# Examining reduced features

## Load libraries

```
library(tidyverse)
```

## Prepare the data

Flattening the list

```
## # A tibble: 35,185 x 2
##    name  value
##    <chr> <list>
##  1 aal   <chr [15]>
##  2 aal   <chr [15]>
##  3 aal   <chr [15]>
##  4 aal   <chr [15]>
##  5 aal   <chr [15]>
##  6 aal   <chr [15]>
##  7 aal   <chr [15]>
##  8 aal   <chr [15]>
##  9 aal   <chr [15]>
## 10 aal   <chr [15]>
## # ... with 35,175 more rows
```

The next stage is providing the IDs and then unnesting further:

```
## # A tibble: 424,629 x 4
##    name  value                     min.features    ID
##    <chr> <chr>                            <int> <int>
##  1 aal   continuant                          15     1
##  2 aal   coronal                             15     1
##  3 aal   distributed                         15     1
##  4 aal   dorsal                             15     1
##  5 aal   front                               15     1
##  6 aal   high                                15     1
##  7 aal   labial                              15     1
##  8 aal   lateral                             15     1
##  9 aal   low                                 15     1
## 10 aal   lowered_larynx_implosive            15     1
## # ... with 424,619 more rows
```

## Questions

### (1) Are there features that are in all answers in all languages?

```
## [1] 1519
```

We have 1519 languages.

What about the distribution of features: we count it here a bit more explicit than necessary:

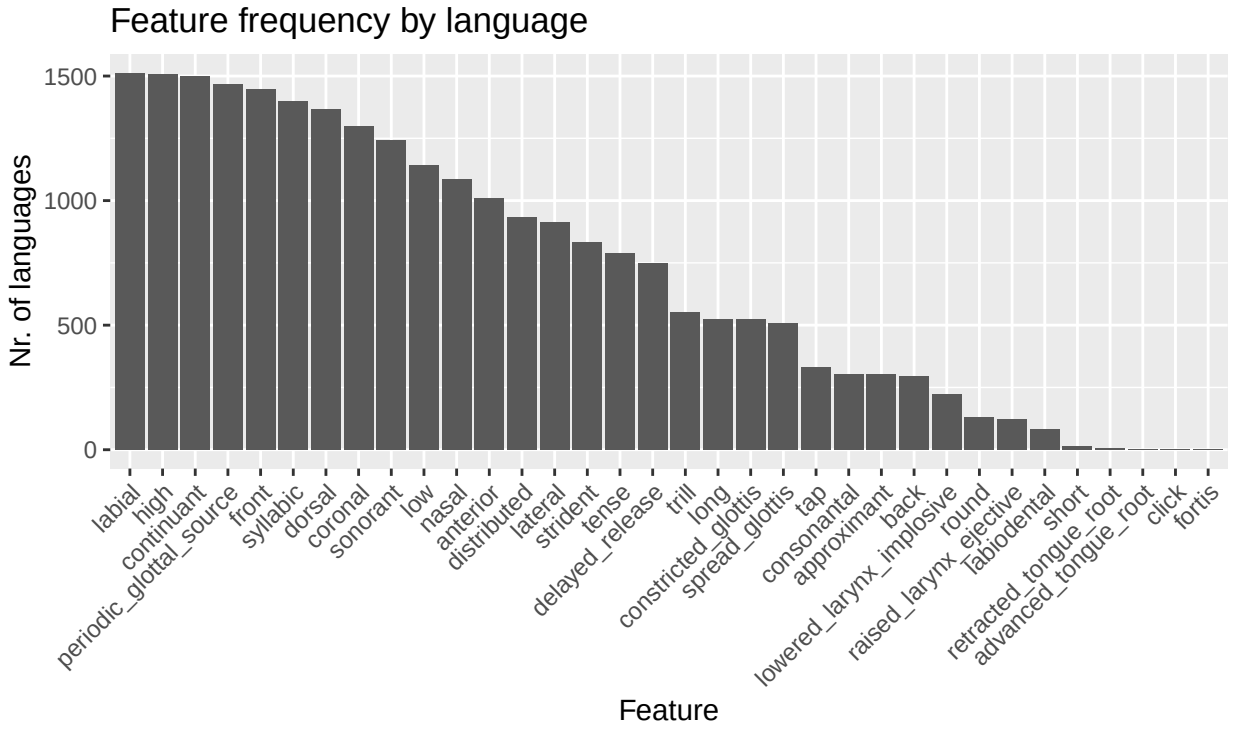
```
## # A tibble: 34 x 2
```

```
##    value                    total
##    <chr>                    <int>
##  1 labial                    1513
##  2 high                      1506
##  3 continuant                1499
##  4 periodic_glottal_source   1466
##  5 front                     1449
##  6 syllabic                  1397
##  7 dorsal                    1367
##  8 coronal                   1299
##  9 sonorant                  1243
## 10 low                       1140
## # ... with 24 more rows
```

Answer: almost but not, labial is the most common but it is missing in 6 languages

## (2) What's the distribution?

We cheat here a bit and take the summarized df **feature.total** to produce the bar plot, otherwise one would have to order the factor before plotting



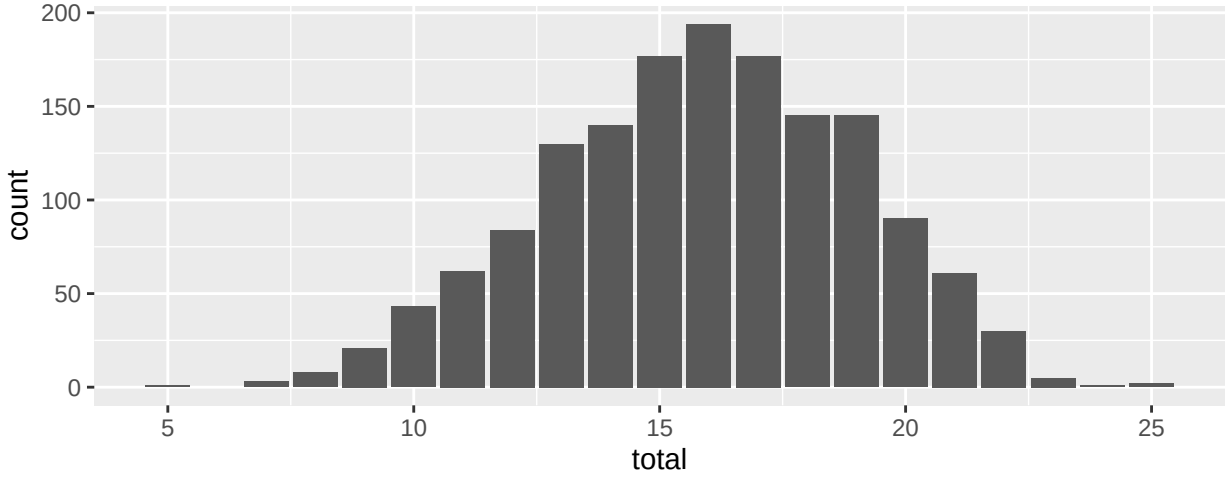

## (3) How many different features for each language?

We take the object **answers.df**, group it by `feature` and `language` and summarize with `n_distinct()`:

```
## # A tibble: 1,519 x 2
##    name  total
##    <chr> <int>
##  1 aal      21
##  2 aap      17
##  3 aar      19
##  4 aau      11
```

```
##  5 abb      17
##  6 abi      13
##  7 abn      18
##  8 abs      11
##  9 abt      10
## 10 abu      13
## # ... with 1,509 more rows
```

A histogram for this needs some bin fixing, so here is a bar plot




## Top features

Our `n` per language is the minimal number of features `min.features` in the different permutations.

We first group by language and feature and sum up how frequent the individual features are in the variable `nr.of.permutations`:

```
## # A tibble: 24,112 x 4
## # Groups:   name, value [24,112]
##    name  value               min.features nr.of.permutations
##    <chr> <chr>                      <int>              <int>
##  1 aal   anterior                      15                 12
##  2 aal   back                          15                  8
##  3 aal   consonantal                   15                  4
##  4 aal   constricted_glottis           15                 16
##  5 aal   continuant                    15                 24
##  6 aal   coronal                       15                 16
##  7 aal   distributed                   15                 12
##  8 aal   dorsal                        15                 24
##  9 aal   front                         15                 16
## 10 aal   high                          15                 24
## # ... with 24,102 more rows
```

Add the variable `feature.rank`, use `ties.method = "min"` so the multiple winners are allowed:

```
## # A tibble: 24,112 x 5
## # Groups:   name [1,519]
##    name  value               min.features nr.of.permutations feature.rank
##    <chr> <chr>                      <int>              <int>        <int>
##  1 aal   anterior                      15                 12           14
```

```
##  2 aal   back                          15                  8           18
##  3 aal   consonantal                   15                  4           20
##  4 aal   constricted_glottis           15                 16           11
##  5 aal   continuant                    15                 24            1
##  6 aal   coronal                       15                 16           11
##  7 aal   distributed                   15                 12           14
##  8 aal   dorsal                        15                 24            1
##  9 aal   front                         15                 16           11
## 10 aal   high                          15                 24            1
## # ... with 24,102 more rows
```

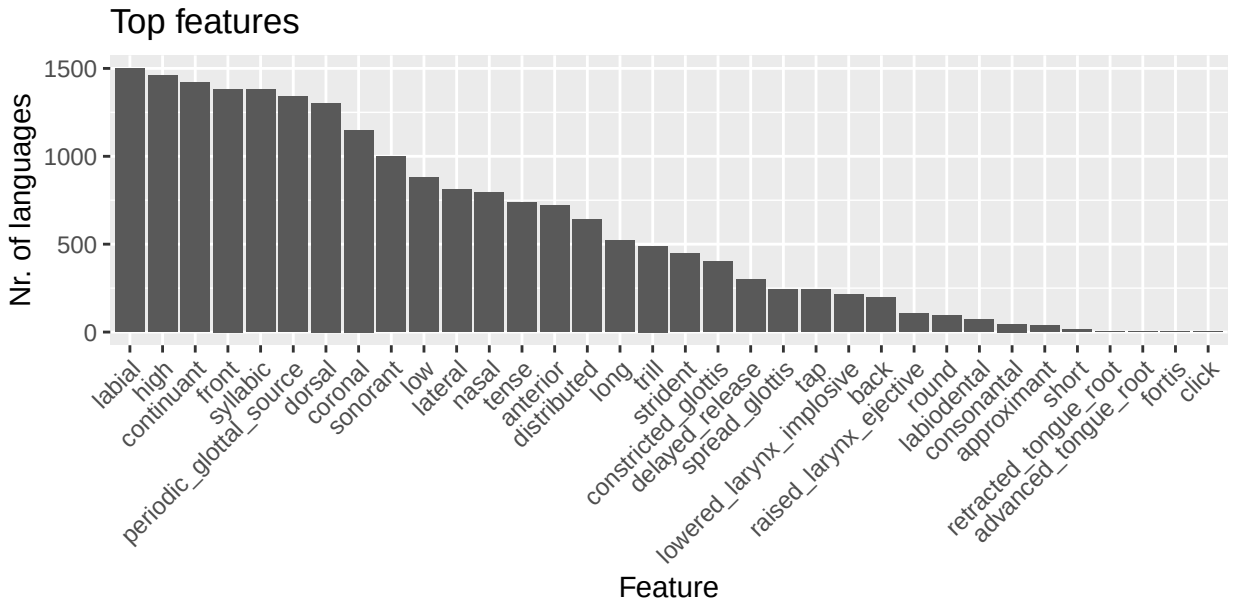
Filter only the top features:

```
## # A tibble: 19,941 x 5
## # Groups:   name [1,519]
##    name  value               min.features nr.of.permutations feature.rank
##    <chr> <chr>                      <int>              <int>        <int>
##  1 aal   anterior                      15                 12           14
##  2 aal   constricted_glottis           15                 16           11
##  3 aal   continuant                    15                 24            1
##  4 aal   coronal                       15                 16           11
##  5 aal   distributed                   15                 12           14
##  6 aal   dorsal                        15                 24            1
##  7 aal   front                         15                 16           11
##  8 aal   high                          15                 24            1
##  9 aal   labial                        15                 24            1
## 10 aal   lateral                       15                 24            1
## # ... with 19,931 more rows
```

Next we summarize the top features across languages:

```
## # A tibble: 34 x 2
##    value                    freq
##    <chr>                   <int>
##  1 labial                   1499
##  2 high                     1458
##  3 continuant               1418
##  4 front                    1383
##  5 syllabic                 1380
##  6 periodic_glottal_source  1341
##  7 dorsal                   1303
##  8 coronal                  1150
##  9 sonorant                  998
## 10 low                       877
## # ... with 24 more rows
```

How's it look?


Top features
Nr. of languages
1500
1000
500
0
labial
high
continuant
front
syllabic
periodic_glottal_source
dorsal
coronal
sonorant
low
lateral
nasal
tense
anterior
distributed
long
trill
strident
constricted_glottis
delayed_release
spread_glottis
tap
lowered_larynx_implosive
back
raised_larynx_ejective
round
labiodental
consonantal
approximant
short
retracted_tongue_root
advanced_tongue_root
fortis
click
Feature