

Searching the CCGbank with **Tregex**

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

2008 Linguistics Summer Mini-Institute

In this activity, we will look at how various phenomena involving long-distance dependencies are handled in the CCGbank, as well as how these interact with the assignment of Propbank argument roles. Some preliminaries:

- The activity requires a version of the CCGbank formatted to work with **tregex**, as well as a patched version of **tregex** that makes it easier to search the corpus. Note that this version of the CCGbank uses angle brackets in categories where parentheses would normally be used, so that **tregex** can parse the trees. However, parentheses in categories are supported in queries, as long as categories are quoted.¹ Kudos galore to Steve Boxwell for making these versions available!
- The patched version of **tregex** is called `stanford-tregex-patched-sab-jul08.jar`. To run it, simply double click on the file (assuming that a recent version of Java is installed).
- There are actually two versions of the CCGbank we'll look at: `CCGbank.Tregex.Without.Sem` and `CCGbank.Tregex.With.Sem`, where the latter has Propbank semantic roles integrated. While for this activity, it will suffice to run the queries on Section 00, the queries can be run on the full corpus if desired.
- The CCGbank User's Manual, which includes a comprehensive description of the algorithm for translating the Penn Treebank syntactic and part-of-speech annotations into CCG derivations, can be downloaded as a University of Pennsylvania technical report from <http://www.cis.upenn.edu/departamental/reports/CCGbankManual.pdf>.
- Propbank framesets can be consulted online: for example, to see what ARG roles the verb *reward* assigns, go to <http://verbs.colorado.edu/framesets/reward-v.html>.

¹The quotes trigger a rewrite of the category into a regular expression with appropriate escaping. Contributing code to **tregex** to support CCGbank out of the box remains for future work.

The activity is inspired in part by the discussion of how the C&C parser fares on long-distance dependencies in Stephen Clark, Mark Steedman and James R. Curran, “Object-Extraction and Question-Parsing using CCG,” *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, 2004 (<http://www.iccs.informatics.ed.ac.uk/~stephenc/papers/emnlp04.pdf>). The integration of Propbank roles, along with corrections to the argument/adjunct choices in the CCGbank, is discussed in Stephen A. Boxwell and Michael White, “Projecting Propbank Roles onto the CCGbank,” *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08)*, 2008 (<http://www.ling.ohio-state.edu/~mwhite/papers/Boxwell-White-LREC08.pdf>).

The steps:

1. First let’s see how the frequency of subject relative clauses, which involve a local dependency, compares to the frequency of object relative clauses, free object relatives, and *tough*-adjectives, which all involve long-distance dependencies. After loading the trees for Section 00, simply search for the categories for the appropriate trigger words:

- subject relative pronoun: $(NP\backslash NP)/(S[dc1]\backslash NP)$
- object relative pronoun: $(NP\backslash NP)/(S[dc1]/NP)$
- free object relative *what*: $NP/(S[dc1]/NP)$
- *tough*-adjective: $(S[adj]\backslash NP)/((S[to]\backslash NP)/NP)$

To find these categories, just include them in quotes, then add a (vacuous) constraint that the node dominates an anonymous node (i.e. `< __`), since `trexex` syntax does not allow single-node queries. Do the counts match your expectations?

2. In the CCGbank, reduced relative clauses (with no relative pronoun) involve a unary rule that has the same effect as an explicit relative pronoun: that is, it maps the category $S[dc1]/NP$ for a sentence missing a noun phrase on the right to an NP post-modifier, $NP\backslash NP$. Are reduced relatives more or less common in Section 00 than ones with an explicit object relative pronoun? To find out, devise a query to match applications of the rule $S[dc1]/NP \implies NP\backslash NP$ in the derivations.
3. You may have noticed in the first step that relative pronouns usually are tagged WDT, but occasionally have IN as their POS tag. The ones tagged IN are incorrect, i.e., not in conformance with the following discussion in the PTB tagging guidelines:

IN or WDT

When *that* introduces complements of nouns, it is a subordinating conjunction (IN).

EXAMPLES:

the fact that/IN you're here
the claim that/IN angels have wings

But when *that* introduces relative clauses, it is a wh-pronoun (WDT), on a par with *which*.

EXAMPLE:

a man that/WDT I know

It's well known that despite the annotators' best efforts, the POS tags in the PTB are far from perfect. So it's not surprising that relative pronouns would sometimes be mistagged as IN instead of WDT. But are the errors evenly distributed? Construct queries to see how often *that* is mistagged as IN when used as a subject relative pronoun versus an object relative pronoun. (You can save your results using `File|Save statistics` if you like.)

4. With Propbank semantic roles, how often do subject NPs realize ARG0 roles versus other roles? Are non-ARG0 roles rare? With the version of CCGbank with integrated Propbank roles, it's easy to find out. First, clear the tree file list and load the trees for the version of the CCGbank with the Propbank info. You can find examples of finite verbs by searching with the query "S[dc1]\NP" < /VB/. Looking at examples other than forms of *be* and auxiliary *have*, you should see that the Propbank roles that a verb assigns are listed in square brackets. As such, you can add to your query that the first argument expresses the ARG0 role by skipping intervening characters, as follows: "S[dc1]\NP.*[ARG0]". Armed with this pattern, you can now check to see how often subject NPs of finite verbs realize ARG0 versus ARG1, ARG2 and ARG3.
5. What roles do object relative pronouns fill? You can search for a relative pronoun and see that the ARG role whose expression is mediated by the relative pronoun is also listed. Adding ARG0, ARG1, and ARG2 to your object relative pronoun query, find out how often these roles are expressed in this way (n.b.: you don't want the left square bracket before ARG this time). Is anything other than ARG1 ever correct?