Phase 1

We’ve mostly discussed searching for content, i.e., trying to find webpages which match a particular meaning. But say you’re a linguist trying to find certain types of structures (e.g., rare wh-word phenomena, strange/atypical word orders, etc.)

For the question below, you can think about any construction you want, but if you’re having trouble making it concrete, think about one or all of these:

- Heavy-NP shift: *The bank robber sent into the safe all of us busybodies who couldn’t keep our mouths shut.*
- Locative inversion: *Out the door ran the thief.*
- Non-constituent coordination: *He gave me a warning shot and the pretty clerk two warning shots.*
- Acceptable subject-verb “disagreement”: *The ham and eggs at table 2 wants more OJ.* (example inspired by Pollard & Sag 1994)

1. How do you design a search engine to find such types of things? In particular:
   (a) What kind of back-end processing do you need (e.g., tokenizers, POS taggers, etc.)?
   (b) What kind of search interface should a linguist have? What should they be able to do? i.e., a keyword search interface seems insufficient here.

Phase 2

(If you have access to a web search, do this on your own; otherwise, come up with ideas for the first part & we’ll do the second part of it together as a class.)

Most corpora do not have much in the way of annotation (e.g., syntactic mark-up), but it is fairly easy to get part-of-speech (POS) categories with a high degree of accuracy.

1. Consider a syntactic phenomenon above and assume you only have POS. How would you approximate your search for the items?

2. Go to [http://corpus.byu.edu](http://corpus.byu.edu) and try your ideas out.

Phase 3

Returning to Phase 1, think about the back-end processing required for your linguistic search engine.

1. Do you think the back-end technology will be helpful for non-linguists who have no idea what these phenomena are? If so, how so? If not, why not? Think about both information retrieval (goal: return relevant documents/webpages) and information extraction (goal: return relevant pieces of information about who did what to whom).

Phase 4

(This one is totally on your own & for your own enjoyment ...)

The Linguist’s Search Engine is the type of tool designed to help for Phase 1 questions above.

- The original work: [http://lse.umiacs.umd.edu](http://lse.umiacs.umd.edu)
- A different tool: [http://wse1.webcorp.org.uk/home/](http://wse1.webcorp.org.uk/home/)