Describing Quantitative Relationships Using Informal Grammar: Appendices

Milo Schield, W. M. Keck Statistical Literacy Project

This is the Appendix to Schield (2011) located at www.StatLit.org/pdf/2011SchieldASA.pdf. This appendix is a separate document because the paper used the 15 page limit set by the ASA.

Appendix A: Accessing WordBanks

As of 2011, the new 500 million word Harper-Collins WordBanks corpus is accessible at http://www.collinslanguage.com/wordbanks/. An earlier version of this database (Cobuild) was used by Schield (2000) and by Schield and Burnham (2001). The current version is much larger and much more powerful. Every word is tagged by their lemma, a detailed part of speech and their high-level part of speech (PoS): adjective, adverb, conjunction, noun, preposition, pronoun, and verb. Each sentence has sentence boundaries.

http://www.collinslanguage.com/wordbanks/termsandconditions.pdf
http://wordbanks.harpercollins.co.uk
http://wordbanks.harpercollins.co.uk/auth/?module=login System login.

WordBanks program documentation is helpful but does not seem to be complete:

http://wordbanks.harpercollins.co.uk/Docs/Help/guide.html General introduction.
http://trac.sketchengine.co.uk/wiki/SkE/CorpusQuerying Details on CQL language.
http://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html Penn Treebank Tag-Set.

The details of how we accessed the Word Banks corpus and the naming convention for our reports along with our list of the word tags are at www.StatLit.org/pdf/2011SchieldASA-WB.pdf. A complete list of all the user-created filenames and the associated selection criteria is at www.StatLit.org/pdf/2011SchieldASA-WB1.pdf. A list of all the files that were created is at www.StatLit.org/pdf/2011SchieldASA-WB2.pdf. The actual data files themselves are not readily available to other researchers under the terms of our signed agreement with Harper-Collins:

> *"This Agreement is personal to You and neither the benefit of this Agreement nor any rights of access to any of the Services nor any right to use the Data may be assigned, licensed or otherwise transferred without the prior written consent of Collins."*

Appendix B: Setup and SubCorpus Selection

Setup included these View and Download Options.

- Attributes: Check Word and Check Tag.
- Structures: Select <s>. References: Select doc.Subcorpus.
- Display Attributes: Select KWIC tokens only,
- Page Size [for viewing]: # lines 100.
- KWIC Context Size [# characters/line]: Select 1000 [Max seems to be 500 + KWIC].
- Icon for 1 click copying: Uncheck.
- SaveOptions: Press to save the above choices.

To maximize the quality of the material, all searches were done on a sub-corpus that excluded all ephemera and excluded all transcriptions of spoken material. On startup, expand KEYWORD to show options and expand TEXT TYPES to show Sub-corpora.

Appendix C: Keyword Selection

There are five textboxes into which selection words or phrases can be entered. In each case Cobuild Query Language (CQL) is generated.

- Query textbox: Words entered into the Query textbox are used to generate lemmas. Entering *make* into the Query textbox generates 894,822 hits: *made* 353801, *make* 346182, *making* 115440, *makes* 79355 and *makin* 44. The CQL generated by this query: word, [word= "(?i)make"|lemma="(?i)make"].
- Keyword Lemma textbox: Entering *make* into the Lemma textbox generates 893,484 hits: *made* 353,801, *make* 344,844, *making* 115440, *makes* 79,355 and *makin* 44. Note the

small difference in “making”. The CQL generated by this query: word,[lemma="make"] The benefit of using the lemma textbox instead of the general Query textbox is that the Lemma text box allows one to couple the lemma selection with a part of speech (PoS). Note that case must be an exact match.

- Keyword Phrase textbox: Entering *make* into the Phrase text box generates 336161 hits: *make* 336161. Entering “Make” generates 9185 hits: Make 9185. The CQL generated by this query: word, "make" To get both upper and lower case, use the word-form or CQL textboxes.
- Keyword Word-form textbox: Entering *make* into the Word-form textbox generates 346,182 hits: *make* 336,161, *Make* 9,185 and *MAKE* 836. The CQL generated by this query is: word,[word="(?i)make"]
- Keyword CQL textbox: Entering “make” into the CQL textbox generates 379,565 hits: *make* 379,565.

Appendix D: CQL Language

The Cobuild Query Language (CQL) is minimally documented. It has similarities and differences with DOS and GREP. The CQL language uses the vertical bar “|” for "or" (disjunction), the ampere sign “&” for "and" (conjunction) and the exclamation mark “!” for "not" (negation). CQL commands are case sensitive. The “t’ in tag and the “w” in word must be lower case. There are three ways/levels of describing an OR relationship involving tags or words using the vertical bar:

- [tag = "IN" | tag = "PP"]; [word = "is" | word = "as"]
- ([tag="IN"] | [tag="PP"]); ([word="is"] | [word="as"])
- [tag = "IN | PP"]; [word = "is" | "as"]

There are three more special CQL characters: period, question mark and asterisk. (1) A period (dot) indicates any non-blank character in that position. So, tag=”NN.” does not return tag = “NN”; tag="N.S" will match NNS and NPS but not NN or NP. (2) A question mark indicates the previous character/instruction is optional. Alternatively, there are 0 or 1 of the previous character or instruction. (3) An asterisk indicates 1 or more repetitions of the previous character/instruction.

The following illustrates how these three special characters can be used:

- CQL for nouns: [tag="N."] matches NN and NP but not NNS. [tag="NN."] matches NNS but not NN. [tag ="N.?"] matches NN and NP; not NNS or NPS. [tag="NN.?"] matches NN and NNS. [tag =”N..?"] matches NN, NP, NNS and NPS. [tag="N*"] matches NN, but not NNS.
- CQL for verbs: [tag="V.*"] matches all verbs. [tag="V.*" & tag!="V.G"] excludes gerunds. [tag="V.*" & tag!= "V.N"] excludes past participles. [tag="VV."] excludes “to be” and “to have”
- CQL for comparatives and superlatives: [tag="..R"] or [tag = "JJR" | "RBR"]. [tag="..S"] or [tag = "JJS" | "RBS"].

CQL spacing may be fixed or variable. CQL fixed spacing has the form {k, k} where k is the number of characters to be skipped. CQL variable spacing has the form {j, k} where j < k. Variable spacing allows for a word or tag to be optional. E.g., the CQL query, "a" [word="big"]{0,1} "box", will match either “a box” or “a big box”.

Double quote = two single quotes. End of sentence is one character. To select all one character tags use "." General advice. Always end a CQL query with “within <s>” to avoid spanning multiple sentences.

Appendix E: Frequency Reports

Frequency reports can be extremely useful in detecting patterns and connections. General advice. (1) Always check “ignore case” when selecting on a word. (2) Include the node so the report clearly distinguishes L1 from R1 (unless the node has too many internal variations). (3) Establish a file naming convention (see next section) that clearly identifies how the report was generated. (4) Generate frequency reports on the full select before randomly sampling text for downloading.

Appendix F: File Naming

WordBanks embeds the selection criteria and sort history in the file header for all saved files. Unfortunately, it does not embed the commands used to create frequency reports. To interpret the content of a frequency file, the file name should contain that information. All filenames of our downloaded files follow this structure:

- Four-digit ID: the first number is always 2 signifying our 2nd analysis. The last three numbers indicate the unique selection used to obtain the associated data.
- Text indicating the select criteria without hyphens.
- Hyphen introducing the specifics of the file. Only one hyphen is permitted per file name to simplify access by other tools (GREP and Excel).
- Either FREQ, LIST or LISTS.
  - FREQ indicates summarized frequency reports. (1) indicates a tag (t) only for a given place; (2) indicates a tag (t) and word (w) or lemma (m) for a given location; (3) indicates just words (w) or lemma (m) for a given place. Place is indicated by location from the node. Node: All text that matches the select criteria. L1 indicates one place left of the node. R1 indicates one place right of the node.
  - LIST. This indicates downloaded lines
  - LISTS This indicates the downloaded lines involve just the sentence. This option is enabled by toggling ViewOptions: to display KWIC/Sentence before SAVE.
- Underscore and numbers after LIST or LISTS introduces # of lines matching select while optional underscore introduces actual # of lines sampled (if different from # selected).

Appendix G: Text File Save

There are three key choices in preparing a textfile for Save. (1) Download the whole file or a part (sample). (2) The sort order of the records. (3) Download a 512K block of text per match or download just the sentence containing the keyword in context (KWIC).

When the downloaded text files are large (over 5,000 lines), two kinds of problems emerge in downloaded files. (1) The files are truncated at 1,000 lines because the preset limit of 1,000 lines was not overwritten. (2) The files are truncated at just a few hundred lines due to some system problem. These can be identified by sorting the files by actual size and looking for mismatches.

Save options: SaveConcordanceAs *Text*. Save Pages: *All*. Check *Include heading*. Check *Include Align KWIC*. Max #Line: *10,000*. Do this for each save. Press: *SaveConcordance*.

Appendix H: File Name Report

A special program (CopyFileNames v 3.1 from www.extrabit.com/copyfilenames) was used to copy the file names of the files into Excel. Once the file names are in Excel, the file name is broken into components so that a summary pivot table can be generated. ID: =Left(cell, 4). TYPE: =MID(cell, SEARCH("-", cell,1)+1, 5). NAME: =MID(cell, 5, SEARCH("-", cell, 1) -5). ID-NAME: =IdCell &"-"& NameCell. The pivot table uses ID-NAME for the rows, TYPE for the columns and counts the IDs. This table has one line per four-digit ID. It summarizes how many files are of the various types (FREQ vs. LIST) and within each type of FREQ (1, 2 or 3). This report is at www.StatLit.org/pdf/2011SchieldASA-WB2.pdf.

Appendix I: WordBanks Recommendations

Although the current version of WordBanks is certainly superior to the Cobuild version in place 10 years ago, there is still room for improvement: (1) In all files include the sub-corpus used. (2) Include the frequency selection in the header of the frequency reports. (3) In the save file header, include the saved format used: full vs. KWIC/sentence. (4) Improve the documentation. (5) Increase the place selections in the frequency reports from 3 to 6. (6) Allow sorting node by node size as well as content. (7) Allow sort on selected contiguous parts of the node. (8) Add “more than” (>) and “less than” (<) as search operators for words tagged as CD to allow a search on (or to eliminate) years.

Appendix J: Tag Set Summary

Tag Description Examples

$ dollar Seen only: $ -$ $A Z$
`` opening quote mark ` ``
" closing quote mark ' "
(opening parenthesis ({
) closing parenthesis) }
, comma ,
: colon or ellipsis : ; ... - --
/ Slash /
SENT sentence terminator . ! ?

CC conjunction, coord and & but or both nor either plus neither yet versus vs. v. et minus less times whether so 'n
CD numeral, cardinal
DT determiner the a an this that some no all any those these another each every both either neither half many
EX existential there there
FW foreign word

IN preposition/sub-conj aboard about above across afore after against albeit along[side] although amid[st] around as at atop because before behind below beneath beside[s] between betwixt beyond by 'cause circa 'cos cum despite down during en except for from if in inside into lest like minus near[est] next notwithstanding of on off once onto opposite out[side] over past per plus re respecting sans since so **than** though thro through[out] thru 'til till toward[s] under[neath] unless unlike until unto up[on] versus via vis-à-vis vs. whereas whereupon whether while whither with[in] without

IN/that that -- most or all are nominalizer

JJ adjective or ordinal #
JJR adjective, comparative
JJS adjective, superlative

LS list item marker

Note: this tag summary is not an official document. It is based on our experience with the documentation and the data.

MD modal auxiliary ca[n't] can cannot could 'd 'll may might must need ought shall should will wo[n't] would NOTE: the [n't] are separate words!!!

NN noun, common, singular or mass
NNS noun, common, plural
NP noun, proper, singular
NPS noun, proper, plural

PDT pre-determiner: all such half quite nary
POS genitive marker ' 's
PP pronoun, personal: 'em he her[s[elf]] him[self] I it[self] me mine myself one[self] our[selves] she theirs them[selves] they us we you[rself]
PP$ pronoun, possessive: his their her its my our your[s]

RB adverb
RBR adverb, comparative
RBS adverb, superlative

RP particle up out off down over around on away through along in aside upon open apart unto whole

SYM symbol *] [/ + = @ _ > | ~ \ < and some single letters
TO "to" as preposition or infinitive marker

UH interjection

In verbs, the x stands for B (to be), H (to have) or V (all others).

Vx verb, base form
VxD verb, past tense
VxG verb, present participle or gerund
VxN verb, past participle
VxP verb, present, not 3rd person single
VxZ verb, present, 3rd person singular

WDT WH-determiner: that what whatever which whichever
WP WH-pronoun what who[m][ever]
WP$ WH-pronoun, possessive whose
WRB Wh-adverb: how[ever] when[ever] where[by] wherever why