CHAPTER 11

Geographic variation of *quite*

Distinctive collexeme analysis

What you will learn from this chapter:

This chapter introduces distinctive collexeme analysis, which employs bidirectional association measures discussed in the previous chapter. This method is based on the co-occurrence frequencies of words that occur in two near-synonymous constructions, or in two or more dialectal or diachronic variants of the same construction. Here we will compare the variants of *quite* + ADJ construction in different national varieties of English. We will first present a canonical distinctive collexeme analysis with only two varieties, British and American English, and then will show how this approach can be extended to more lects, presenting a unified approach to multiple distinctive collexeme analysis.

11.1 Introduction to distinctive collexeme analysis

The method of distinctive collexeme analysis (Gries & Stefanowitsch 2004; Wulff 2006; Wulff et al. 2007), henceforth DCA, belongs to the family of corpus-based collostructional methods developed by Stefan Th. Gries and Anatol Stefanowitsch (Stefanowitsch & Gries 2003 and later works). DCA is a technique designed specifically to compare two or more constructions by finding the distinctive slot fillers (collexemes) that are significantly attracted to one construction and repelled by the other. One can examine two (or more) near-synonymous constructions in one variety, e.g. *go*-V vs. *go-and*-V in Wulff (2006) or English analytic causatives (Gilquin 2006). Another option, which is demonstrated here, is comparison of formally identical constructions in two varieties, e.g. the *into*-causative in British and American English (Wulff et al. 2007). The result of the statistical procedure is two lists of distinctive collexemes, which are typical of each lectal variant. For instance, Wulff et al. (2007) found out that many verbs that are distinctive of the British variant of the *into*-causative designate negative emotions (e.g. *She terrified me into doing it*) and threatening (*He blackmailed me into doing it*). These semantic classes are not typical of collexemes in the American English variant of the same construction. On the other hand, many distinctive American collexemes refer to communication (e.g. *She talked me into doing it*). Wulff et al. (2007) hypothesize that these and other differences may reflect the

varying degrees of prominence of specific semantic scenarios in the two cultures. These scenarios are expressed in the title of their paper, 'Brutal Brits and Persuasive Americans'.

A similar analysis was carried out in Levshina at al. (2011), where the Netherlandic and Belgian variants of the causative construction with *doen* were compared. An inspection of the constructional slots reveals that the Netherlandic causative *doen* seems to be specialized in the so-called affective causation, i.e. situations when a stimulus produces a mental reaction, e.g. *Zijn kapsel doet me denken aan een vogelnest* 'His hairstyle reminds me of (lit. makes me think) of a bird's nest'. Since Netherlandic Dutch is considered to be the leader in language change, and Belgian Dutch is believed to be more conservative, the more limited semantic repertoire of Netherlandic *doen* ties in well with the previous observations of the ongoing qualitative and quantitative shrinking of the auxiliary *doen* in Dutch (e.g. Speelman & Geeraerts 2009).

There exists special software (Gries 2004) that can help you perform this and other types of collostructional analysis. The aim of this chapter, however, is to demonstrate the basic principles of DCA in a sequence of simple steps. In DCA it is common to apply a significance test (traditionally, the Fisher exact test is used) to find out which collexemes are over- and underrepresented in each variety at a statistically significant level. The log-transformed *p*-value is frequently used as a measure of distinctiveness. The results are symmetric: if one collexeme is significantly overrepresented in Construction/Variety A, then it is significantly underrepresented in Construction/Variety B, and vice versa. The next section will discuss simple DCA, whereas a unified approach to multiple DCA will be introduced in Section 11.3.

11.2 Distinctive collexeme analysis of *quite* + ADJ in different varieties of English: A unified approach

11.2.1 Theoretical background and data

To reproduce the code in this case study, you will need two datasets from the add-on companion package `Rling`.

```
> library(Rling)
```

The method will be illustrated with a case study of the construction *quite* + ADJ in British and American English. A few examples of the construction are below:

(1) a. *The restaurant is quite good*.
b. *All art is quite useless.* (Oscar Wilde)
c. *The result is quite extraordinary*.

Quite has complex semantics. It can function both as a maximizer ('entirely'), and a moderator, similar to *fairly* or *rather*. Often, these functions can be disambiguated only in

context. However, there are also distributional cues, most importantly, semantic classes of modified adjectives (Paradis 1997). With scalar adjectives (e.g. *good*, *nice*, *interesting*, *difficult*, *rich*), as in (1a), *quite* can function as a moderator. With limit adjectives, which imply a clear boundary (e.g. *useless, sure*, *cooked*, *clear*, *different*, *wrong*, *dead*), and extreme adjectives, which describe a high degree of some quality (e.g. *extraordinary*, *huge*, *scorching*, *marvellous*, *astounding*), *quite* tends to be a maximizer, as in (1b) and (1c).

These properties are characteristic of contemporary British English. However, there is evidence that the moderating function emerged not earlier than the 18th century. Thus, there are reasons to expect that this function would not be so prominent in American English. In fact, some lexicographic sources observe that this is indeed so. First, *quite* is not used as a moderator in American English. Instead, American *quite* serves as a booster with scalar adjectives, similar to *very* or *extremely*. The sentence (1a) could therefore only mean in American English that the restaurant is very good. Second, American *quite* is not used with extreme adjectives.

On the basis of the second observation, we can formulate the hypothesis that American *quite* will contain fewer adjectives with extreme semantics than its British counterpart. Another prediction can be made based on the history of *quite* as a degree modifier: since the maximizer function with limit adjectives seems to be the only one fully developed before English was exported to America, we will expect more limit adjectives among the collexemes of American *quite*.

To test the hypotheses, we will use adjectives that occur immediately after *quite* in the Corpus of Global Web-based English (GloWbE), which represents geographic varieties of English in twenty countries (Davies 2013). The lists can be found in the `quite_Br` and `quite_Am` datasets.

```
> data(quite_Br)
> data(quite_Am)
> head(quite_Br)
  Adj        BrE
1 DIFFERENT  2313
2 SURE       1916
3 HAPPY      1710
4 GOOD       1614
5 CLEAR      1470
6 RIGHT      1162
```

The datasets have similar structure. The rows correspond to the adjectives (lemmas). The column `BrE` displays the frequencies in the British variant of the construction, and `AmE` in the dataset `quite_Am` corresponds to the American data. The number of unique adjectives is different: the British dataset contains 3702 collexemes, and its American counterpart has 3046 adjectives.

```
> nrow(quite_Br)
[1] 3702
> nrow(quite_Am)
[1] 3049
```

The frequencies of the constructional variants differ, as well: the British variant has almost twice as many instances as the American one (61722 vs. 37699), although the subcorpora are of almost equal size:

```
> sum(quite_Br$BrE)
[1] 61722
> sum(quite_Am$AmE)
[1] 37699
```

Are there any qualitative differences between the variants? This is a question for simple DCA, which will be performed in the next subsection. After that, the Canadian variant of the construction will be added to illustrate a unified approach to multiple DCA.

11.2.2 Simple distinctive collexeme analysis of *quite* + ADJ in British and American English

To perform the analysis, one needs a 2-by-2 table of co-occurrence frequencies shown in Table 11.1 (cf. Table 10.1 in the previous chapter).

Table 11.1 Frequencies required for simple distinctive collexeme analysis

	Construction/Variety A	**Construction/Variety B**
Collexeme *X*	*a*	*b*
¬ *X* all other collexemes	*c*	*d*

To perform the analysis, one needs to create the vectors that contain the frequencies *a*, *b*, *c* and *d* for each adjective, following the approach introduced in Chapter 10. The first step is to create a data frame which will contain all adjectives and will have two columns, with the British and American frequencies:

```
> quite <- merge(quite_Br, quite_Am, by = "Adj", all = TRUE)
> 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
```

The 'NA' values in the table show that the corresponding adjective is missing in one of the lists. The missing values should be replaced with zeroes:

```
> quite[is.na(quite)] <- 0
```

The next step is to obtain four frequencies from Table 11.1. The columns in `quite` are in fact vectors with the frequencies *a* and *b*. The vector with frequencies *c* can be obtained as the sum of all instances of the construction in the British data minus *a*, and the vector with *d* is the sum of all instances of the American variant of *quite* + ADJ minus *b*:

```
> a <- quite$BrE
> b <- quite$AmE
> c <- sum(quite$BrE) - quite$BrE
> d <- sum(quite$AmE) - quite$AmE
```

You will also need the expected frequency in the British variety:

```
> aExp <- (a + b)*(a + c)/(a + b + c + d)
```

Now you are ready to run the statistical tests and compare the geographical variants of *quite*. Since the relationships between the varieties are symmetric, it suffices to compute attraction/repulsion scores for British English. You can use the `pv.fisher.collostr()` function, which was presented in the previous chapter, to compute the Fisher exact test *p*-values for all adjectives. This function is available in the `Rling` package.

```
> pvF <- pv.Fisher.collostr(a, b, c, d)
```

For better interpretability, it is common to take a negative logarithm with base 10 of the *p*-values. It is also necessary to add information about the direction of association, i.e. whether a given adjective is overrepresented or underrepresented in the British variant of the construction. If the observed frequency in the British data is smaller than the expected frequency, then the log-transformed score will remain negative. If the observed frequency in the British construction is greater than the expected frequency, the log-transformed score will become positive (see more details in Chapter 10):

```
> logpvF <- ifelse(a < aExp, log10(pvF), -log10(pvF))
```

The greater the `logpvF` score, the more overrepresented the collexeme in the British data and the more underrepresented in the American data. Now the scores can be added to the data as a new column, and the data frame can be sorted according to the scores. The function `order()` can be used to sort a data frame in ascending and descending order. In the latter case, one can use the minus sign before the name of the variable by which the data are sorted. The collexemes that are the most distinctive of the British variant can be found at the top of the list.

```
> quite$logp <- logpvF
> quite <- quite[order(-quite$logp),]
```

```
> quite[1:20,]
      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
798   DIFFICULT       848   366  7.836994
2516  QUICK           87    12   7.782684
2517  QUIET           61    5    7.543155
2651  RIGHT           1162  537  7.309045
2969  STAGGERING      67    7    7.308632
1808  KEEN            105   20   6.856004
2115  NICE            557   229  6.570673
991   EMOTIONAL       119   26   6.545695
1073  EXCITING        161   44   6.280060
3236  TRICKY          146   38   6.228315
3689  WORRYING        56    6    5.969637
1630  INCREDIBLE      119   29   5.633430
2412  PREPARED        166   49   5.546736
1922  LUCKY           132   35   5.523337
```

To obtain the top twenty collexemes that are distinctive of the American variant, one can simply remove the minus sign before the name of the variable, so that the scores are ordered from the smallest to the largest one:

```
> quite <- quite[order(quite$logp),]
> quite[1:20,]
      Adj           BrE   AmE   logp
413   CERTAIN       175   281   -23.787625
2390  POSSIBLE      791   791   -22.028837
793   DIFFERENT     2313  1872  -19.421232
1131  FAMILIAR      87    168   -18.763244
228   AWARE         97    173   -17.416095
3070  SURE          1916  1492  -11.916458
3558  VALUABLE      23    64    -10.727512
2547  REAL          52    97    -10.615434
3046  SUCCESSFUL    124   161   -9.558394
2546  READY         301   307   -9.473267
1238  FOND          65    104   -9.112043
2837  SIMILAR       420   393   -8.831172
958   EFFECTIVE     108   140   -8.431429
```

```
2851  SKEPTICAL     8     36    -8.415016
26    ACCURATE      113   143   -8.179308
3119  TASTY         20    51    -8.025411
3665  WILLING       123   149   -7.721011
1465  HELPFUL       94    122   -7.347363
2913  SOMETIME      47    74    -6.580660
3969  FAVORABLE     0     15    -6.317976
```

It is a popular practice in distinctive collexeme analysis to classify all distinctive collexemes (i.e. those whose absolute scores are above the cut-off point) into some semantic classes and then sum up the scores for the classes. Because of space limitations, we will only look at the top twenty collexemes in each variety.

An informal inspection of the top collexemes seems to support the theoretical expectations. First, about a half of the top twenty collexemes in the British list can be considered scalar (e.g. *big*, *nice*, *difficult*). The exceptions are extreme adjectives *extraordinary*, *daunting*, *staggering* and *incredible*, and limit adjectives *right* and *prepared*. In contrast, the American top twenty list contains mostly limit adjectives (*certain*, *possible*, *different*, *aware*, *sure*, etc.),[1] and no extreme adjectives.

It is also surprising that the American top collexemes are in general more positive than the British ones, which may have negative connotations (*difficult*, *hard*, *tricky*, *worrying*). This observation will also hold if the lists beyond the top 20 limit are inspected. This finding suggests interesting differences in the semantic prosody of *quite*.

The absolute standard cut-off value, which corresponds to the *p*-value of 0.05 is approximately 1.3 for log-transformed values (with 10 as the logarithm base). Only 182 adjectives from the entire list are distinctive of the British variant, and 185 are distinctive of the American variant at the significance level 0.05.

```
> nrow(quite[quite$logp > 1.3,])
[1] 182
> nrow(quite[quite$logp < (-1.3),])
[1] 185
```

The overwhelming majority of the collexemes are therefore not distinctive of either variety at the significance level of 0.05. Most of them are simply not sufficiently frequent, but some of them are nearly equally distributed in the two varieties.

To summarize, British *quite* seems to be more frequently used with scalar adjectives where the attenuation function is the most natural (*quite nice*, *quite big*). It also tends to attract extreme adjectives (*quite extraordinary*) more strongly than its American

1. The word *sometime* ended up in the list due to a parsing error in the corpus.

counterpart, which, in its turn, more frequently functions as an intensifier of non-gradable limit adjectives (*quite sure*, *quite different*). This means that the British modifier is more polysemous than American *quite*, which has to do with the relatively recent development of the new functions of *quite* in addition to its oldest meaning 'absolutely, entirely'. These changes happened in British English in the second part of the 18th–19th centuries, after English was exported to America (see Levshina 2014 for more information).

11.2.3 Multiple distinctive collexeme analysis: *Quite* + ADJ in the British, American and Canadian varieties of English

This subsection shows how to extend simple DCA to cases with more than two varieties or near-synonymous constructions. This time, the Canadian variant of *quite* + ADJ will be added. Canadian English is known to have retained some properties of British English, but at the same time it has been strongly influenced by American English as a part of North American English. The list of adjectives from the GloWbE corpus can be found in the dataset `quite_Ca` in `Rling`. To compare all three lists, one should add the adjectives from `quite_Ca` to the list of the British and American collexemes stored in the data frame `quite`. To do so, we will use again the `merge()` function. The `logp` column from `quite` can be discarded because the old distinctiveness scores are no longer needed:

```
> data(quite_Ca)
> quite1 <- merge(quite[, -4], quite_Ca, by = "Adj", all = TRUE)
> quite1[is.na(quite1)] <- 0
> str(quite1)
'data.frame': 4856 obs. of 4 variables:
$ Adj: Factor w/ 4856 levels "ABASHED","ABBREVIATED",..: 1 2 3 4 5
6 7 8 9 10 …
$ BrE: num 1 1 91 2 1 6 1 7 4 1 …
$ AmE: num 0 1 46 2 0 3 1 6 1 0 …
$ CE: num 0 0 17 3 1 0 0 1 2 0 …
```

The main principle of multiple DCA is the same as in the simple DCA. One compares the proportions of every collexeme in different constructions or variants of the same construction. The main difference is that multiple DCA, as implemented in our unified approach, compares each construction/variant against all others.

This case study will focus only on the distinctive collexemes of the Canadian *quite* + ADJ construction against the British and American variants taken together. We will need the frequencies shown in Table 11.2. Variety *A* in the table is then Canadian English, and the 'other' varieties are British and American.

Table 11.2 Frequencies required for multiple distinctive collexeme analysis

	Construction/Variety A	**Other constructions/varieties**
Collexeme *X*	*a*	*b*
¬ *X* (all other collexemes)	*c*	*d*

To obtain the frequencies, one can use a procedure that is very similar to the one described in the previous subsection. The only difference is that one should sum up the frequencies of collexemes in both British and American data to obtain the count *b*

```
> a <- quite1$CE
> b <- quite1$BrE + quite1$AmE
> c <- sum(a) - a
> d <- sum(b) - b
```

You will also need the expected frequency of every collexeme in Canadian English.

```
> aExp <-(a + b)*(a + c)/(a + b + c + d)
```

Gries (2004) uses the binomial test to compute distinctiveness scores for multiple distinctive collexeme analysis, but we will use the Fisher exact test, as before, for the sake of consistency, presenting a unified approach. The correlation between the original Gries' method scores and ours is nearly perfect ($r = 0.98$). The R code is the same as above:

```
> pvF <- pv.Fisher.collostr(a, b, c, d)
> logpvF <- ifelse(a < aExp, log10(pvF), -log10(pvF))
> quite1$logpCE <- logpvF
```

The top twenty most distinctive collexemes in the Canadian variant of the construction can be obtained as follows:

```
> quite1 <- quite1[order(-quite1$logpCE),]
> quite1[1:20,]
      Adj               BrE   AmE   CE   logpCE
1833  LARGE             334   247   128  4.894402
870   DISTINGUISHABLE   1     0     6    4.646948
4642  GOOD-NATURED      0     0     5    4.536043
374   BUSY              134   75    57   4.468647
2546  READY             301   307   128  4.095098
793   DIFFERENT         2313  1872  690  3.802790
1061  EVIDENT           94    101   51   3.688441
638   CRAPPY            2     1     6    3.664810
1592  IMPRESSIVE        198   141   78   3.616966
```

```
978   ELITE            2     0     5    3.308234
3119  TASTY            20    51    24   3.298938
508   COMFORTABLE      178   175   78   3.254352
1909  LOW              328   202   109  3.240176
1744  INTRIGUED        28    10    16   3.215085
415   CHALLENGING      111   75    47   3.176938
2837  SIMILAR          420   393   155  3.071796
1186  FILLING          4     4     7    2.937432
816   DISAPPOINTED     88    60    38   2.770482
4535  APPRECIATIVE     0     0     3    2.721544
4550  BOOKISH          0     0     3    2.721544
```

The list bears more resemblance to the top distinctive American collexemes than to the British ones, containing *different* (and its synonym *distinguishable*), *similar*, *ready* and *tasty*. We can also examine the collexemes that are significantly underrepresented in the Canadian variant:

```
> quite1 <- quite1[order(quite1$logpCE),]
> quite1[1:20,]
      Adj          BrE   AmE   CE   logpCA
1424  HAPPY        1710  545   208  -9.422766
2651  RIGHT        1162  537   151  -8.223901
1426  HARD         659   160   62   -6.595266
2724  SCARY        175   58    8    -5.756694
348   BRILLIANT    131   39    6    -4.297981
281   BIG          224   53    15   -4.277109
3694  WRONG        282   103   26   -4.186445
3236  TRICKY       146   38    9    -3.380667
1922  LUCKY        132   35    8    -3.117908
1304  FUNNY        413   179   53   -3.025242
132   ANNOYING     122   75    11   -3.006459
3689  WORRYING     56    6     1    -2.428485
617   CORRECT      244   220   42   -2.357467
2920  SORRY        31    15    0    -2.271914
1283  FRIGHTENING  82    27    5    -2.197128
1343  GLAD         74    36    5    -2.190326
3070  SURE         1916  1492  421  -2.121099
3062  SUPERB       37    4     0    -2.119141
117   AMUSING      165   86    20   -2.015827
2179  ODD          104   62    11   -1.943230
```

Some of the repelled collexemes were in the British top list (*happy*, *right*, *hard*, *big, tricky*, *lucky*, *worrying*). Only one of them (*sure*) was ranked high on the American list. There are also extreme adjectives, which are typical of the British construction: *superb* and *brilliant*.

So, Canadian *quite* + ADJ seems to be more similar to the American variant than to the British one. This is not surprising, considering the history of the North American varieties and the diachronic development of *quite* (see the previous subsection). To obtain the distinctive collexemes for the British and American varieties, one can repeat the procedure by comparing each variety against the rest.

11.3 Summary

This chapter has introduced the basic principles of distinctive collexeme analysis. They were illustrated by a case study of *quite* as a pre-adjectival modifier in British, American and Canadian English. The approach is based on the computation of association measures for 2-by-2 tables, which were introduced in the previous chapter. Although log-transformed Fisher exact test *p*-values are usually computed due to their robustness to low frequencies, in principle, any bidirectional measure that was mentioned in Chapter 10 can be used. Since the *p*-value is a hypothesis-testing statistic, which is influenced by the overall sample size, it may be more appropriate to use a measure of effect size, such as the odds ratio, especially if one wants to compare results based on corpora of different sizes. From this also follows that the length of a list of distinctive collexemes depends on the sample size.

How to write up the results of distinctive collexeme analysis

To report the results of DCA, you can make a table with top distinctive collexemes (or all, if the space permits) for each constructional or lectal variant and provide their distinctiveness scores. It is also recommended that you classify the distinctive collexemes into several theoretically interpretable classes, and add up their distinctiveness scores. These scores can then be compared across the varieties and presented in a table.