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, xlab = 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

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

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>H_A</th>
<th>M_A</th>
<th>H_B</th>
<th>M_B</th>
<th>effect_size</th>
<th>CI_lower</th>
<th>CI_upper</th>
<th>p_value</th>
<th>p_adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:</td>
<td>back</td>
<td>my</td>
<td>0</td>
<td>1890</td>
<td>17</td>
<td>1823</td>
<td>Inf</td>
<td>5.841370e-06</td>
<td>5.671970e-03</td>
<td></td>
</tr>
<tr>
<td>2:</td>
<td>eyes</td>
<td>CHIMERA</td>
<td>30</td>
<td>1620</td>
<td>0</td>
<td>1790</td>
<td>Inf</td>
<td>-3.074373 Inf</td>
<td>2.331613e-10</td>
<td>2.163737e-07</td>
</tr>
<tr>
<td>3:</td>
<td>eyes</td>
<td>my</td>
<td>1</td>
<td>1649</td>
<td>30</td>
<td>1760</td>
<td>4.812183</td>
<td>2.220141 10.156672</td>
<td>6.462279e-08</td>
<td>2.104017e-05</td>
</tr>
<tr>
<td>4:</td>
<td>hand</td>
<td>CHIMERA</td>
<td>40</td>
<td>2500</td>
<td>0</td>
<td>2580</td>
<td>Inf</td>
<td>-3.414695 5.686176e-13</td>
<td>6.493613e-10</td>
<td></td>
</tr>
<tr>
<td>5:</td>
<td>hand</td>
<td>my</td>
<td>7</td>
<td>2533</td>
<td>51</td>
<td>2529</td>
<td>2.866900</td>
<td>1.717219 4.254416</td>
<td>2.109590e-09</td>
<td>2.04576e-06</td>
</tr>
<tr>
<td>6:</td>
<td>hands</td>
<td>CHIMERA</td>
<td>22</td>
<td>1338</td>
<td>0</td>
<td>1590</td>
<td>Inf</td>
<td>-2.698213 3.646064e-08</td>
<td>3.021750e-05</td>
<td></td>
</tr>
<tr>
<td>7:</td>
<td>head</td>
<td>CHIMERA</td>
<td>40</td>
<td>1940</td>
<td>0</td>
<td>1970</td>
<td>Inf</td>
<td>-3.390409 1.497662e-12</td>
<td>1.572564e-09</td>
<td></td>
</tr>
<tr>
<td>8:</td>
<td>head</td>
<td>my</td>
<td>4</td>
<td>1976</td>
<td>27</td>
<td>1943</td>
<td>2.778609</td>
<td>1.253757 4.756674</td>
<td>1.727748e-05</td>
<td>9.070676e-03</td>
</tr>
<tr>
<td>9:</td>
<td>shoulder</td>
<td>CHIMERA</td>
<td>16</td>
<td>354</td>
<td>0</td>
<td>420</td>
<td>Inf</td>
<td>-2.178354 4.495079e-06</td>
<td>1.240642e-03</td>
<td></td>
</tr>
</tbody>
</table>

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