# clean_for_r

October 29, 2022

```
[ ]: import pandas as pd

     data_df = pd.read_csv('regilaul_vowels.csv')

     data_df.head()
```

```
[ ]:    Unnamed: 0.1  Unnamed: 0       word  syll shape  index segment  quantity  \
     0             0           0       sain  sain  CVVC      0      ai         3
     1             1           1   mal´lika,  mal´  CVCC      0    a(i)         2
     2             2           2   mal´lika,    li    CV      1       i         2
     3             3           3  mal´likast  mal´  CVCC      0    a(i)         2
     4             4           4  mal´likast    li    CV      1       i         2

        stressed  ictus  duration  midpoint           f1           f2  fileid  \
     0         1  ictus  0.217500  4.263002   714.071536  1129.944778      41
     1         1    off  0.179055  4.614738   946.021821  1448.348662      41
     2         0  ictus  0.130760  4.936603   574.096119  1864.237985      41
     3         1  ictus  0.193353  5.891856  1075.429489  1570.329209      41
     4         0    off  0.122714  6.186455   589.298998  1688.844658      41

       performer
     0        LK
     1        LK
     2        LK
     3        LK
     4        LK
```

```
[ ]: r_df =␣
      ↪data_df[["word","syll","shape","index","segment","quantity","stressed","ictus",␣
      ↪"duration","f1","f2","fileid","performer"]].copy()
     r_df.tail()
```

```
[ ]:            word  syll shape  index segment  quantity  stressed  ictus  \
     753      südant  dant  CVCC      1                 3         1    off
     754       sülle   sül   CVC      0      yl         2         1  ictus
     755       sülle    le    CV      1       e         1         0    off
     756  rabadaie.    ra    CV      0       a         1         0  ictus
     757  rabadaie.    ba    CV      1                 1         0    off
     duration          f1           f2  fileid performer
753  0.172132  654.258343  1112.070873      65        LK
754  0.336953  640.080936  1706.830954      65        LK
755  0.308579  721.464844  1631.467177      65        LK
756  0.333783  935.850426  1478.448942      65        LK
757  0.319876  824.139440  1571.052387      65        LK
```

[ ]:
```
labels = {"index": "foot_i", "fileid":"song"}
#make labels more descriptive for summary stats
pd.DataFrame.rename(r_df, columns=labels, inplace=True)
r_df.tail()
```

[ ]:
```
           word  syll  shape  foot_i segment  quantity  stressed  ictus  \
753     südant  dant   CVCC       1                 3         1    off
754      sülle   sül    CVC       0      yl          2         1  ictus
755      sülle    le     CV       1       e          1         0    off
756  rabadaie.    ra     CV       0       a          1         0  ictus
757  rabadaie.    ba     CV       1                 1         0    off

     duration          f1           f2  song performer
753  0.172132  654.258343  1112.070873    65        LK
754  0.336953  640.080936  1706.830954    65        LK
755  0.308579  721.464844  1631.467177    65        LK
756  0.333783  935.850426  1478.448942    65        LK
757  0.319876  824.139440  1571.052387    65        LK
```

[ ]:
```
r_df.segment.unique()
```

[ ]:
```
array(['ai', 'a(i)', 'i', 'e', 'a', 'i ', 'u', ' ', ' ', 'o ', 'y', ' ',
       ' ', 'au', 'u ', 'e ', 'l', ' ', 'æ', 'al', 'ø ', 'o', 'ei', 'ae',
       'ee', 'a ', 'el', 'ell', 'ju', 'ull', 'ul', 'æi', ' ii', 'yy ',
       'ee ', 'oi', ' i', 'ui', 'ii ', 'ui ', ' ', 'il', 'yi', 'ja',
       'æ ', 'yy', ' ', 'ææ', 'oo', 'aa', ' u', 'ol', ' e', 'oe', 'ea',
       'øø ', 'eil', 'ei ', 'yl'], dtype=object)
```

[ ]:
```
#r_df = r_df[r_df['segment'].str.len() == 1]

#put off approximant codas until rhyme analysis:
r_df = r_df[r_df['segment'].str.contains('l')==False]
r_df = r_df[r_df['segment'].str.contains('j')==False]
r_df = r_df[r_df['segment'].str.contains(' ')==False]
#r_df = r_df[r_df['segment'].str.contains('y')==False]



# r_df = r_df[r_df['segment'].str.contains('ei')==False]
# r_df = r_df[r_df['segment'].str.contains('ea')==False]
# r_df = r_df[r_df['segment'].str.contains('æi')==False]
# r_df = r_df[r_df['segment'].str.contains('ui')==False]
# r_df = r_df[r_df['segment'].str.contains('oi')==False]
# r_df = r_df[r_df['segment'].str.contains('ae')==False]
r_df = r_df[r_df['segment'].str.contains('\(')==False]

# r_df = r_df[r_df['segment'].str.contains('oe')==False]

# r_df = r_df[r_df['segment'].str.contains('au')==False]
# r_df = r_df[r_df['segment'].str.contains('ai')==False]
# r_df = r_df[r_df['segment'].str.contains('yi')==False]
# r_df = r_df[r_df['segment'].str.contains(' i')==False]
# r_df = r_df[r_df['segment'].str.contains(' e')==False]
# r_df = r_df[r_df['segment'].str.contains(' u')==False]




r_df.segment.unique()
```

[ ]:

```
array(['ai', 'i', 'e', 'a', 'i ', 'u', ' ', ' ', 'o ', 'y', ' ', ' ',
       'au', 'u ', 'e ', ' ', 'æ', 'ø ', 'o', 'ei', 'ae', 'ee', 'a ',
       'æi', ' ii', 'yy ', 'ee ', 'oi', ' i', 'ui', 'ii ', 'ui ', ' ',
       'yi', 'æ ', 'yy', ' ', 'ææ', 'oo', 'aa', ' u', ' e', 'oe', 'ea',
       'øø ', 'ei '], dtype=object)
```

[ ]:

```
r_df["qual_id"] = r_df['segment'].str[0]
r_df.head()
```

[ ]:

```
             word  syll shape  foot_i segment  quantity  stressed  ictus  \
0            sain  sain  CVVC       0      ai         3         1  ictus
2       mal´lika,    li    CV       1       i         2         0  ictus
4      mal´likast    li    CV       1       i         2         0    off
5            sain  sain  CVVC       0      ai         3         1    off
7  man´nipil´li,    ni    CV       1       i         2         0    off

   duration          f1           f2  song performer qual_id
0  0.217500  714.071536  1129.944778    41        LK       a
2  0.130760  574.096119  1864.237985    41        LK       i
4  0.122714  589.298998  1688.844658    41        LK       i
5  0.159839  726.248682  1464.830515    41        LK       a
7  0.130770  493.465339  1222.836439    41        LK       i
```

[ ]:

```
#vowel space center of gravity for each performer
dictS = {'LO': (569.6259050003295, 1415.8521323345387),
 'LK': (693.7620755684583, 1438.6183172232195),
 'MH': (758.0741804312802, 1679.2886667630698)}
```

[ ]:

```
fonelist = r_df.f1.tolist()
ftwoList = r_df.f2.tolist()
perfList = r_df.performer.tolist()

tupList = list(zip(fonelist,ftwoList,perfList))
```

[ ]:

```
from scipy.spatial import distance
#calculate euclidean distance from F1, F2 values
euclid = []
for element in tupList:
    fx, fy = element[0], element[1]
    key = element[2]
    p_s = dictS.get(key)
    euc = distance.euclidean(p_s,[fx,fy])
    euclid.append(euc)


r_df["euclid"] = euclid
r_df.tail()
```

[ ]:

```
          word  syll shape  foot_i segment  quantity  stressed  ictus  \
752     südant    sü    CV       0       y         1         0  ictus
753     südant  dant  CVCC       1                3         1    off
755      sülle    le    CV       1       e         1         0    off
756  rabadaie.    ra    CV       0       a         1         0  ictus
757  rabadaie.    ba    CV       1                1         0    off

     duration          f1           f2  song performer qual_id      euclid
752  0.385577  520.775479  1749.135073    65        LK       y  355.450444
753  0.172132  654.258343  1112.070873    65        LK          328.928226
755  0.308579  721.464844  1631.467177    65        LK       e  194.828453
756  0.333783  935.850426  1478.448942    65        LK       a  245.343123
757  0.319876  824.139440  1571.052387    65        LK          185.841438
```

[ ]:

```
import os
vowel_df = "/Users/sarah/Git/regilaul_project/manuscript/results/vowel_master.
 ↪csv"
regi_clean = (open(vowel_df,'w'))
r_df.to_csv(regi_clean)
regi_clean.close()
```

[ ]:

```
r_df
```

[ ]:

```
              word  syll shape  foot_i segment  quantity  stressed  ictus  \
0             sain  sain  CVVC       0      ai         3         1  ictus
```

```
2        mal´lika,    li    CV        1       i         2         0  ictus
4       mal´likast    li    CV        1       i         2         0    off
5             sain  sain  CVVC        0      ai         3         1    off
7    man´nipil´li,    ni    CV        1       i         2         0    off
..             …   …   …       …      …       …       …   …
752         südant    sü    CV        0       y         1         0  ictus
753         südant  dant  CVCC        1                3         1    off
755          sülle    le    CV        1       e         1         0    off
756      rabadaie.    ra    CV        0       a         1         0  ictus
757      rabadaie.    ba    CV        1                1         0    off

     duration          f1           f2  song performer qual_id      euclid
0    0.217500  714.071536  1129.944778    41        LK       a  309.340958
2    0.130760  574.096119  1864.237985    41        LK       i  442.122204
4    0.122714  589.298998  1688.844658    41        LK       i  271.156332
5    0.159839  726.248682  1464.830515    41        LK       a   41.742771
7    0.130770  493.465339  1222.836439    41        LK       i  294.415695
..        …           …            …   …      …      …         …
752  0.385577  520.775479  1749.135073    65        LK       y  355.450444
753  0.172132  654.258343  1112.070873    65        LK          328.928226
755  0.308579  721.464844  1631.467177    65        LK       e  194.828453
756  0.333783  935.850426  1478.448942    65        LK       a  245.343123
757  0.319876  824.139440  1571.052387    65        LK          185.841438

[720 rows x 15 columns]
```