Kucera Corpus Handout  
*Get your Corpus on*!

Getting Started

The first thing to know about working with these datasets is how to use Regular Expressions (RegEx). There are many resources out there that will help walk you from basics to advanced, scripted searches. For the purpose of this workshop, I will provide you with some entry level explanation.

Regular expression 101: Basic Searches

* Use quotations to find a specific word: “word”
  + This will only find *exactly* what is within the quotes; it is even case sensitive.
  + You can string these together to search for fixed phrases like: “kick” “the” “bucket” will give *only* “kick the bucket”

Regular expression 201: Special Characters

* “ . ” The period is a “wild card” character: “.eat”
  + Gives Feat, Meat, Beat, etc
  + Does NOT produce Sweat, Cleat, Bleat
* “ ? “ The question mark treats what it follows as optional: “Mrs?”
  + There can be 0 or 1 of the preceding characters
  + Gives results for both Mr and Mrs
* “ + “ The asterisk treats the previous character as *one or more* occurrences: “s+uper”
  + Gives results for, super, ssuper, sssuper, etc.
* “ \* “ The asterisk treats the previous character as *zero or more* occurrences: “noo\*b”
  + Gives results for nob, noob, nooob, noooob, etc,
  + Very similar to the *plus* *sign*, with a small but, crucial difference.

Regular expression 202: Support Characters

* “( )” Parenthesis are used to group things together. “super” “(fun)?”
  + This produces either “super” or “super fun”
  + Here it groups all the letters of “fun” together so that they are optional all together or not at all.
* “ | ” This is used as an *or* operator. “Thank” “(you|god|heaven)”
  + Yields “Thank you” *and* “Thank god” *and* “Thank heaven”
* “[ ]” The bracket are an array and match any single character inside it: [aln]
  + This produces a match to *either* a *or* l  *or* n.

Regular Expression 301: Part of Speech (PoS)

* It uses the format of [ pos = “CODE” ]
* Available codes search for *types* of words such as
  + **vvd** – past tense verb
  + **nn** – single common noun
  + **pn** – nominal pronoun (e.g. anyone, everything)
  + **to** – infinitive “To”
  + **vbn** – past participle of lexical verb
  + jj – general adjective

Regular Expression 302: Lemma

* It uses the format of [ lemma = “word” ]
* This searches *all* iterations of a word: [ lemma = “red” ]
  + This produces: red, redden, reds, redneck, etc.

If you are interested in using this resource, we would be happy to get you setup with a user account and some training. Below are a few additional simple commands and features you can look forward too.

NAVIGATION:

* Use the arrow keys to move line by line
* Use the space bar or “PgDn” to move down a page at a time
* Use “PgUp” to go up a page at a time
* “q” exits the current search

USABILITY:

You can:

* save your searches to “.txt” files for export to your own computer.
* modify how many words or letters appear on the either side of your search
* turn on the PoS and Lemma tags so you can see them.
* see what each corpus supports
* create repeatable saved searches that you can edit or repeat.