corpora are "pretty printed" across many lines – and this returns all on one line)

-t prints only the terminals

To make the search, first call tgrep, and then the option/s followed by the pattern you are looking for:

elaine42:~> tgrep –aw 'S < NP << S'

More information about tgrep

A useful place to find more information is the tgrep manual. Just type "man tgrep" at the prompt. For example:

elaine42:~> man tgrep