

Searching the Penn Treebank with **Tregex**

Corpus-Based Computational Linguistics
(Chris Brew and Michael White)

2008 Linguistics Summer Mini-Institute

In this activity, we will work through progressively more complex searches of the trees in the Penn Treebank (PTB) using the Stanford **tregex** tool. Some preliminaries:

- **Tregex** can be downloaded from <http://nlp.stanford.edu/software/tregex.shtml>. To run it, simply unpack the archive and double click on the `stanford-tregex.jar` file (assuming that a recent version of Java is installed).
- A **tregex** tutorial can be found at http://nlp.stanford.edu/javanlp/tutorials/tregex/The_Wonderful_World_of_Tregex.ppt.
- With the Penn Treebank, we'll be using the Wall Street Journal (WSJ) part of the corpus, with merged syntactic and part-of-speech annotation. Annotation guidelines are available from the Penn Treebank page: <http://www.cis.upenn.edu/~treebank/>. An HTML version of the bracketing guidelines is available at <http://purl.org/net/dm/07/autumn/795.10/ptb-annotation-guide/root.html>.

The activity is inspired in part by the discussion of comparative correlatives in Philip Resnik, Aaron Elkiss, Ellen Lau, and Heather Taylor, "The Web in Theoretical Linguistics Research: Two Case Studies Using the Linguist's Search Engine," *31st Meeting of the Berkeley Linguistics Society*, February 2005, pp. 265–276 (<http://umiacs.umd.edu/~resnik/pubs/bls2005.pdf>).

1. Are there any comparative correlatives (CCs) in Section 00 of the PTB? Comparative correlatives, also known as conditional correlatives and *more-more* constructions, are found in pairs of clauses such as *The more pizza Romeo eats, the fatter he gets*. To find out, start **tregex** and load the trees from Section 00, using the **Load trees...** item under the **File** menu. Then, enter a query in the **Pattern:** box that looks for a determiner (DT) that has *the* as its child and has a comparative adjective (JJR) as a sibling (< specifies immediate dominance; \$ specifies siblinghood; a summary of the **tregex** query syntax can be found by pressing the **Help** button). Next press **Search**, and after the search completes, click on each matching sentence in the **Matches:** pane to see whether this search turned up any comparative correlatives.

2. Though not strictly necessary, clauses in comparative correlatives often contain *the JJR* with no following head noun, as in *the fatter he gets*. With this in mind, let's expand and refine the search. First, load the trees from all sections of the WSJ part of the PTB, using **File|Clear tree file list** and **File|Load trees...** Next, search for some constituent that dominates the determiner *the* and which has a comparative adjective as its final child. To match any node in the tree, you can use `_`, i.e. two underscores; to specify the last child, use `<-`. Note also that the query $X < Y < Z$ means an X that dominates a Y and a Z ; to specify an X that dominates a Y which in turn dominates a Z , you use parentheses around the Y and Z as in $X < (Y < Z)$. Did your search turn up any CCs?

3. As it turns out, the PTB has enough instances of CCs to warrant their discussion in the bracketing guidelines, but not enough to warrant too much thought about how to best analyze them. Accordingly, the PTB guidelines state that the comparative phrases in CCs should be placed under a node with label X (unanalyzed, unexplained, etc.). We can take advantage of this annotation to look for such comparative phrases by searching for nodes with this label. However, as the X nodes sometimes have function tags, we'll want to search for nodes whose labels start with X; to do so, we can use the regular expression `/^X.*`. Additionally, as we'd like to include sentence-initial *The*, we could use a regular expression to allow for a capital *t*; in this case though, it suffices to just look for a determiner under the X node, which is invariably *the* (capitalized or not). Finally, as you may have noticed, the comparative phrases sometimes involve comparative adverbs (RBR) instead of comparative adjectives, which need not appear as immediate children of X (cf. *the more "earthquake-resistant"*, which contains an adjective phrase); accordingly, we'll want to specify not-necessarily-immediate dominance (<<) of a node that's either a comparative adjective or adverb (JJR|RBR). Putting all this together into a query, how many CCs does your search find?

4. McCawley’s (1988) generalization that copula deletion in CCs is only licit when the subject of the clause is generic, rather than specific, has long been accepted without challenge in the literature. However, using data found with the help of the Linguist’s Search Engine (LSE), Resnik et al. (2005) suggest that parallelism is also an important factor:

A search for naturally occurring CCs, however, showed that there is more to the story. Using the “query by example” process, a search of the static LSE Web collection (3.5 million sentences) yielded an unexpected result. While it was true that, in instances of copula deletion, CCs commonly occurred with generics in their subject, it was more striking that *all* instances of copula deletion included deletion of a main copular verb *in both clauses*. [emphasis original]

Do the CCs in the PTB support this observation? To find out, we can refine our query for CCs to look for comparative phrases that are not accompanied by a verb. One way to do so is to add a constraint that the X node **not** have a sibling that dominates any verb at all. Negation can be specified by prefixing a relation with an exclamation mark, so !\$ means no sibling; to look for a verb in any form, we can use the regular expression /VB.*/. How many sentences does your search turn up? Do they all involve parallel copula deletion?