

Assignment 5

L615

Due Thursday, April 30

1. Using TIGERSearch and the Penn Treebank corpus, provide:
 - (a) The number of relative clauses modifying subject NPs, as in:
The guy *that you met* is an old friend of mine.
 - (b) The number of relative clauses modifying object NPs, as in:
That guy hates the author *that I was telling you about*.
 - (c) The number of headless relative clauses modifying subject NPs, as in:
The guy *you met* is an old friend of mine.
 - (d) The number of headless relative clauses modifying object NPs, as in:
That guy hates the author *I was telling you about*.For each case, also provide the TIGERSearch query you used and any difficulties you encountered.
2. Using the BooTCaT tools, create 2 small comparable corpora from the web, which are in 2 of your favorite languages (or your 2 favorite regional dialects). In other words, create corpora for the same topic in 2 different languages, and when I say “small”, that means you can use the default settings (e.g., 10 URLs per query).
 - (a) Report exactly what you did, including, e.g., the seed terms for each language.
 - (b) Report the following basic statistics for each corpus:
 - Number of tokens
 - Number of types
 - Number of URLs
 - Average document length (in words)
 - (c) Compare the 20 most frequent words in each corpus. Do you notice any particular trends?
3. I’m not going to give you an additional assignment for your final project, but will just outline the final details for what you need to do.
 - (a) Your final paper (10-20 pages) will be due on Tuesday, May 5 at 10:15am.
 - (b) Also at 10:15am on Tuesday, May 5, you will give about a 10-minute presentation to the class on your corpus project. This is less something that you’ll be graded on and more a chance to share work with each other.

If you have any questions about any components of the project, let me know.