Lesson 4 - Distinctive Collexeme Analysis

Erin M. Buchanan

01/31/2019

Language Topics Discussed

- Grammar versus the Lexicon
- Constructions and collexemes
- Examples

Grammar versus Lexicon

- Generally, the lexicon is considered the mental dictionary with understandings of meaning, phonetics, orthographics
- While, grammar was viewed as the abstract syntactic rules.
- Construction based approaches, such as Pattern Grammar and Lexical Chunks suggest an integration of lexicon and grammar, implying statistical language that does not distinguish the two.

Constructions and Collexemes

- Collostructional methods: investigating the interaction of words and constructions or understanding the structure of things together
- Constructions: construction-based approach to language
  - Grammar consists of form-meaning pairs, which is not different from the semantic lexicon
- You can apply these methods to all levels of language - words, phrases, tense

Constructions and Collexemes

- These methods help convert linguistic units into numbers, and therefore, allow you to apply statistical tests
- Similar in nature to the association measures discussed in the last chapter
- Objective way of identifying meaning of grammatical construction (what does *into the night* mean?)

Constructions and Collexemes

- What restrictions does grammar create for what might go next? (i.e., are there slots to fill in the grammar and what lexemes fill those slots?)
- "A word may occur in a construction if it is semantically compatible with the meaning of the construction (or, more precisely, with the meaning assigned by the construction to the particular slot in which the word appears)"

Types of Collexeme Analyses

- Collexeme analysis: measures the degree of attraction/repulsion of a word in a construction
- Distinctive Collexeme analysis: measures the preference of one word over another in that particular construction
  - Multiple DCA: expands to more constructions (i.e. not one just versus another)
- Covarying Collexeme analysis: measures the attraction of a word in one slot of a construction to words in another slot of the same construction

Examples

- Therefore, this method uses co-occurrence frequencies to show preference of construction combinations
- Wulff (2006) searched the British National Corpus for:
  - go-and-V: Now, just keep polishing those glasses while I go and *check* the drinks.
  - go-V: Go *find* the books and show me.
- Findings:
  - The verbs of each construction are not a subset of each other (therefore, not synonymous, and different semantically).

Examples

go-and-V
(92)

collect, live, visit, talk, watch, ask, sort, wash, hide, stand, stay, knock, eat, spoil, lay, tidy, feed, babysit, powder, pee, change, lock, baste, socialize, regurgitate, re-clean, re-credit, book, rouse, milk, lie, nick, vandalize, clean

Examples

- Wulff et al. (2007): a similar analysis with American and British English
  - into (negative): He blackmailed me into doing it.
  - into (persuasive): She talked me into doing it.
- This example will focus on American and British English versions of *quite*-ADJ combinations.
  - This restaurant is quite good.
  - The results is quite extraordinary.

Understanding Quite

- Quite can operate in several ways (British):
  - Maximizer: usually paired with limit (sure, clear) and extreme adjectives (huge, astounding) - aking to increasing the adjective.
  - Moderator: usually paired with scalar adjectives like good, nice, interesting - akin to using it as a rather/fairly.

Understanding Quite

- However, in American English, we use to use quite to maximize scalar adjectives, similar to saying very or extremely (good).
- Often not used with extreme adjectives (maybe instead we use really?)
- Hypothesis: American quite constructions will include less extreme adjectives than British English.
- Hypothesis: More limit adjectives with quite in American English because it was around before we split from Britian.

Analysis

- Pulled data from Corpus of Global Web Based English (GloWbE)
- Geographic differences of English in 20 countries

```r
library(Rling)
data(quite_Am)
data(quite_Br)
```

Data

```r
head(quite_Am)
```

```
##         Adj  AmE
## 1 DIFFERENT 1872
## 2      SURE 1492
## 3     CLEAR  938
## 4      GOOD  901
## 5  POSSIBLE  791
## 6    SIMPLE  599
```

Data

```r
head(quite_Br)
```

```
##         Adj  BrE
## 1 DIFFERENT 2313
## 2      SURE 1916
## 3     HAPPY 1710
## 4      GOOD 1614
## 5     CLEAR 1470
## 6     RIGHT 1162
```

Basic Differences

```
nrow(quite_Br)
```

```
## [1] 3702
```

```
nrow(quite_Am)
```

```
## [1] 3049
```

```
sum(quite_Br$BrE)
```

```
## [1] 61722
```

```
sum(quite_Am$AmE)
```

```
## [1] 37699
```

Simple DCA

- Use a 2X2 table like last week

	Construction A	Construction B
Collexeme X	a	b
X all other collexeme	c	d

Merge the data

- Since we aren't entering the data ourselves, let's merge these two datasets

```
quite = merge(quite_Br, quite_Am, #two datasets to merge
              by = "Adj", #how to match them
              all = TRUE) #return rows that don't match
head(quite)
```

```
##             Adj BrE AmE
## 1       ABASHED   1  NA
## 2 ABBREVIATED   1   1
## 3          ABLE  91  46
## 4      ABNORMAL   2   2
## 5    ABOMINABLE   1  NA
## 6      ABRASIVE   6   3
```

Clean up the data

```
quite[is.na(quite)] = 0
head(quite)
```

```
##             Adj BrE AmE
## 1       ABASHED   1   0
## 2 ABBREVIATED   1   1
## 3          ABLE  91  46
## 4      ABNORMAL   2   2
## 5    ABOMINABLE   1   0
## 6      ABRASIVE   6   3
```

Summarize the data

```r
a = quite$BrE #all quite to adj British constructions
b = quite$AmE #all quite to adj American constructions
c = sum(quite$BrE) - quite$BrE #overall all other combinat
d = sum(quite$AmE) - quite$AmE #overa0l all other combinat
head(cbind(as.character(quite$Adj), a, b, c, d))
```

```
##                      a    b    c       d
## [1,] "ABASHED"       "1"  "0"  "61721" "37699"
## [2,] "ABBREVIATED"   "1"  "1"  "61721" "37698"
## [3,] "ABLE"          "91" "46" "61631" "37653"
## [4,] "ABNORMAL"      "2"  "2"  "61720" "37697"
## [5,] "ABOMINABLE"    "1"  "0"  "61721" "37699"
## [6,] "ABRASIVE"      "6"  "3"  "61716" "37696"
```

Use an Association Measure

```r
#Calculated expected value of A
#Given row and column and sum totals, what should we expec
aExp = (a + b)*(a + c) / (a + b + c + d)
head(cbind(as.character(quite$Adj), a, aExp, b, c, d))
```

```
##                    a    aExp                b    c
## [1,] "ABASHED"     "1"  "0.620814516047917" "0"  "61721
## [2,] "ABBREVIATED" "1"  "1.24162903209583"  "1"  "61721
## [3,] "ABLE"        "91" "85.0515886985647"  "46" "61631
## [4,] "ABNORMAL"    "2"  "2.48325806419167"  "2"  "61720
## [5,] "ABOMINABLE"  "1"  "0.620814516047917" "0"  "61721
## [6,] "ABRASIVE"    "6"  "5.5873306443126"   "3"  "61716
```

logPF

```
#Calculate a chi-square for every combination
pvF = pv.Fisher.collostr(a, b, c, d)
#Convert to effect size measure
logpvF = ifelse(a < aExp, log10(pvF), -log10(pvF))
head(cbind(as.character(quite$Adj), a, aExp, b, c, d, logpv
```

```
##                      a    aExp               b    c
## [1,] "ABASHED"      "1"  "0.620814516047917" "0"  "61721"
## [2,] "ABBREVIATED" "1"  "1.24162903209583"  "1"  "61721"
## [3,] "ABLE"         "91" "85.0515886985647"  "46" "61631"
## [4,] "ABNORMAL"     "2"  "2.48325806419167"  "2"  "61720"
## [5,] "ABOMINABLE"  "1"  "0.620814516047917" "0"  "61721"
## [6,] "ABRASIVE"     "6"  "5.58733064443126"  "3"  "61716"
##      logpvF
## [1,] "0"
## [2,] "-4.82163733276644e-17"
## [3,] "0.478157443860978"
## [4,] "-0.195801176525502"
## [5,] "0"
```

Interpreting the numbers

- Higher positive scores indicate that the collexeme is represented more in British English (attraction to BE)

```
quite$logp = logpvF
quite = quite[ order(-quite$logp), ]
topBE = quite$Adj[1:20]
head(quite)
```

```
##                Adj  BrE AmE      logp
## 1424         HAPPY 1710 545 44.054249
## 1426          HARD  659 160 29.375670
## 1111 EXTRAORDINARY  217  46 12.160975
## 281            BIG  224  53 10.761473
## 2586       RELAXED   82   7  9.762904
## 698       DAUNTING  103  17  7.869958
```

Interpreting the numbers

- Higher negative scores indicate that the collexeme is represented more in American English (repulsion to BE)

```
quite = quite[ order(quite$logp), ]
topAE = quite$Adj[1:20]
head(quite)
```

```
##            Adj  BrE  AmE      logp
## 413    CERTAIN  175  281 -23.78763
## 2390  POSSIBLE  791  791 -22.02884
## 793  DIFFERENT 2313 1872 -19.42123
## 1131  FAMILIAR   87  168 -18.76324
## 228      AWARE   97  173 -17.41609
## 3070      SURE 1916 1492 -11.91646
```

Examine the Tops

- BE: Scalar (big, nice, difficult); extreme (daunting, staggering, incredible); limit (right, prepared)
- AE: Limit (certain, possible different, aware)
- BE more negative than AE

```
as.character(topBE)
```

```
##  [1] "HAPPY"          "HARD"           "EXTRAORDINARY" "BI
##  [5] "RELAXED"        "DAUNTING"       "DIFFICULT"     "QU
##  [9] "QUIET"          "RIGHT"          "STAGGERING"    "KI
## [13] "NICE"           "EMOTIONAL"      "EXCITING"      "TI
## [17] "WORRYING"       "INCREDIBLE"     "PREPARED"      "LU
```

```
as.character(topAE)
```

```
##  [1] "CERTAIN"    "POSSIBLE"   "DIFFERENT"  "FAMILIAR"
##  [6] "SURE"       "VALUABLE"   "REAL"       "SUCCESSFUL'
## [11] "FOND"       "SIMILAR"    "EFFECTIVE"  "SKEPTICAL"
## [16] "TASTY"      "WILLING"    "HELPFUL"    "SOMETIME"
```

Summary

- We can extend the association measures learned last week to collexeme analysis, where the focus is specifically on grammatical slot differences.
- These analyses can show cultural or structural differences in a language.
- Applied use of chi-square analysis and interpretation of the qualitative results implies that grammar and semanticity are interwoven.