Searching word forms

Searching on the word level means searching for:
- word forms
- lemmata
- POS tags
- word senses
- ...

We will focus today on online search interfaces
- Word form searches will be aided by knowing regular expressions

BYU Search Interface

Mark Davies has a search interface at:
http://corpus.byu.edu/
- BNC, COCA, etc.
- Note the meta-data for each corpus.
- Some search options have different meanings than with regular expressions
  - * allows for anything after a character (e.g., [n*])
  - ? is a single character wildcard
  - . allows one to specify that two pieces of information are true about the same token (e.g., [cool],[i]*)
  - = for synonym search (e.g., [waste] [=energy])

BNCWeb

BNCWeb allows access to the BNC
- http://corpora.lancs.ac.uk/BNCweb/
- simple query
- CQP

BNCWeb

Simple Query

- wildcards: * for zero or more characters, + for one or more characters, ? for one character, and ??+ for three or more characters
- examples:
  - un+able
  - un??+able
- specify words: + for one word, ++ for two words, ...
- examples:
  - pick + up
  - pick ++ up
- sets over words:
  - on [his,her] advice
- search for POS tags:
  - _AJS results
  - beer can_VM0
  - beer can_{V}
CQP (Corpus Query Processor) is the underlying technology for BNCWeb

▶ A part of the Corpus Workbench (CWB):
   http://cwb.sourceforge.net
▶ A Perl API is available
▶ It works for any word-based annotation (POS, lemma, etc.)

CQPweb is a web-based graphical user interface (GUI):
http://cwb.sourceforge.net/cqpweb.php

▶ Some corpora using CQP: https://cqpweb.lancs.ac.uk, https://corpling.uis.georgetown.edu/cqp/

Some notes on the syntax:

▶ Double quotes indicate word boundaries (shorthand)
  ▶ Longer form: `[word="in"] [word="spite"] [word="of"]`
  ▶ Order-independent search not possible
▶ Bracket notation allows for specifying multiple properties
  ▶ `[word="beer"] [word="can" & pos="V.*"]`