Bilag 1: Analyse af Ludvig Holbergs forfatterskab

Victor Harbo Olesen

06/01/2021

Indledning

I dette dokument vil jeg lave forskellige analyser ud fra Holbergs samlede forfatterskab. Jeg fik ideen til at undersøge hele Holbergs forfatterskab, da jeg fandt ud af, at det ligger digitaliseret på holbergskrifter.dk. Jeg sendte en forespørgsel til folkene bag sitet og fik igennem dem tilsendt hele hans forfatterskab, på 109 tekster.

Import og cleaning af data

Dataen her er hentet hos holberskrifter.dk, med tilladelse fra Det Danske Sprog- og Litteraturselskab og Universitetet i Bergen, som begge skal have en stor tak for adgangen. Teksterne er hentet i XML format, det vil sige at dataen lige nu ikke ligger som den mest læsbare tekst, hverken for mennesket eller for computeren, Heldigvis kan man med et par linjer kode trække tekst ud af disse filer, som både er nemmere for os at læse og som giver mere mening at arbejde videre med.

```
read_tei <- function(folder) {
  list.files(folder, pattern = '\\.xml$', full.names = TRUE) %>%
    map_dfr(~.x %>% parseTEI(.,node = "text") %>%tibble())
}

tekster <- read_tei("/Users/vhol/Documents/Holbergskaffe/forfatterskab")
```

Dataanalysen er baseret på Tidy Data Principet fra tidytext pakken. Ideen bag dette princip er at man tager tekster og bryder dem ned i individuelle ord. Når man gør dette vil der være et ord per række i datasættet.

Det næste der skal ske med dataen er, at den skal transformeres til tidytext formatet, som kort er nævnt ovenfor. Derudover fjernes forstyrrende elementer som punktumer uden mellemrum mellem ordene, så alle ord tælles hver for sig. Grunden til, at punktumer fjernes fra teksten, er fordi de er forstyrende for maskinlæsningen af teksten. Eksempel: I sætningen "... og sådan blev det.Nu måtte han ..." tæller computeren syv ord, fordi den læser "det.Nu", som et ord og ikke to forskellige.

```
tekster <- rename(tekster, text = .)
data.frame(lapply(tekster, function(x) {gsub("[.]", " ", x)})) -> tekster
data.frame(lapply(tekster, function(x) {gsub("[:]", " ", x)})) -> tekster
data.frame(lapply(tekster, function(x)
  {str_replace_all(x, "([a-å])([0-9])","\\1 \\2")})) -> tekster
data.frame(lapply(tekster, function(x)
  {str_replace_all(x, "([0-9])([a-åA-Å])","\\1 \\2")})) -> tekster
tekster %>%
```

```
  unnest_tokens(word, text) %>%
  select(word, everything()) -> tekster_tidy
```

Analyse

Nu hvor dataen er indlæst og sat sammen til en stor tekst mængde er det muligt at regne på, hvor mange gange Holberg nævner kaffe i løbet af hans forfatterskab. Her skal vi igen huske, at kaffe staves anderledes på Holbergs tid end nu til dags. I koden har jeg brugt 17 forskellige kafferelaterede begreber. Disse ord er alle hentet fra https://holbergordbog.dk/, det samme gør sig gældende for the og tobaksordene længere nede.

```
kaffe_ord <- c("caffee", "caffe", "cafe", "café", "caffée", "coffee", "caffee-bord",
               "caffee-bønner", "caffee-drik", "caffee-drikken", "caffee-drikker",
               "caffee-huus", "kafé", "caffee-tand", "cafee", "caffeé")
kaffe_total <- filter(tekster_tidy, word %in% kaffe_ord)
```

Resultatet er 99 observationer af ordet kaffe på den ene eller anden måde. Kigger vi nærmere på de resultater, der er fundet i dataframen "kaffe_antal" ser vi, at computeren kun har fundet de steder hvor der er enkelte ord. Det skyldes vores data format, tidy-data princippet har brudt bindingerne op, således, at caffee-drikker ikke findes. istedet findes "caffee" og "drikker". I dette tilfælde tælles der altså en ekstra forekomst af "caffee". I kodestykket nedenfor vises de 99 fremkomster af ord der har noget med kaffe at gøre.

```
kaffe_total %>%
  count(word, sort = TRUE) -> kaffe_antal
kaffe_antal
```

```
##     word  n
## 1 caffee 62
## 2 caffée 12
## 3   café 10
## 4  caffe  7
## 5 coffee  6
## 6 caffeé  2
```

Hvis man vil forsøge at sætte antallet af gange kaffe er nævnt op imod hele Holbergs forfatterskab, kan man udregne hvor stor en procentdel ordene om kaffe udgør af hans samlede antal ord.

```
99/4844038*100
```

```
## [1] 0.002043749
```

Ordene om kaffe udgør altså 0.002043749% af Holbergs samlede forfatterskab.

Kontekst

Man kan spørge sig selv om ordene om kaffe udgør en specielt stor mængde af Holbergs forfatterskab. Dette kan imidlertid være en smule svært at svare på, da ord aldrig står alene. Konteksten rundt om ordene kan ikke fanges med denne metode. Derimod kan man sammenligne resultatet med undersøgelser af andre luksusvarer, her the og tobak.

```
the_ord <- c("thee", "the", "thée", "thé", "tee")
the_total <- filter(tekster_tidy, word %in% the_ord)
the_total %>%
  count(word, sort = TRUE) -> the_antal

tobak_ord <- c("tobak", "toback", "tobac", "tobach", "tabak",
               "tabac", "tobaks", "tabaks", "tabacs")
tobak_total <- filter(tekster_tidy, word %in% tobak_ord)
tobak_total %>%
  count(word, sort = TRUE) -> tobak_antal

# The procent:
182/4844038*100
```

```
## [1] 0.003757196
```

```
# Tobak procent:
130/4844038*100
```

```
## [1] 0.002683711
```

Her ses det, at Holberg nævner the og tobak henholdsvis 182 og 130 gange. De nævnes altså en smule mere end kaffe. Det svarer til at 0.003757196% af Holbergs forfatterskab er ord om the og 0.002683711% er ord om tobak.

Det kunne være interessant at se hvordan disse tal om luksusvarer ser ud i forhold til andre ting, som Ludvig Holberg skriver om. For at vi kan det, er vi nødt til at have en ide om hvilke ord, der bruges mest i Holbergs tekster.

Mest brugte ord hos Holberg

I det følgende stykke kode, ser vi hvilke ord der fremkommer hyppigst i Holbergs forfatterskab.

```
tekster_tidy %>%
  count(word, sort = TRUE) %>%
  top_n(150)
```

```
## Selecting by n
```

```
##           word      n
## 1           at 150521
## 2           og 146841
## 3           de  81189
## 4          som  75372
## 5          til  67634
## 6          den  67217
## 7           af  66020
## 8          det  61209
## 9          udi  59365
## 10          en  56583
```

```
## 11         han  51378
## 12           i  50194
## 13         med  42843
## 14          er  42363
## 15         sig  41878
## 16         for  40812
## 17         paa  39188
## 18        ikke  37846
## 19         jeg  33597
## 20         var  33297
## 21         saa  32495
## 22         men  28416
## 23       havde  26108
## 24          da  22874
## 25         man  22674
## 26          om  22674
## 27         der  22578
## 28        hans  21903
## 29         ved  21788
## 30          et  20497
## 31        blev  19470
## 32         ham  19017
## 33        kand  18739
## 34        have  16926
## 35         dem  16884
## 36       samme  16222
## 37         har  15215
## 38       deres  15056
## 39       denne  14548
## 40         thi  14350
## 41         sin  13413
## 42       eller  12832
## 43         end  12176
## 44       andre  12060
## 45       efter  12036
## 46       kunde  11928
## 47       dette  11911
## 48        alle  11526
## 49         mig  10662
## 50         fra  10643
## 51        hvad  10266
## 52       haver   9747
## 53       ingen   9649
## 54         dog   9510
## 55       ogsaa   9414
## 56        hvor   9125
## 57         mod   9077
## 58        være   9068
## 59         hun   9006
## 60      saadan   8988
## 61        hand   8915
## 62         ere   8744
## 63        over   8648
## 64      skulde   8612
```

```
## 65        vare   8476
## 66        uden   8366
## 67       anden   7985
## 68         maa   7947
## 69          nu   7900
## 70        naar   7882
## 71       vilde   7828
## 72        selv   7827
## 73         aar   7809
## 74      saasom   7762
## 75         vil   7624
## 76         min   7505
## 77        skal   7468
## 78       igien   7423
## 79       været   7345
## 80     hvilken   7301
## 81       store   7275
## 82      kongen   7163
## 83     alleene   6923
## 84        stor   6916
## 85        hvis   6681
## 86       intet   6595
## 87    saaledes   6529
## 88         lod   6468
## 89       nogle   6232
## 90           u   6195
## 91        kong   6179
## 92       disse   6176
## 93       giøre   6104
## 94     hvilket   6045
## 95        folk   6042
## 96     saadant   6021
## 97         vel   5980
## 98       bleve   5953
## 99       mange   5915
## 100    efterdi   5911
## 101      andet   5724
## 102       tiid   5660
## 103         du   5575
## 104         vi   5530
## 105         in   5500
## 106      under   5495
## 107      siden   5470
## 108       mand   5425
## 109        sit   5386
## 110          3   5231
## 111      nogen   5142
## 112     første   5132
## 113       sine   5083
## 114      meget   5069
## 115        see   4979
## 116        her   4953
## 117       sige   4929
## 118      konge   4927
```

```
## 119        hos   4911
## 120        een   4908
## 121 adskillige   4839
## 122          2   4711
## 123     maatte   4547
## 124     derfor   4541
## 125      blive   4535
## 126       seer   4444
## 127      strax   4440
## 128     hvilke   4425
## 129     danske   4353
## 130      heele   4285
## 131        kom   4142
## 132      komme   4123
## 133      gamle   4120
## 134          a   4057
## 135      noget   3973
## 136       navn   3925
## 137      meere   3868
## 138      siger   3829
## 139       ting   3829
## 140         ey   3821
## 141     derpaa   3820
## 142        tid   3764
## 143     giorde   3680
## 144      hende   3625
## 145      blant   3608
## 146      giort   3557
## 147   historie   3532
## 148      herre   3530
## 149       icke   3526
## 150        ord   3515
```

For det første kan vi se en masse stopord, men hvis man bladrer i listen ser man også ord som konge og historie, der er brugt henholdsvis 4927 og 3532 gange. Nedenfor udregnes hvor stor en procentdel af forfatterskabet disse to ord fylder.

```
# procent for konge
4927/4844038*100
```

```
## [1] 0.1017127
```

```
# procent for historie
3532/4844038*100
```

```
## [1] 0.07291437
```

Resultatet er her, at selve ordet konge udgør 0.1017127% af Holbergs forfatterskab, mens ordet historie udgør 0.07291437% af forfatterskabet. Begge ord bruges altså væsentligt mere, end hele ordforrådet om luksusvarerne.

Datasæt uden stopord

Her vil jeg forsøge, at fjerne stopordene fra datasættet, så det, som er vigtigt for dataen nemmere kan ses i visualiseringer. Den stopordsliste der bruges, er en moderne dansk liste, hvor der er tilføjet de 150 mest brugte stopord fra Holbergs forfatterskab.

```
stopord <- read_csv("stopord.txt")
tekster_tidy %>%
  anti_join(stopord, by = "word") %>%
  count(word, sort = TRUE) %>%
  select(word, n) -> tekster_tidy_nostops
```

Ordet "caffee" indtager en 5728'ne plads ud af de 161330 unikke ord, som Holberg har skrevet, der ikke er stopord.

```
slice(tekster_tidy_nostops, 5728)
```

```
##     word  n
## 1 caffee 62
```

Visualiseringer

I dette afsnit kommer Holbergs forfatterskab frem på en anden måde end hvis man læste det som vi normalt gør. Jeg vil forsøge, at lave forskellige visualiseringer, der kan sige noget om hele hans forfatterskab.

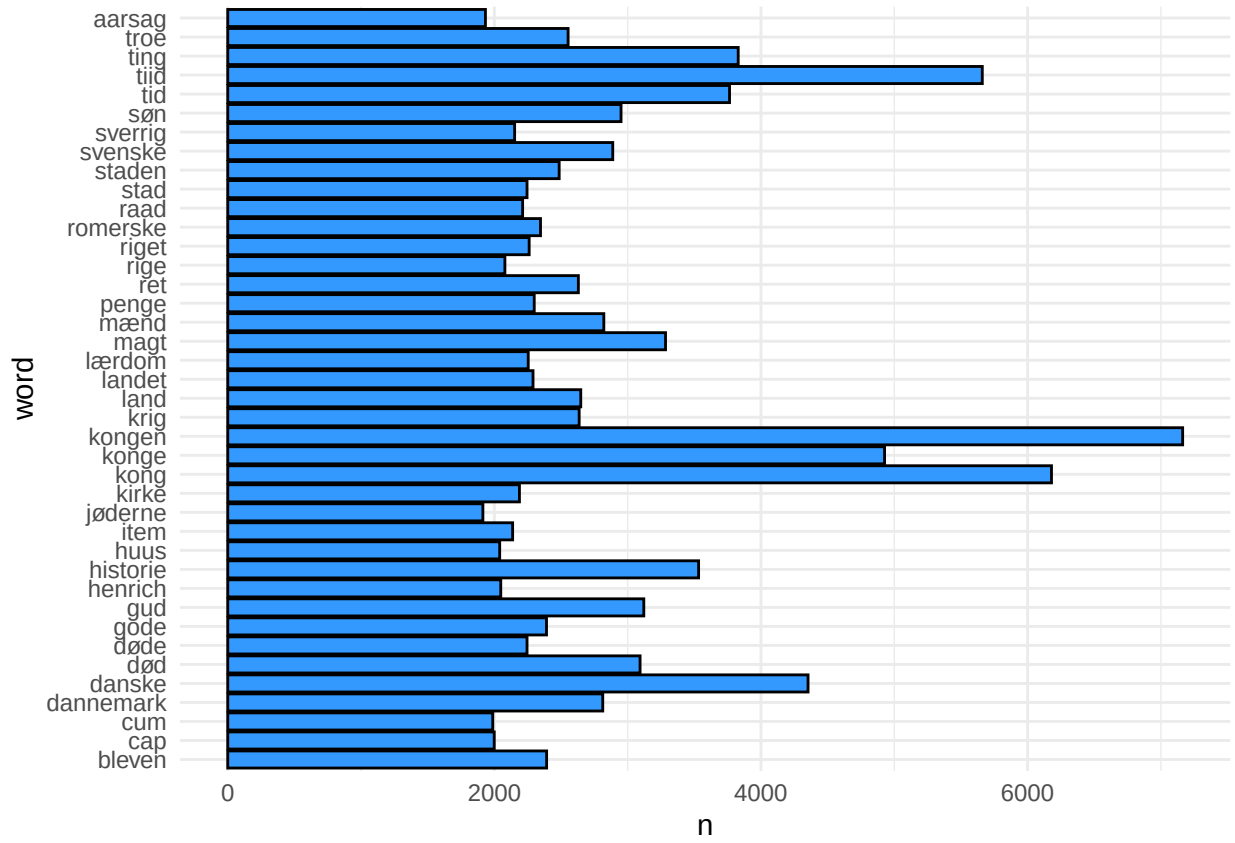
```
tekster_tidy_nostops %>%
  top_n(40) %>%
  ggplot(aes(label = word, size = n, color = n)) +
  geom_text_wordcloud() +
  scale_size_area(max_size = 10) +
  theme_minimal() +
  scale_color_gradient(low = "blue", high = "red")
```

```
## Selecting by n
```

I denne wordcloud ses de 40 mest brugte ord i Holbergs forfatterskab. Jo større ordet er og jo mere rødt det er, betyder at ordet er brugt meget. Hvis man hellere vil have nogle tal på, kan man få dataen frem i et histogram, som gøres i kodeblokken nedenunder.

```
tekster_tidy_nostops %>%
  top_n(40) %>%
  ggplot(aes(x = word, y = n)) +
  geom_col(fill="#3399FF", colour="black") +
  coord_flip() +
  theme_minimal()
```

```
## Selecting by n
```

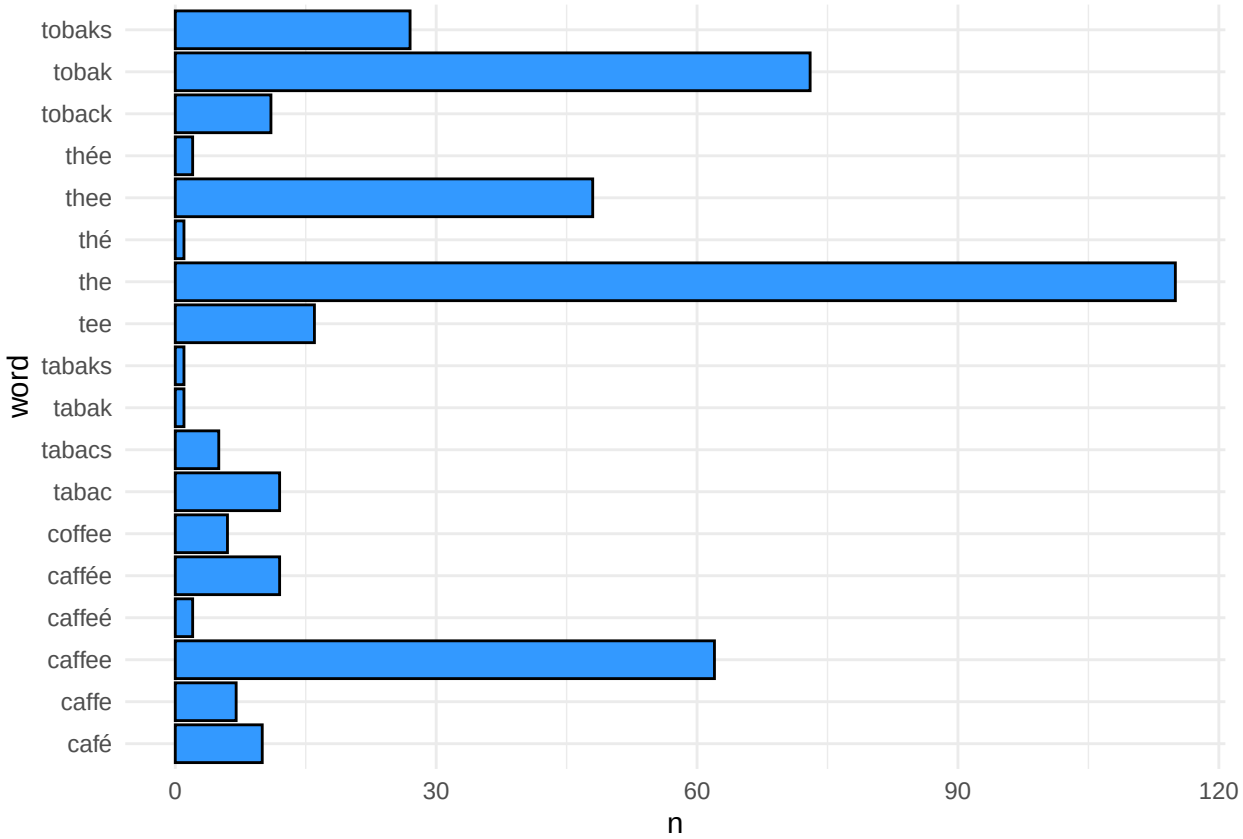


Ovenfor ses de 40 mest brugte ord i Holbergs forfatterskab på en anden måde. Her vises dataen i et histogram, hvor man også kan få en ide om, hvor mange gange det enkelte ord har været brugt. I de næste visualiseringer sættes luksusvarerne fra Holbergs forfatterskab op i forskellige visualiseringer.

Visualisering af luksus

I de følgende visualiseringer vises, hvor meget Holberg har skrevet om de forskellige luksusvarer.

```
luksusvarer <- rbind(kaffe_antal, the_antal, tobak_antal)
luksusvarer %>%
  ggplot(aes(x = word, y = n)) +
  geom_col(fill="#3399FF", colour="black") +
  coord_flip() +
  theme_minimal()
```



Dette histogram viser brugen af de forskellige stavemåder, det ville måske give mere mening som en word-cloud, hvor man hurtigt får et overblik over, hvad der bruges mest.

```
luksusvarer <- arrange(luksusvarer, desc(n))
luksusvarer %>%
  ggplot(aes(label = word, size = n, color = n)) +
  geom_text_wordcloud() +
  scale_size_area(max_size = 20) +
  theme_minimal() +
  scale_color_gradient(low = "black", high = "red")
```

Her står det altså klart, at ordet “the” blev brugt mest at ordene om kaffe, the og tobak. The blev efterfulgt, af “tobak” som nr to og “caffee” som nr tre.

Referencer

Denne analyse er lavet på baggrund af materiale fra Det Danske Sprog- og Litteraturselskab og Universitetet i Bergen.
Hvis man er interesseret i at dykke ned i Holbergs forfatterskab ligger det tilgængeligt på: http://holbergsskrifter.dk/
Analysen har også gjort brug af Holbergordbogen, der kan findes på: https://holbergordbog.dk