Proof of Concept

CorporaCoCo v1.0-2 (2017-03-31)

Anthony Hennessey
Statistics and Probability, School of Mathematical Sciences
University of Nottingham
anthony.hennessey@nottingham.ac.uk

Viola Wiegand
Centre for Corpus Research, College of Arts and Law
University of Birmingham
v.wiegand@bham.ac.uk

Michaela Mahlberg
Centre for Corpus Research, College of Arts and Law
University of Birmingham
m.a.mahlberg@bham.ac.uk

Christopher R. Tench
Division of Clinical Neurosciences, School of Medicine
University of Nottingham
christopher.tench@nottingham.ac.uk

Jamie Lentin
Shuttle Thread
Manchester
jm@ravingmantis.com

Load the `CorporaCoCo` package.

```
library(CorporaCoCo)
```

Create tokenized copies of 'Great Expectations' and 'A Tale of Two Cities' novels. The texts are available in the CorporaCorpus package which is available from github at https://github.com/ravingmantis/CorporaCorpus, there are installation instructions on the front page. (The CorporaCorpus package is not available on CRAN as at 17MB it exceeds the CRAN data package size limit of 5MB). The tokenization we use here is very simplistic, but it will do for our purposes. The `stringi` package has a solid implementation of UTF-8 word boundaries so although this is simple tokenization it should do a reasonable job for text in any language.

```
library(CorporaCorpus)
library(stringi)
GE  <- unlist( stri_extract_all_words( stri_trans_tolower( readLines(corpus_filepaths('DNov', 'GE')) ) ))
TTC  <- unlist( stri_extract_all_words( stri_trans_tolower( readLines(corpus_filepaths('DNov', 'TTC')) ) ))
```

Choose the set of nodes.

```
nodes <- c('back', 'eye', 'eyes', 'forehead', 'hand', 'hands', 'head', 'shoulder')
```

First we want to check that there are no significant results under the null. We create two corpora from alternate chunks of 1000 tokens of the two novels and check that there are no significant co-occurrence differences between our two sets of chunks.

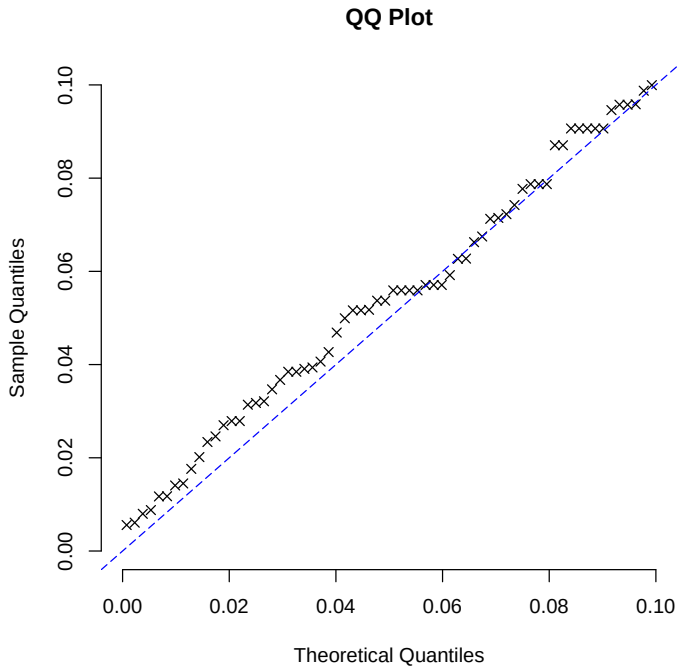
```
chunks <- split(c(GE, TTC), ceiling(seq_along(c(GE, TTC)) / 1000))
corpus_a <- unlist( chunks[seq(1, length(chunks), 2)] )
corpus_b <- unlist( chunks[seq(2, length(chunks), 2)] )
corpus_a_c <- surface(corpus_a, span = '5LR')
corpus_b_c <- surface(corpus_b, span = '5LR')
results <- coco(corpus_a_c, corpus_b_c, nodes = nodes, fdr = 0.01)
results
```

```
Empty data.table (0 rows) of 11 cols: x,y,H_A,M_A,H_B,M_B...
```

This gives us the opportunity to check an assumption of FDR that the p-values are uniformly distributed.

```
results_all <- coco(corpus_a_c, corpus_b_c, nodes = nodes, fdr = 1.0)
test_p_values <- results_all$p_value[results_all$p_value <= 0.1]
```

```
plot(
    qunif(ppoints(test_p_values), min = 0, max = 0.1),
    sort(test_p_values),
    bty = 'n', pch = 4, xlim = c(0.0, 0.1), ylim = c(0.0, 0.1),
    main = "QQ Plot", xlab = "Theoretical Quantiles", ylab = "Sample Quantiles"
)
abline(a = 0, b = 1, col = 'blue', lty = 5)
```



Next we check that if we make some changes to one of our corpora that the method can spot them. Let us change about 90% of the '*my*' tokens to '*CHIMERA*' tokens in `corpus_a` and comfirm that the method notices

```
corpus_a_mod <- corpus_a
mys <- which(corpus_a_mod == 'my')
corpus_a_mod[sample(mys, floor(length(mys)*0.9))] <- 'CHIMERA'
corpus_a_mod_c <- surface(corpus_a_mod, span = '5LR')
results <- coco(corpus_a_mod_c, corpus_b_c, nodes = nodes, fdr = 0.01)
results
```

```
           x       y H_A  M_A H_B  M_B effect_size CI_lower  CI_upper      p_value   p_adjusted
 1:     back CHIMERA  17 1758   0 1947        -Inf     -Inf -2.188961 3.278590e-06 3.170396e-03
 2:     eyes CHIMERA  29 1592   0 1776        -Inf     -Inf -3.035804 4.212795e-10 3.871558e-07
 3:     eyes      my   2 1619  32 1744    3.892194 1.917290  7.003039 2.078478e-07 9.550608e-05
 4:     hand CHIMERA  50 2556   0 2493        -Inf     -Inf -3.668557 2.331889e-15 2.672345e-12
 5:     hand      my   4 2602  43 2450    3.512370 2.049233  5.452421 7.534736e-10 4.317404e-07
 6:    hands CHIMERA  24 1387   0 1489        -Inf     -Inf -2.687507 2.799724e-08 2.264976e-05
 7:    hands      my   1 1410  24 1465    4.528933 1.908262  9.886466 1.690464e-06 6.837928e-04
 8:     head CHIMERA  30 2055   0 1937        -Inf     -Inf -2.845673 2.757629e-09 1.478089e-06
 9:     head      my   1 2084  40 1897    5.455854 2.892648 10.784870 5.260298e-12 5.639039e-09
10: shoulder CHIMERA  20  424   0  334        -Inf     -Inf -1.943756 1.207272e-05 3.283779e-03
```

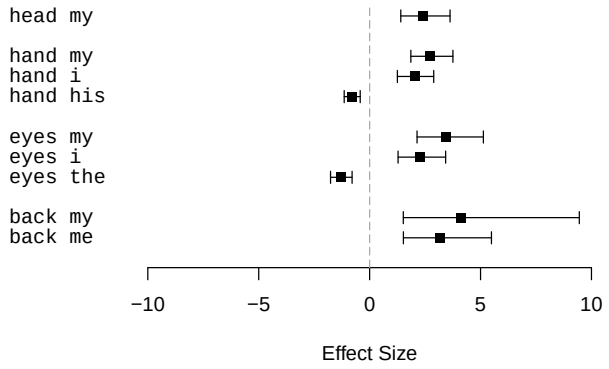
Next a more realistic example (and the reason we chose that set of `nodes`). Here we check that the results indicate the different narative voice, third and first person, used in the two novels; the body part nouns are expected to be found in suspensions (Mahlberg, 2013).

```
results <- surface_coco(TTC, GE, span = '5LR', nodes = nodes, fdr = 0.01)
results
```

```
      x   y H_A  M_A H_B  M_B effect_size  CI_lower   CI_upper      p_value   p_adjusted
1: back  me   3 1316  48 2355    3.159998  1.521928  5.4917238 9.754793e-07 9.423130e-04
2: back  my   1 1318  31 2372    4.105901  1.517363  9.4521419 1.987134e-05 9.597855e-03
3: eyes   i  10 1611  52 1724    2.280107  1.281850  3.4267531 2.247538e-07 6.869976e-05
4: eyes  my   5 1616  58 1718    3.446625  2.137003  5.1270592 1.061195e-11 9.731159e-09
5: eyes the 120 1501  57 1719   -1.269288 -1.761782 -0.7909003 4.323172e-08 1.982175e-05
6: hand his 175 2267 114 2543   -0.783898 -1.147324 -0.4250235 1.158348e-05 4.413307e-03
7: hand   i  17 2425  74 2583    2.030509  1.250655  2.8889719 7.519299e-09 4.297280e-06
8: hand  my  12 2430  85 2572    2.742060  1.858321  3.7535208 1.043073e-13 1.192232e-10
9: head  my   9 1732  62 2219    2.426331  1.404175  3.6251454 3.575486e-08 3.822194e-05
```

and plot of the results (TTC is on the left)

```
plot(results)
```

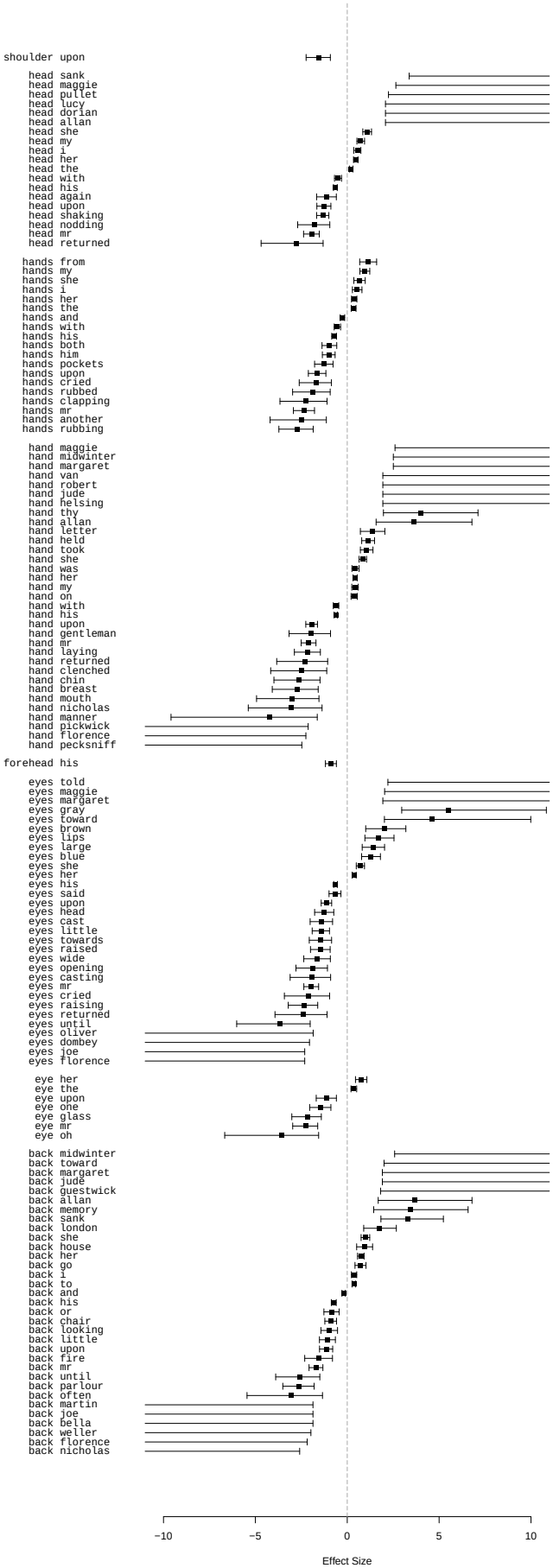


Finally we compare all of Dickens novels against a set of 19th century novels to check if we can reproduce the observations from Mahlberg (2013) that Dickens uses descriptions of body language more frequently that other authors of the time.

```
DICKENS <- unlist(stri_extract_all_words(stri_trans_tolower(do.call(c, lapply(corpus_filepaths('DNov'), readLines)))))
NCNB <- unlist(stri_extract_all_words(stri_trans_tolower(do.call(c, lapply(corpus_filepaths('19C'), readLines)))))
results <- surface_coco(DICKENS, NCNB, span = '5LR', nodes = nodes, fdr = 0.01)
```

Here is a plot of the results; Dickens is on the left.

```
plot(results)
```


shoulder upon
head sank
head maggie
head pullet
head lucy
head dorian
head allan
head she
head my
head i
head her
head the
head with
head his
head again
head upon
head shaking
head nodding
head mr
head returned
hands from
hands my
hands she
hands i
hands her
hands the
hands and
hands with
hands his
hands both
hands him
hands pockets
hands upon
hands cried
hands rubbed
hands clapping
hands mr
hands another
hands rubbing
hand maggie
hand midwinter
hand margaret
hand van
hand robert
hand jude
hand helsing
hand thy
hand allan
hand letter
hand held
hand took
hand she
hand was
hand her
hand my
hand on
hand with
hand his
hand upon
hand gentleman
hand mr
hand laying
hand returned
hand clenched
hand chin
hand breast
hand mouth
hand nicholas
hand manner
hand pickwick
hand florence
hand pecksniff
forehead his
eyes told
eyes maggie
eyes margaret
eyes gray
eyes toward
eyes brown
eyes lips
eyes large
eyes blue
eyes she
eyes her
eyes his
eyes said
eyes upon
eyes head
eyes cast
eyes little
eyes towards
eyes raised
eyes wide
eyes opening
eyes casting
eyes mr
eyes cried
eyes raising
eyes returned
eyes until
eyes oliver
eyes dombey
eyes joe
eyes florence
eye her
eye the
eye upon
eye one
eye glass
eye mr
eye oh
back midwinter
back toward
back margaret
back jude
back guestwick
back allan
back memory
back sank
back london
back she
back house
back her
back go
back i
back to
back and
back his
back or
back chair
back looking
back little
back upon
back fire
back mr
back until
back parlour
back often
back martin
back joe
back bella
back weller
back florence
back nicholas
−10
−5
0
5
10
Effect Size

References

Mahlberg, M. (2013). *Corpus stylistics and dickens's fiction*. Routledge.