**Natural Language Processing**

Until now we have considered only ‘bag of words’ implementations where words are things without meaning other than their frequency. We haven’t incorporated any information about the meanings of words (except for removing certain ‘stopwords), the contexts in which they are being used etc. ‘Semantic’ methods consider these things. The options are broad, from identifying groups of words associated with particular topics or sentiment, to weighting words based on their similarities, to identifying the subjects or objects referenced in sentences.

Political scientists haven’t used semantic methods much to date. The main exception is in event data analysis that use subject-verb-object relationships in news stories. There are valuable opportunities here for creative minds! For example:

**N-grams**

* N-grams: e.g. bigrams (‘not happy’) versus a unigram approach (‘not’, ‘happy’) – big difference in meaning!

**Dictionaries/LIWC**

* Linguistic inquiry and word count provides a dictionary of about 6000 commonly used words that experts have grouped into a large number of categories (e.g. religiousity; sadness). It is a commercial site that originated with the [General Inquirer](http://www.wjh.harvard.edu/~inquirer/homecat.htm) lexicon. The advantages of preexisting dictionaries like LWIC are that they are broadly accepted and cover a lot of ground. The disadvantage it does not do any one category really well. (In addition, I believe that its word list is limited to unigrams?)
* Researchers have constructed lots of dictionaries in the past so it is worth searching before you start to construct your own. To create a dictionary from scratch, consider generating a word frequency list from your documents; sort the list by word frequency; and then tag the relevant words for each category of interest (starting with the most common words – after removing stopwords (and possibly stemming) of course. You can do the same thing for bigrams etc.
* Quanteda offers an interface for doing [LIWC like](https://github.com/kbenoit/LIWCalike) analyses (but you will need to provide the dictionaries)

**Parts of speech (POS) tagging**

The Python [NLTK](http://www.nltk.org/) package makes it possible to tag the words in a text for grammatical syntax. There is also an R [OpenNLP](https://cran.r-project.org/web/packages/openNLP/openNLP.pdf) package that I’ve never used. The [tag options](https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html) allow you to extract all sorts of information:

* Proper nouns (NNP): Extract or count names from documents
* Verbs and adjectives (VB, VBZ etc): Extract only words that are most likely to capture sentiment
* Pronouns (PRP, PRPS): Compare gender references or politicians’ tendency to frame politics as about us versus them
* Subjects/verbs/objects (who did what to whom?)

The options are truly endless: https://sites.google.com/site/partofspeechhelp/home/prp\_prp

**Wordnet** is a networkmap of words where word similarity and hierarchy is defined quantitatively. It can be used to identify closely related word and to discover different usages of the same word. For example, different ways of referring to dogs could be grouped (puppies) into a single category, while a research could also check for ways in which counting references to ‘dog’ might be misleading.

Try this. Go to <http://wordnetweb.princeton.edu/perl/webwn>

Enter ‘dog’ in the search. This provides a list of all of the distinct usages of dog as noun and verb. Then click on the S for the first case. Now you have a bunch of synset options (or similar words). Click on direct hypernyms – these are the most immediate more general categories of which dog is a part. Click on direct hyponyms – these are more words that refer to the more general category of dog. Etc!

And here’s an interactive visualization of the WordNet you just explored:

<http://people.csail.mit.edu/torralba/research/LabelMe/wordnet/test.html>

**Where does this stuff come from?** Before statistical machine learning, the objective of artificial intelligence research was to create a complete mapping of language (this is a bit oversimplified but think of a French-English translation tool that includes a map between every French sentence and every English sentence). Tons of money went into this effort during the 1970s and 80s, but it lost support due to slow progress. Statistical machine learning reinvigorated artificial intelligence research and these resources developed have proved to be very valuable complement to machine learning research.