Get freqs for NWR YD project

AC

9/8/2020

Initial clean up

This is a better than what we had in divime, because it does more cleanup. To run it, you need to be in raw_YEL

```
grep "^[FM]A" txt/*.txt | # use only adult speech
cut -f 6 | #take transcription
sed "s/\[: /\[:_/g" |  sed "s/\[=! /\[:_/g" | sed "s/\[- /\[-_/g" | #remove spaces that are not real on
grep -v "\[-_" | #remove sentences in Eng and Pidgin
 grep -v "he " | grep -vw "they" | grep -vw "I"  | grep -vw "you" | grep -vw "your"| grep -vw "his" | g
grep -vw "alright" |grep -vw "already" |grep -vw "almost" |grep -vw "another" |
grep -vw "bye" |grep -vw "calling" |grep -vw "coming" |grep -vw "clothes" |
grep -vw "complete" |grep -vw "counting" |grep -vw "enough" |grep -vw "fight" |
grep -vw "first" |grep -vw "finished" |grep -vw "there"|grep -vw "which" |grep -vw "witchcraft" |grep -
grep -vw "window" |grep -vw "want"|grep -vw "sister"|grep -vw "sleeping"|
grep -vw "sometimes" | grep -vw "somehwere"|grep -vw "scared"|grep -vw "saw" |
grep -vw "inside" | grep -vw "gonna" | grep -vw "her" | grep -vw "him"|
grep -vw "here" | grep -vw "easy" | grep -vw "early" | grep -vw "eye"|
grep -vw "cook" | grep -vw "dog" | grep -vw "area" | grep -vw "around"|
grep -v "a@l" | grep -v "e@l" | grep -v "d@l"|
sed 's/\[[^[]*\]//g' | #delete comments
tr ' ' '\n' | #cut at word boundaries
tr -d '?' | tr -d '.' | tr -d '!' | tr -d '-' | tr -d ',' | #clean up punctuation
grep -v "&" | grep -v "@s" | #remove switches
grep -v "xxx" | grep -vw "xx" | grep -vw "hm" | grep -vw "mm" | grep -vw "mmhm" | grep -vw "oh" | grep
grep -v "[A-Z]" | grep -vw "Ńaamońo" | #get rid of all names
sed "s/aa+/aa/g" | sed "s/ee+/ee/g" | sed "s/ii+/ii/g" | sed "s/oo+/oo/g" | sed "s/uu+/aa/g" |
sed "s/êê+/êê/g" | sed "s/ââ+/ââ/g" | sed "s/áá+/áá/g" | sed "s/óó+/óó/g" | #exaggeration of vowel leng
sed "s/aaaa/aa/g" |
tr -d '>' | tr -d '<' |tr -d '(' |tr -d ')' | sed "s/@c//g" | # final cleaning and write out
sed "s/che/te/g" |  sed "s/chi/ti/g" | sort | grep -v "[0-9]" | sed '/^$/d' > words_corpus.txt
```

Getting frequencies for segments in our stimuli

Middy gave me an onset list that converts all vowels and consonants to common representations for the purposes of counting syllables (rossel-ortho-replacements.txt). I thought of the following changes:

- remove all rewrites for vowels
- but that destroys the context for the following rules, eg not removing colon means knw does not find a match in knw:a
- but why weren't these defined with regular expressions anyway?

- it looks like I should be careful with the 4- and 3-letter phonemes, because I do not have those in my rewrites but not otherwise

So all things considered, it looks like it may be easier to just do the replacement of the 4- and 3-letter segments here, just being careful to map these to something that is not used in the stimuli, such as ngm.

Also, this reveals my rewrites were incomplete, so I added some lines for the onsets that were missing in my initial correspondance list.

The correspondances here looks similar to the one in wrangling but it has more entries because we want to separate each phoneme, and we want to distinguish between short and long

```
# get frequencies in raw_YEL
#note, you should have ran the code documented in get_freq_cor.Rmd

scan("words_corpus.txt",what="char")->wds

wds[1:1000]
```

```
##    [1] "ń:aa"    "ń:ââ"    "ń:ââ"    "ń:ââ"    "ńaa"     "ńe"      "ńeekuwo"
##    [8] "ńeńe"    "ńoo"     "ńoo"     "ńuu"     "ńuw:o"   "ńââ"     "a"
##   [15] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##   [22] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##   [29] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##   [36] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##   [43] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##   [50] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##   [57] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##   [64] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##   [71] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##   [78] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##   [85] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##   [92] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##   [99] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [106] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [113] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [120] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [127] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [134] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [141] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [148] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [155] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [162] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [169] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [176] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [183] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [190] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [197] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [204] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [211] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [218] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [225] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [232] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [239] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [246] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [253] "a"       "a"       "a"       "a"       "a"       "a"       "a"
```

```
##  [260] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [267] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [274] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [281] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [288] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [295] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [302] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [309] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [316] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [323] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [330] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [337] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [344] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [351] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [358] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [365] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [372] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [379] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [386] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [393] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [400] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [407] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [414] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [421] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [428] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [435] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [442] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [449] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [456] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [463] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [470] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [477] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [484] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [491] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [498] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [505] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [512] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [519] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [526] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [533] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [540] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [547] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [554] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [561] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [568] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [575] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [582] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [589] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [596] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [603] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [610] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [617] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [624] "a"  "a"  "a"  "a"  "a"  "a"  "a"
##  [631] "a"  "a"  "a"  "a"  "a"  "a"  "a"
```

```
##  [638] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [645] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [652] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [659] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [666] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [673] "a"       "a"       "a"       "a"       "a"       "a"       "a"
##  [680] "a"       "a"       "a"       "aa"      "aa"      "aa"      "aa"
##  [687] "aa"      "aa"      "aa"      "aa"      "aa"      "aa"      "aa"
##  [694] "aa"      "aa"      "aa"      "aa"      "aa"      "aa"      "aa"
##  [701] "aa"      "aa"      "aa"      "aa"      "aa"      "aa"      "aa"
##  [708] "aa"      "aa"      "aa"      "aa"      "aa"      "aa"      "aa"
##  [715] "aa"      "aa"      "aa"      "aa"      "aa"      "aa"      "aa"
##  [722] "aa"      "aa"      "aa"      "aa"      "aa"      "aa"      "aa"
##  [729] "aa"      "aa"      "aa"      "aa"      "aa"      "aa"      "aa"
##  [736] "aa"      "aa"      "aa"      "aa"      "aa"      "aa"      "aa"
##  [743] "aa"      "aa"      "aa"      "aa"      "aa"      "aa"      "aa"
##  [750] "aa"      "aa"      "aa"      "aa"      "aa"      "aa"      "aa"
##  [757] "aa"      "aa"      "aa"      "aa"      "aa"      "aa"      "aa"
##  [764] "aa"      "aa"      "aa"      "aa"      "aa"      "aa"      "aa"
##  [771] "aa"      "aa"      "aa"      "aa"      "aa"      "aa"      "aa"
##  [778] "aa"      "aa"      "aa"      "aa"      "aa"      "aa"      "aa"
##  [785] "aa"      "aa"      "aa"      "aa"      "aa"      "aa"      "aa"
##  [792] "aa"      "aa"      "aa"      "aa"      "aa"      "aa"      "aa"
##  [799] "aa"      "aa"      "aaé"     "ada"     "adê"     "adê"     "adê"
##  [806] "adê"     "adê"     "adê"     "adê"     "adê"     "adê"     "adê"
##  [813] "aka"     "aka"     "aka"     "aka"     "aka"     "aka"     "aka"
##  [820] "aka"     "aka"     "aka"     "aka"     "aka"     "aka"     "aki"
##  [827] "aki"     "aki"     "aki"     "aki"     "aki"     "aki"     "aki"
##  [834] "aki"     "aki"     "aki"     "aki"     "aki"     "aki"     "aki"
##  [841] "aki"     "aki"     "aki"     "aki"     "aki"     "aki"     "aki"
##  [848] "aki"     "aki"     "aki"     "akê"     "akê"     "akê"     "akê"
##  [855] "akê"     "akê"     "akê"     "akê"     "akê"     "al:ii"   "al:ii"
##  [862] "al:ii"   "al:ii"   "al:ii"   "al:ii"   "al:ii"   "al:ii"   "al:ii"
##  [869] "al:ii"   "al:ii"   "al:ii"   "al:ii"   "al:ii"   "al:ii"   "al:ii"
##  [876] "al:ii"   "al:ii"   "al:ii"   "al:ii"   "al:ii"   "al:ii"   "al:ii"
##  [883] "al:ii"   "al:ii"   "al:ii"   "al:ii"   "al:ii"   "al:ii"   "al:ii"
##  [890] "ala"     "ala"     "ala"     "ala"     "ala"     "ala"     "ala"
##  [897] "ala"     "ala"     "ala"     "ala"     "ala"     "ala"     "ala"
##  [904] "ala"     "ala"     "ala"     "ala"     "ala"     "ala"     "ala"
##  [911] "ala"     "ala"     "ala"     "ala"     "ala"     "ala"     "ala"
##  [918] "ala"     "ala"     "ala"     "ala"     "ala"     "ala"     "ala"
##  [925] "ala"     "ala"     "ala"     "ala"     "ala"     "ala"     "ala"
##  [932] "ala"     "ala"     "ala"     "ala"     "ala"     "ala"     "ala"
##  [939] "ala"     "ala"     "ala"     "ala"     "ala"     "ala"     "ala"
##  [946] "ala"     "ala"     "ala"     "ala"     "ala"     "ala"     "ala"
##  [953] "ala"     "ala"     "ala"     "ala"     "ala"     "ala"     "ala"
##  [960] "ala"     "ala"     "ala"     "ala"     "ala"     "ala"     "ala"
##  [967] "ala"     "ala"     "ala"     "ala"     "ala"     "ala"     "ala"
##  [974] "ala"     "ala"     "ala"     "ala"     "ala"     "ala"     "ala"
##  [981] "ala"     "ala"     "ala"     "ala"     "ala"     "ala"     "ala"
##  [988] "ala"     "ala"     "ala"     "ala"     "ala"     "ala"     "ala"
##  [995] "ala"     "ala"     "ala"     "ala"     "ala"     "ala"
```

```r
#initialize the unichar phono-like representation
wds_uni=wds

for(i in 1:dim(correspondances)[1]) { #transform into unicharacter
  wds_uni=gsub(correspondances[i,1],correspondances[i,2],wds_uni,fixed=T,useBytes = T)
}
wds_uni[1:1000]
```

```
##    [1] "ń 3 "           "ń 5 "           "ń 5 "             "ń 5 "
##    [5] "ń 20 "          "ń 32 "          "ń 21 k 35 w 34 " "ń 32 ń 32 "
##    [9] "ń 23 "          "ń 23 "          "ń 24 "            "ń 35 w 13 "
##   [13] "ń 16 "          "31 "            "31 "              "31 "
##   [17] "31 "            "31 "            "31 "              "31 "
##   [21] "31 "            "31 "            "31 "              "31 "
##   [25] "31 "            "31 "            "31 "              "31 "
##   [29] "31 "            "31 "            "31 "              "31 "
##   [33] "31 "            "31 "            "31 "              "31 "
##   [37] "31 "            "31 "            "31 "              "31 "
##   [41] "31 "            "31 "            "31 "              "31 "
##   [45] "31 "            "31 "            "31 "              "31 "
##   [49] "31 "            "31 "            "31 "              "31 "
##   [53] "31 "            "31 "            "31 "              "31 "
##   [57] "31 "            "31 "            "31 "              "31 "
##   [61] "31 "            "31 "            "31 "              "31 "
##   [65] "31 "            "31 "            "31 "              "31 "
##   [69] "31 "            "31 "            "31 "              "31 "
##   [73] "31 "            "31 "            "31 "              "31 "
##   [77] "31 "            "31 "            "31 "              "31 "
##   [81] "31 "            "31 "            "31 "              "31 "
##   [85] "31 "            "31 "            "31 "              "31 "
##   [89] "31 "            "31 "            "31 "              "31 "
##   [93] "31 "            "31 "            "31 "              "31 "
##   [97] "31 "            "31 "            "31 "              "31 "
##  [101] "31 "            "31 "            "31 "              "31 "
##  [105] "31 "            "31 "            "31 "              "31 "
##  [109] "31 "            "31 "            "31 "              "31 "
##  [113] "31 "            "31 "            "31 "              "31 "
##  [117] "31 "            "31 "            "31 "              "31 "
##  [121] "31 "            "31 "            "31 "              "31 "
##  [125] "31 "            "31 "            "31 "              "31 "
##  [129] "31 "            "31 "            "31 "              "31 "
##  [133] "31 "            "31 "            "31 "              "31 "
##  [137] "31 "            "31 "            "31 "              "31 "
##  [141] "31 "            "31 "            "31 "              "31 "
##  [145] "31 "            "31 "            "31 "              "31 "
##  [149] "31 "            "31 "            "31 "              "31 "
##  [153] "31 "            "31 "            "31 "              "31 "
##  [157] "31 "            "31 "            "31 "              "31 "
##  [161] "31 "            "31 "            "31 "              "31 "
##  [165] "31 "            "31 "            "31 "              "31 "
##  [169] "31 "            "31 "            "31 "              "31 "
##  [173] "31 "            "31 "            "31 "              "31 "
##  [177] "31 "            "31 "            "31 "              "31 "
##  [181] "31 "            "31 "            "31 "              "31 "
```

```
##  [185] "31 "          "31 "          "31 "          "31 "
##  [189] "31 "          "31 "          "31 "          "31 "
##  [193] "31 "          "31 "          "31 "          "31 "
##  [197] "31 "          "31 "          "31 "          "31 "
##  [201] "31 "          "31 "          "31 "          "31 "
##  [205] "31 "          "31 "          "31 "          "31 "
##  [209] "31 "          "31 "          "31 "          "31 "
##  [213] "31 "          "31 "          "31 "          "31 "
##  [217] "31 "          "31 "          "31 "          "31 "
##  [221] "31 "          "31 "          "31 "          "31 "
##  [225] "31 "          "31 "          "31 "          "31 "
##  [229] "31 "          "31 "          "31 "          "31 "
##  [233] "31 "          "31 "          "31 "          "31 "
##  [237] "31 "          "31 "          "31 "          "31 "
##  [241] "31 "          "31 "          "31 "          "31 "
##  [245] "31 "          "31 "          "31 "          "31 "
##  [249] "31 "          "31 "          "31 "          "31 "
##  [253] "31 "          "31 "          "31 "          "31 "
##  [257] "31 "          "31 "          "31 "          "31 "
##  [261] "31 "          "31 "          "31 "          "31 "
##  [265] "31 "          "31 "          "31 "          "31 "
##  [269] "31 "          "31 "          "31 "          "31 "
##  [273] "31 "          "31 "          "31 "          "31 "
##  [277] "31 "          "31 "          "31 "          "31 "
##  [281] "31 "          "31 "          "31 "          "31 "
##  [285] "31 "          "31 "          "31 "          "31 "
##  [289] "31 "          "31 "          "31 "          "31 "
##  [293] "31 "          "31 "          "31 "          "31 "
##  [297] "31 "          "31 "          "31 "          "31 "
##  [301] "31 "          "31 "          "31 "          "31 "
##  [305] "31 "          "31 "          "31 "          "31 "
##  [309] "31 "          "31 "          "31 "          "31 "
##  [313] "31 "          "31 "          "31 "          "31 "
##  [317] "31 "          "31 "          "31 "          "31 "
##  [321] "31 "          "31 "          "31 "          "31 "
##  [325] "31 "          "31 "          "31 "          "31 "
##  [329] "31 "          "31 "          "31 "          "31 "
##  [333] "31 "          "31 "          "31 "          "31 "
##  [337] "31 "          "31 "          "31 "          "31 "
##  [341] "31 "          "31 "          "31 "          "31 "
##  [345] "31 "          "31 "          "31 "          "31 "
##  [349] "31 "          "31 "          "31 "          "31 "
##  [353] "31 "          "31 "          "31 "          "31 "
##  [357] "31 "          "31 "          "31 "          "31 "
##  [361] "31 "          "31 "          "31 "          "31 "
##  [365] "31 "          "31 "          "31 "          "31 "
##  [369] "31 "          "31 "          "31 "          "31 "
##  [373] "31 "          "31 "          "31 "          "31 "
##  [377] "31 "          "31 "          "31 "          "31 "
##  [381] "31 "          "31 "          "31 "          "31 "
##  [385] "31 "          "31 "          "31 "          "31 "
##  [389] "31 "          "31 "          "31 "          "31 "
##  [393] "31 "          "31 "          "31 "          "31 "
##  [397] "31 "          "31 "          "31 "          "31 "
```

```
##  [401] "31 "           "31 "           "31 "           "31 "
##  [405] "31 "           "31 "           "31 "           "31 "
##  [409] "31 "           "31 "           "31 "           "31 "
##  [413] "31 "           "31 "           "31 "           "31 "
##  [417] "31 "           "31 "           "31 "           "31 "
##  [421] "31 "           "31 "           "31 "           "31 "
##  [425] "31 "           "31 "           "31 "           "31 "
##  [429] "31 "           "31 "           "31 "           "31 "
##  [433] "31 "           "31 "           "31 "           "31 "
##  [437] "31 "           "31 "           "31 "           "31 "
##  [441] "31 "           "31 "           "31 "           "31 "
##  [445] "31 "           "31 "           "31 "           "31 "
##  [449] "31 "           "31 "           "31 "           "31 "
##  [453] "31 "           "31 "           "31 "           "31 "
##  [457] "31 "           "31 "           "31 "           "31 "
##  [461] "31 "           "31 "           "31 "           "31 "
##  [465] "31 "           "31 "           "31 "           "31 "
##  [469] "31 "           "31 "           "31 "           "31 "
##  [473] "31 "           "31 "           "31 "           "31 "
##  [477] "31 "           "31 "           "31 "           "31 "
##  [481] "31 "           "31 "           "31 "           "31 "
##  [485] "31 "           "31 "           "31 "           "31 "
##  [489] "31 "           "31 "           "31 "           "31 "
##  [493] "31 "           "31 "           "31 "           "31 "
##  [497] "31 "           "31 "           "31 "           "31 "
##  [501] "31 "           "31 "           "31 "           "31 "
##  [505] "31 "           "31 "           "31 "           "31 "
##  [509] "31 "           "31 "           "31 "           "31 "
##  [513] "31 "           "31 "           "31 "           "31 "
##  [517] "31 "           "31 "           "31 "           "31 "
##  [521] "31 "           "31 "           "31 "           "31 "
##  [525] "31 "           "31 "           "31 "           "31 "
##  [529] "31 "           "31 "           "31 "           "31 "
##  [533] "31 "           "31 "           "31 "           "31 "
##  [537] "31 "           "31 "           "31 "           "31 "
##  [541] "31 "           "31 "           "31 "           "31 "
##  [545] "31 "           "31 "           "31 "           "31 "
##  [549] "31 "           "31 "           "31 "           "31 "
##  [553] "31 "           "31 "           "31 "           "31 "
##  [557] "31 "           "31 "           "31 "           "31 "
##  [561] "31 "           "31 "           "31 "           "31 "
##  [565] "31 "           "31 "           "31 "           "31 "
##  [569] "31 "           "31 "           "31 "           "31 "
##  [573] "31 "           "31 "           "31 "           "31 "
##  [577] "31 "           "31 "           "31 "           "31 "
##  [581] "31 "           "31 "           "31 "           "31 "
##  [585] "31 "           "31 "           "31 "           "31 "
##  [589] "31 "           "31 "           "31 "           "31 "
##  [593] "31 "           "31 "           "31 "           "31 "
##  [597] "31 "           "31 "           "31 "           "31 "
##  [601] "31 "           "31 "           "31 "           "31 "
##  [605] "31 "           "31 "           "31 "           "31 "
##  [609] "31 "           "31 "           "31 "           "31 "
##  [613] "31 "           "31 "           "31 "           "31 "
```

```
##  [617] "31 "           "31 "           "31 "           "31 "
##  [621] "31 "           "31 "           "31 "           "31 "
##  [625] "31 "           "31 "           "31 "           "31 "
##  [629] "31 "           "31 "           "31 "           "31 "
##  [633] "31 "           "31 "           "31 "           "31 "
##  [637] "31 "           "31 "           "31 "           "31 "
##  [641] "31 "           "31 "           "31 "           "31 "
##  [645] "31 "           "31 "           "31 "           "31 "
##  [649] "31 "           "31 "           "31 "           "31 "
##  [653] "31 "           "31 "           "31 "           "31 "
##  [657] "31 "           "31 "           "31 "           "31 "
##  [661] "31 "           "31 "           "31 "           "31 "
##  [665] "31 "           "31 "           "31 "           "31 "
##  [669] "31 "           "31 "           "31 "           "31 "
##  [673] "31 "           "31 "           "31 "           "31 "
##  [677] "31 "           "31 "           "31 "           "31 "
##  [681] "31 "           "31 "           "20 "           "20 "
##  [685] "20 "           "20 "           "20 "           "20 "
##  [689] "20 "           "20 "           "20 "           "20 "
##  [693] "20 "           "20 "           "20 "           "20 "
##  [697] "20 "           "20 "           "20 "           "20 "
##  [701] "20 "           "20 "           "20 "           "20 "
##  [705] "20 "           "20 "           "20 "           "20 "
##  [709] "20 "           "20 "           "20 "           "20 "
##  [713] "20 "           "20 "           "20 "           "20 "
##  [717] "20 "           "20 "           "20 "           "20 "
##  [721] "20 "           "20 "           "20 "           "20 "
##  [725] "20 "           "20 "           "20 "           "20 "
##  [729] "20 "           "20 "           "20 "           "20 "
##  [733] "20 "           "20 "           "20 "           "20 "
##  [737] "20 "           "20 "           "20 "           "20 "
##  [741] "20 "           "20 "           "20 "           "20 "
##  [745] "20 "           "20 "           "20 "           "20 "
##  [749] "20 "           "20 "           "20 "           "20 "
##  [753] "20 "           "20 "           "20 "           "20 "
##  [757] "20 "           "20 "           "20 "           "20 "
##  [761] "20 "           "20 "           "20 "           "20 "
##  [765] "20 "           "20 "           "20 "           "20 "
##  [769] "20 "           "20 "           "20 "           "20 "
##  [773] "20 "           "20 "           "20 "           "20 "
##  [777] "20 "           "20 "           "20 "           "20 "
##  [781] "20 "           "20 "           "20 "           "20 "
##  [785] "20 "           "20 "           "20 "           "20 "
##  [789] "20 "           "20 "           "20 "           "20 "
##  [793] "20 "           "20 "           "20 "           "20 "
##  [797] "20 "           "20 "           "20 "           "20 "
##  [801] "20 28 "        "31 d 31 "      "31 d 26 "      "31 d 26 "
##  [805] "31 d 26 "      "31 d 26 "      "31 d 26 "      "31 d 26 "
##  [809] "31 d 26 "      "31 d 26 "      "31 d 26 "      "31 d 26 "
##  [813] "31 k 31 "      "31 k 31 "      "31 k 31 "      "31 k 31 "
##  [817] "31 k 31 "      "31 k 31 "      "31 k 31 "      "31 k 31 "
##  [821] "31 k 31 "      "31 k 31 "      "31 k 31 "      "31 k 31 "
##  [825] "31 k 31 "      "31 k 33 "      "31 k 33 "      "31 k 33 "
##  [829] "31 k 33 "      "31 k 33 "      "31 k 33 "      "31 k 33 "
```

```
##  [833] "31 k 33 "        "31 k 33 "        "31 k 33 "        "31 k 33 "
##  [837] "31 k 33 "        "31 k 33 "        "31 k 33 "        "31 k 33 "
##  [841] "31 k 33 "        "31 k 33 "        "31 k 33 "        "31 k 33 "
##  [845] "31 k 33 "        "31 k 33 "        "31 k 33 "        "31 k 33 "
##  [849] "31 k 33 "        "31 k 33 "        "31 k 26 "        "31 k 26 "
##  [853] "31 k 26 "        "31 k 26 "        "31 k 26 "        "31 k 26 "
##  [857] "31 k 26 "        "31 k 26 "        "31 k 26 "        "31 l 1 "
##  [861] "31 l 1 "         "31 l 1 "         "31 l 1 "         "31 l 1 "
##  [865] "31 l 1 "         "31 l 1 "         "31 l 1 "         "31 l 1 "
##  [869] "31 l 1 "         "31 l 1 "         "31 l 1 "         "31 l 1 "
##  [873] "31 l 1 "         "31 l 1 "         "31 l 1 "         "31 l 1 "
##  [877] "31 l 1 "         "31 l 1 "         "31 l 1 "         "31 l 1 "
##  [881] "31 l 1 "         "31 l 1 "         "31 l 1 "         "31 l 1 "
##  [885] "31 l 1 "         "31 l 1 "         "31 l 1 "         "31 l 1 "
##  [889] "31 l 1 "         "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [893] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [897] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [901] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [905] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [909] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [913] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [917] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [921] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [925] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [929] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [933] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [937] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [941] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [945] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [949] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [953] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [957] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [961] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [965] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [969] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [973] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [977] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [981] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [985] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [989] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [993] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
##  [997] "31 l 31 "        "31 l 31 "        "31 l 31 "        "31 l 31 "
```

```
wds_uni=unlist(strsplit(wds_uni,split=" "))
counts=data.frame(table(wds_uni))
counts$wds_uni=as.character(counts$wds_uni)

colnames(correspondances)<-c("ortho","fake")
correspondances=data.frame(correspondances)
correspondances$fake=gsub(" ","",as.character(correspondances$fake))

#backtranslate
merge(counts,correspondances,by.x="wds_uni",by.y="fake")->counts
counts=counts[order(counts$Freq),]
```

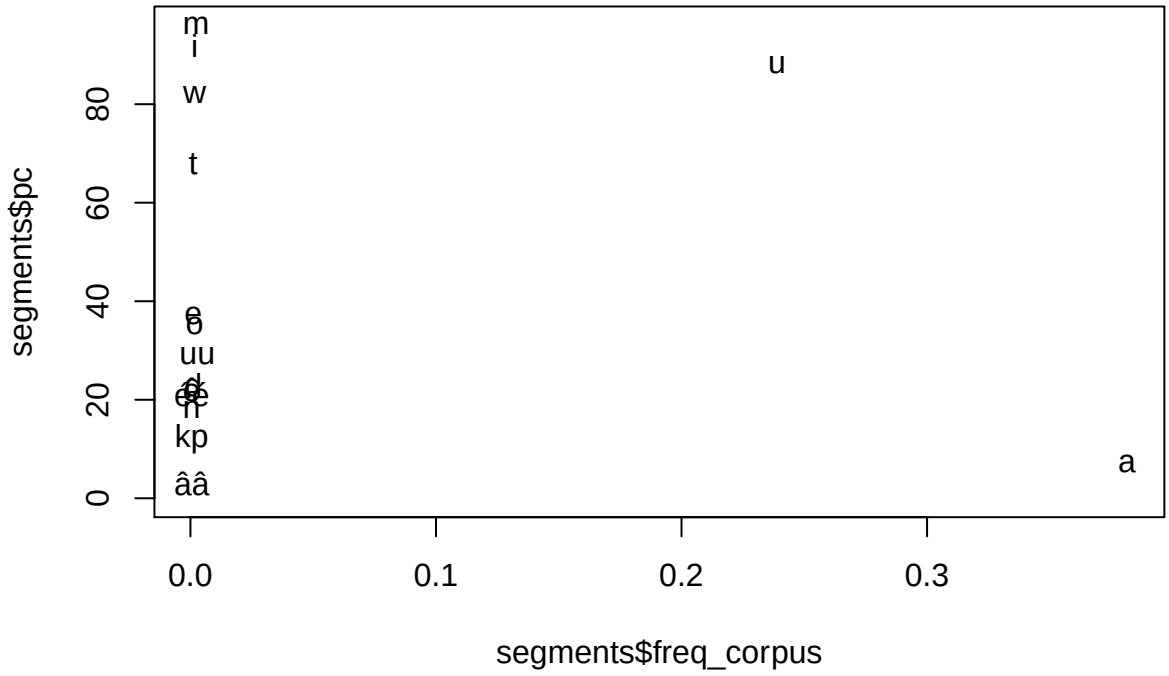
```
colnames(counts)[2]<-"counts_corpus"
counts$freq_corpus=counts$counts_corpus/sum(counts$counts_corpus)

write.table(counts,"segment-counts.txt",col.names = T,row.names = F,quote=T,sep="\t")
```

Analyses

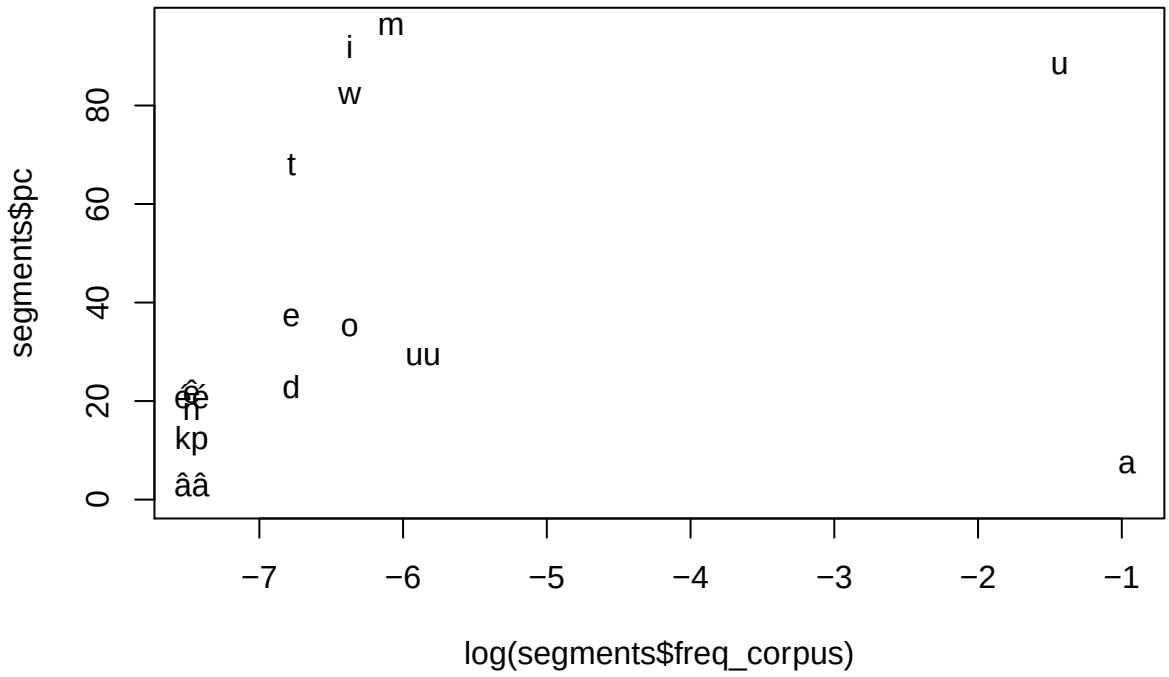
I ran the code in wrangling again, to integrate these counts & freqs into the main data set.

In addition, I wrote out a new version of segments which has the corpus frequency

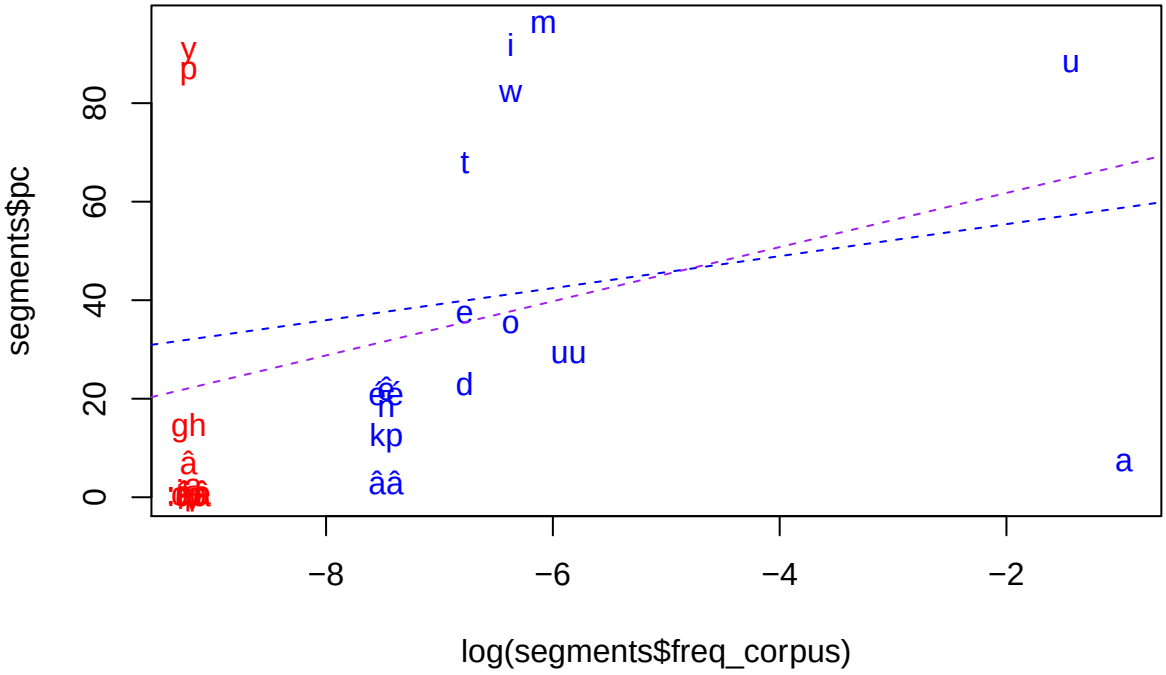
```
read.table("segments_with_cor_freq.txt",header=T)->segments # not ran
plot(segments$pc~segments$freq_corpus,type="n")
text(segments$pc~segments$freq_corpus,labels=segments$ortho)
```




```
plot(segments$pc~log(segments$freq_corpus),type="n")
text(segments$pc~log(segments$freq_corpus),labels=segments$ortho)
```




```
#some stimuli sounds do not occur in the corpus at all
#so we'll give them a really small frequency, just so that they show up
#and we tag them in red
segments$freq_corpus[is.na(segments$freq_corpus)]<-.0001
plot(segments$pc~log(segments$freq_corpus),type="n",main="blue fitted to sounds in corpus, purple to al
text(segments$pc~log(segments$freq_corpus),labels=segments$ortho,col=ifelse(segments$freq_corpus<.00057
#abline(lm(segments$pc~log(segments$freq_corpus),subset=c(segments$freq_corpus<.00057)),col="red",lty=2
abline(lm(segments$pc~log(segments$freq_corpus),subset=c(segments$freq_corpus>=.00057)),col="blue",lty=
abline(lm(segments$pc~log(segments$freq_corpus)),col="purple",lty=2)
```


blue fitted to sounds in corpus, purple to all
segments$pc
0
20
40
60
80
−8
−6
−4
−2
log(segments$freq_corpus)
m
i
y
p
u
w
t
e
o
uu
d
gh
kp
a
â
ââ