

# Proof of Concept

## CorporaCoCo v1.1-0 (2017-11-22)

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  
jamie.lentin@shuttlethread.com

Fetch the ordered tokens for 'Great Expectations' and 'A Tale of Two Cities' novels from the CLiC API.

```
library(jsonlite)
get_book_tokens <- function(shortname) {
  base_uri <- 'http://clic.bham.ac.uk/api'
  json <- fromJSON(paste0(base_uri, "/subset?corpora=", shortname))
  tokens <- tolower( unlist( sapply(json$data, function(x) {
    head(x[[1]], -1)[as.integer(tail(x[[1]], 1)[[1]])+1]
  } ) ) )
}
GE <- get_book_tokens('GE')
TTC <- get_book_tokens('TTC')
```

Load the CorporaCoCo package.

```
library(CorporaCoCo)
```

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.

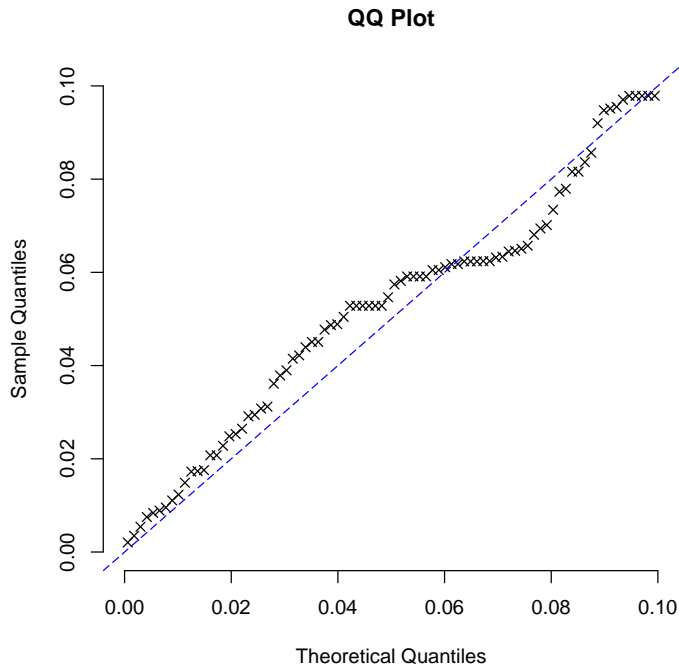
```
chunks <- split(c(GE, TTC), ceiling(seq_along(c(GE, TTC)) / 1000))
corpus_a <- unlist( chunks[seq(1, length(chunks), 2)] )
corpus_b <- unlist( chunks[seq(2, length(chunks), 2)] )
corpus_a_c <- surface(corpus_a, span = '5LR')
corpus_b_c <- surface(corpus_b, span = '5LR')
results <- coco(corpus_a_c, 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 confirm 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	my	0	1890	17	1823		Inf	2.093968	Inf	5.841370e-06	5.671970e-03
2:	eyes	CHIMERA	30	1620	0	1790		-Inf	-Inf	-3.074373	2.331613e-10	2.163737e-07
3:	eyes	my	1	1649	30	1760	4.812183	2.220141	10.156672	4.642279e-08	2.154017e-05	
4:	hand	CHIMERA	40	2500	0	2580		-Inf	-Inf	-3.414695	5.686176e-13	6.493613e-10
5:	hand	my	7	2533	51	2529	2.866900	1.717219	4.254416	2.109590e-09	1.204576e-06	
6:	hands	CHIMERA	22	1338	0	1590		-Inf	-Inf	-2.698213	3.645054e-08	3.021750e-05
7:	head	CHIMERA	40	1940	0	1970		-Inf	-Inf	-3.390409	1.497662e-12	1.572545e-09
8:	head	my	4	1976	27	1943	2.778609	1.253757	4.756674	1.727748e-05	9.070676e-03	
9:	shoulder	CHIMERA	16	354	0	420		-Inf	-Inf	-2.178354	4.495079e-06	1.240642e-03

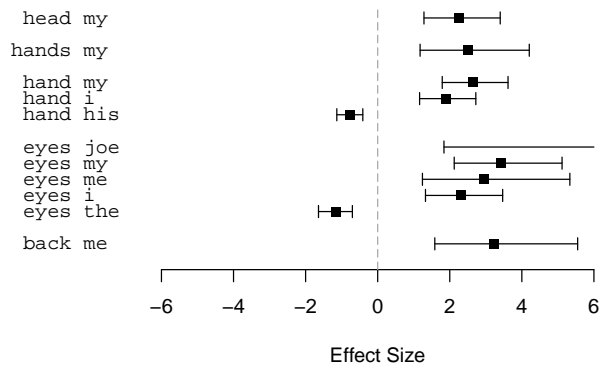
Next a more realistic example (and the reason we chose that set of nodes). Here we check that the results indicate the different narrative voice, third and first person, used in the two novels; the body part nouns are expected to be found in suspensions (Mahlberg, 2013).

```
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	1337	49	2341	3.221181	1.584866	5.5489805	5.440975e-07	5.283187e-04	
2:	eyes	i	10	1640	53	1737	2.322489	1.326370	3.4680980	1.290817e-07	5.963576e-05	
3:	eyes	joe	0	1650	16	1774		Inf	1.839353	Inf	3.552572e-05	6.691836e-03
4:	eyes	me	3	1647	25	1765	2.958423	1.241832	5.3326117	3.621123e-05	6.691836e-03	
5:	eyes	my	5	1645	57	1733	3.434699	2.123620	5.1159658	9.752564e-12	9.011369e-09	
6:	eyes	the	123	1527	62	1728	-1.166398	-1.642460	-0.7024399	2.098712e-07	6.464034e-05	
7:	hand	his	176	2294	114	2536	-0.771065	-1.133959	-0.4126876	1.250677e-05	4.744234e-03	
8:	hand	i	19	2451	75	2575	1.909259	1.162857	2.7232409	1.629910e-08	9.274188e-06	
9:	hand	my	13	2457	85	2565	2.646457	1.791317	3.6168202	1.860637e-13	2.117405e-10	
10:	hands	my	5	1125	45	1775	2.511311	1.177037	4.2063750	1.127123e-05	9.321308e-03	
11:	head	my	10	1710	61	2169	2.265311	1.284027	3.3998354	1.607393e-07	1.689370e-04	

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

```
plot(results)
```

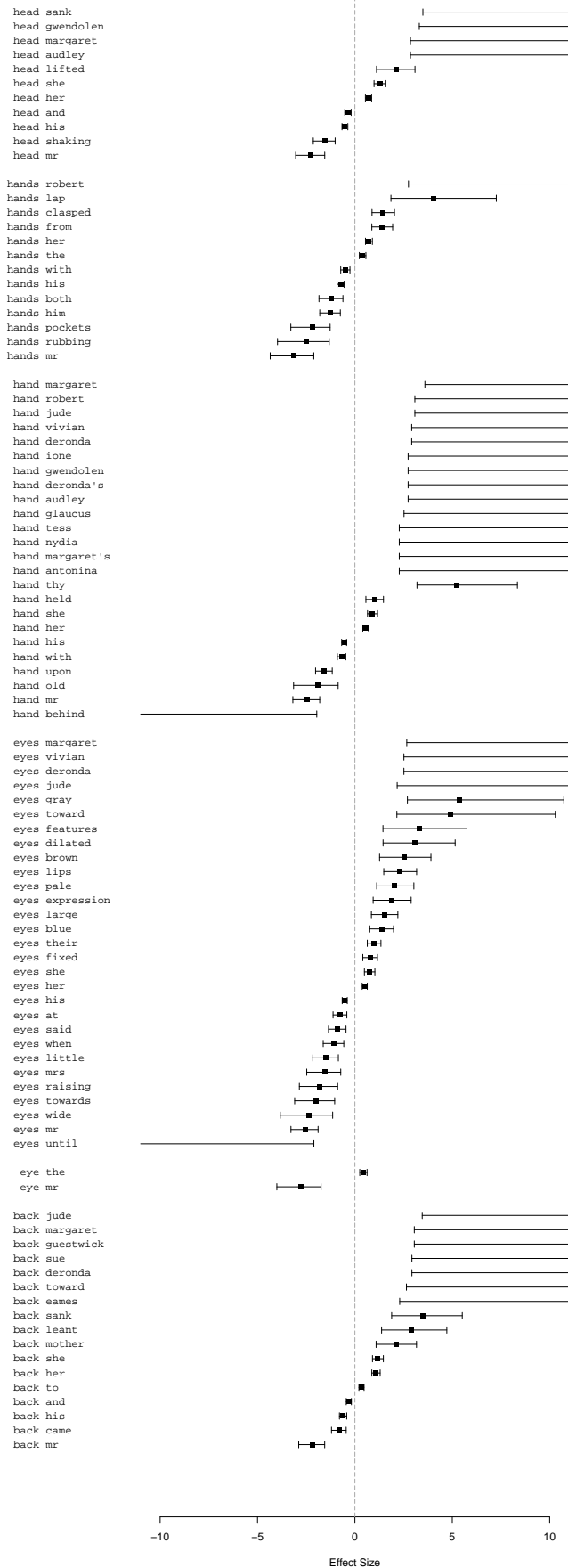


Finally we compare all of Dickens' novels against a set of 19th century novels to check if we can reproduce the observations from [Mahlberg \(2013\)](#) about Dickensian body language patterns. Practically we see this in terms such as *rubbing* co-occurring more frequently with *hands* in Dickens than the other 19th century novels.

```
DICKENS <- get_book_tokens('dickens')
NCNB <- get_book_tokens('ntc')
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)
```



## References

Mahlberg, M. (2013). *Corpus Stylistics and Dickens's Fiction*. London: Routledge.