# Trends in grammatical tense of the English language: irregular vs regular verbs

*Artyom Stepanov, Yulia Kolomenskaya*

*13.06.18*

## Contents



### Materials

Link to the data set (csv file)
Link to the additional data set (csv file)

### Introduction

In modern English language we can seemingly more often find regular forms of the traditionally irregular verbs used in various types of speech. This phenomenon can be explained by the language natural tendency to grammaticalize (and thus adapt) irregular forms, especially frequently used ones. In our research we aimed to analyze the balance in usage of regular/irregular forms of one and the same verb.

The previous research of the correlation between regular and irregular verb forms seems to be more focused on the neurolinguistic aspect of this phenomenon such as difference in acquisition and mental processing of these forms rather than on pure linguistic study of the shift from irregular to regular form for one and the same initially irregular verb.

### Research hypothesis

Our research hypothesis is based on the assumption that the percentage of irregular form usage decreases over time and might also depend on the genre it is used in. Thus the null hypothesis will be that there is no correlation between the choice of the form type and the factors listed above (time and text genre).

### Data

The dataset used in this project was collected from the two well-known English language corpora: BNC (British National Corpus) and COCA (Corpus of Contemporary American English). The search through these corpora was based on the list of the most frequently used irregular English verbs. The final dataset contains information on the source, date, genre, verb type, verb form and context sentence.

- Dependent variable: 'Normalized frequency' = Relative value; it is calculated in a separate dataset for each 'genre - year' combination
- Predictor variables: 'Date' - numeric; year of the publication 'Genre' - categorical; 'SPOK' - spoken, 'ACPROSE' - academic prose, 'NONAC' - non-academic prose, 'OTHERPUB' - other publications (includes magazine publications from COCA), 'FICTION', 'NEWS'
- Number of observations is 49416 in total

#### Data collection and annotation

The main challenge in data annotation was to create a universal genre classification based on the division initially provided by the corpora. Thus 'ACADEMIC' ''NEWS' and "FICTION' are found in both BNC and COCA, whereas 'SPOKEN' is found in COCA, but not in BNC. Furthermore we united 'MAG' (magazine) in COCA and 'OTHERPUB' in BNC.

Another issue lies in the corpora date misalignment: while BNC includes texts for the period 1970s-1993, COCA contains materials for 1990-2017 time period. This time discrepancy prevents us from taking language origin (American vs British) as another predictor variable. Though we can check the dynamics for American and British versions of English separately.

```
data=read.csv("/Users/juliakolomenskaya/Downloads/fin_verbs.csv")
data=data[-1]

#We convert numeric values to categorical ones:
copy=data
data$Type[data$Type==1]='irregular'
data$Type[data$Type==0]='regular'
```

## R libraries in use

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.2.1 --

## √ ggplot2 2.2.1     √ purrr   0.2.4
## √ tibble  1.4.2     √ dplyr   0.7.4
## √ tidyr   0.8.0     √ stringr 1.3.1
## √ readr   1.1.1     √ forcats 0.2.0

## Warning: package 'stringr' was built under R version 3.4.4

## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidyverse)
library(plotly)
```

```
##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```r
library(lme4)
```

```
## Warning: package 'lme4' was built under R version 3.4.4
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following object is masked from 'package:tidyr':
##
##     expand
```

```r
library(ggfortify)
```

```
## Warning: package 'ggfortify' was built under R version 3.4.4
```

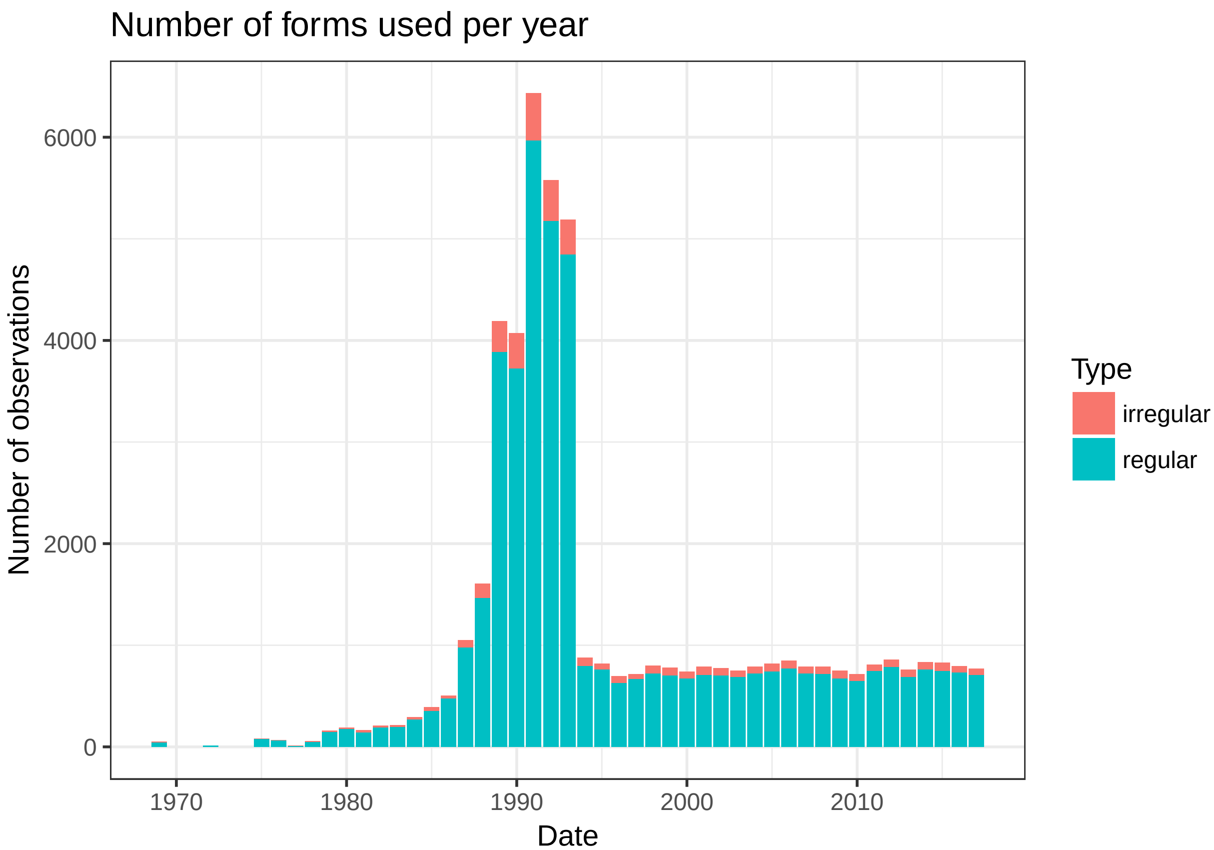
```r
library(Rtsne)
library(FactoMineR)
```

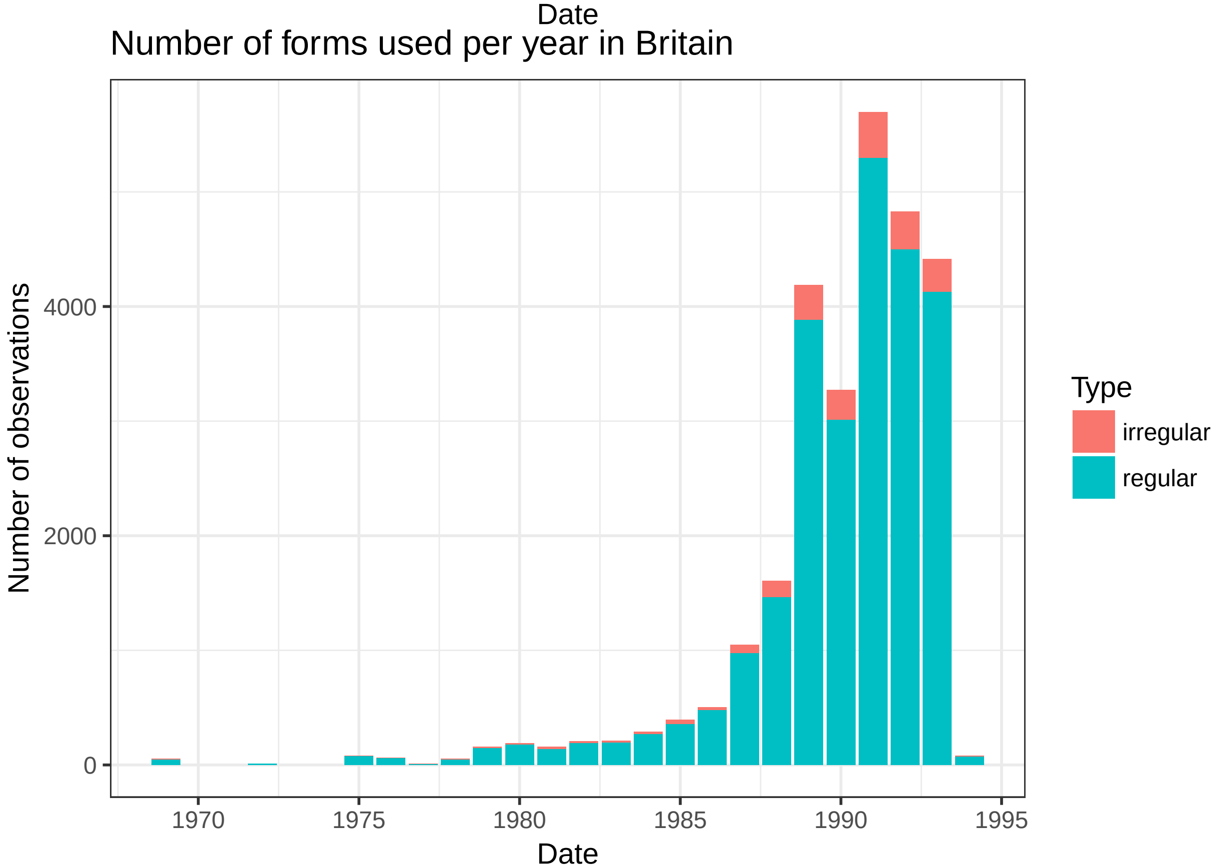
```
## Warning: package 'FactoMineR' was built under R version 3.4.4
```

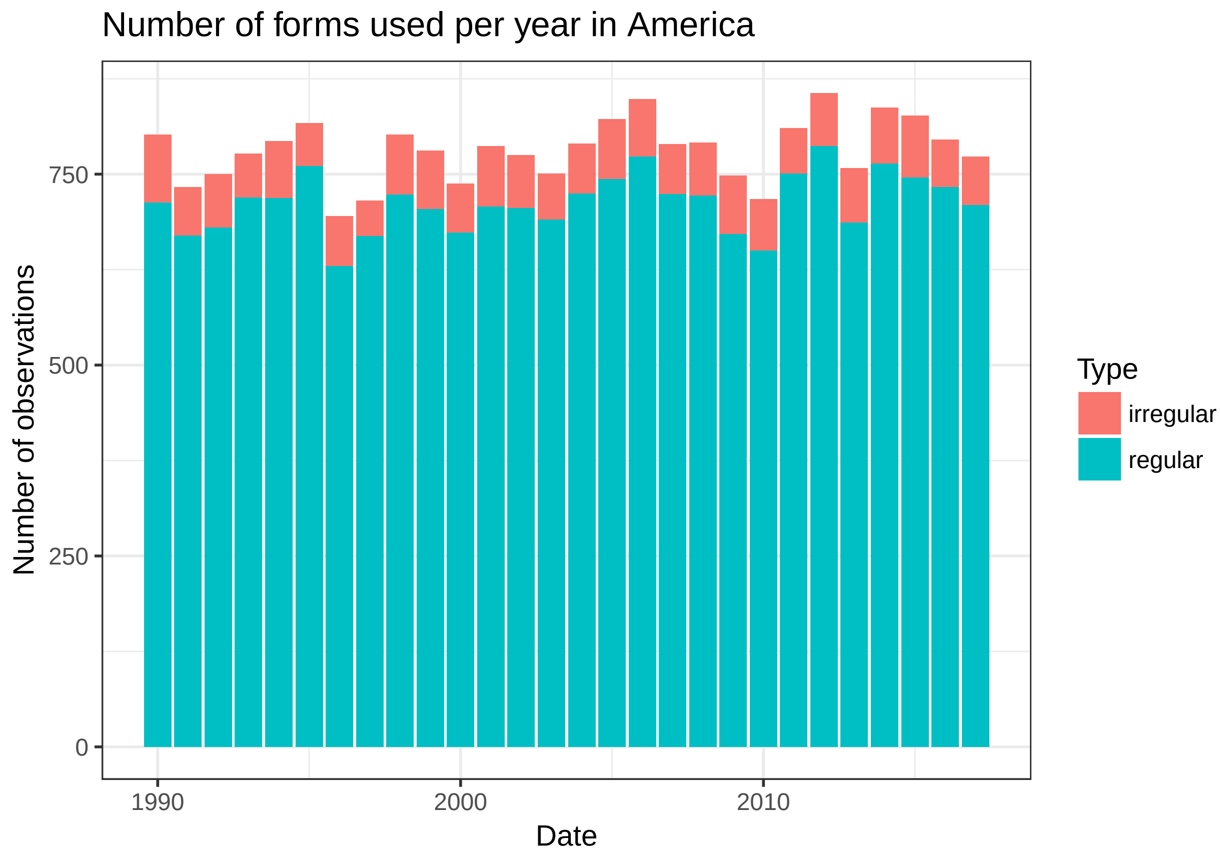
```r
# include R libraries here or later
```

## Analysis: descriptive statistics

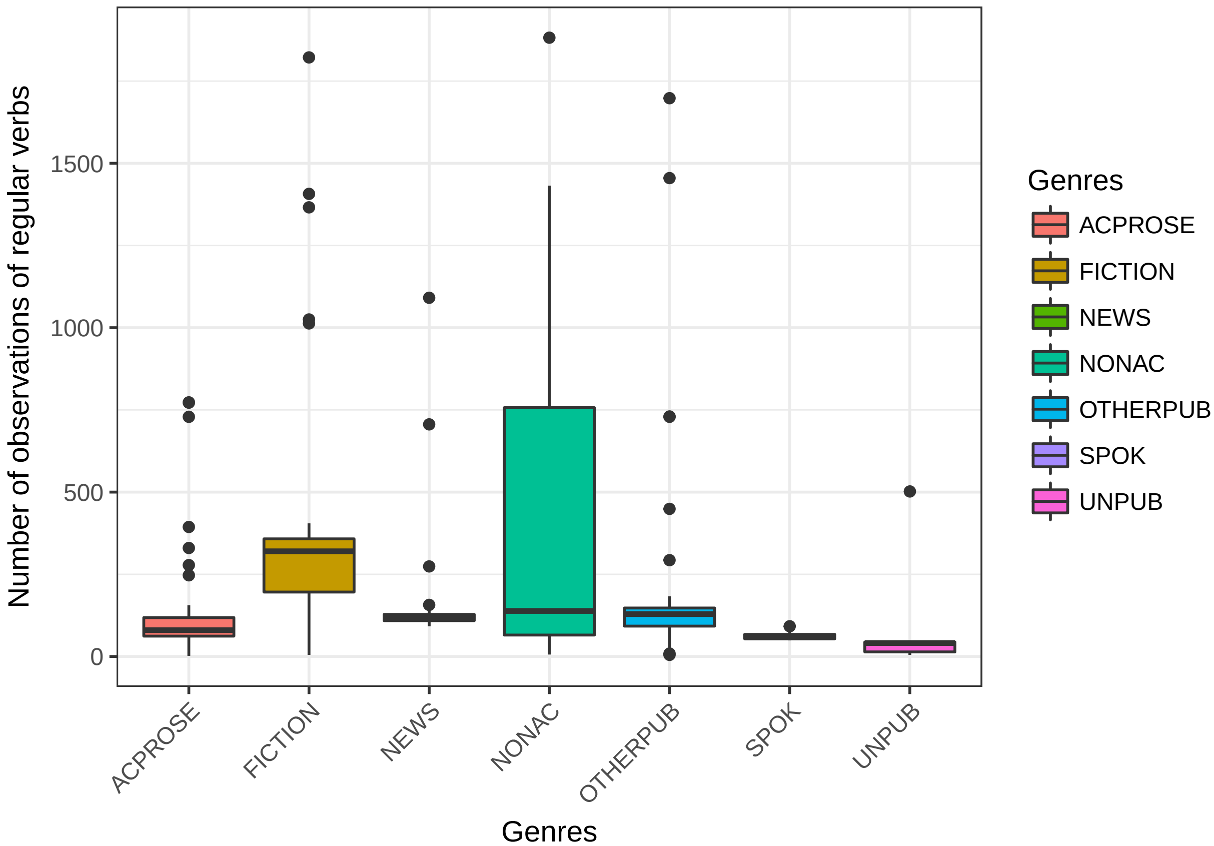
Here we show, how our data is distributed across the given time period (we provide overall results and the results for each of the corpora):


Number of forms used per year
Number of observations
0
2000
4000
6000
1970
1980
1990
2000
2010
Date
Type
irregular
regular

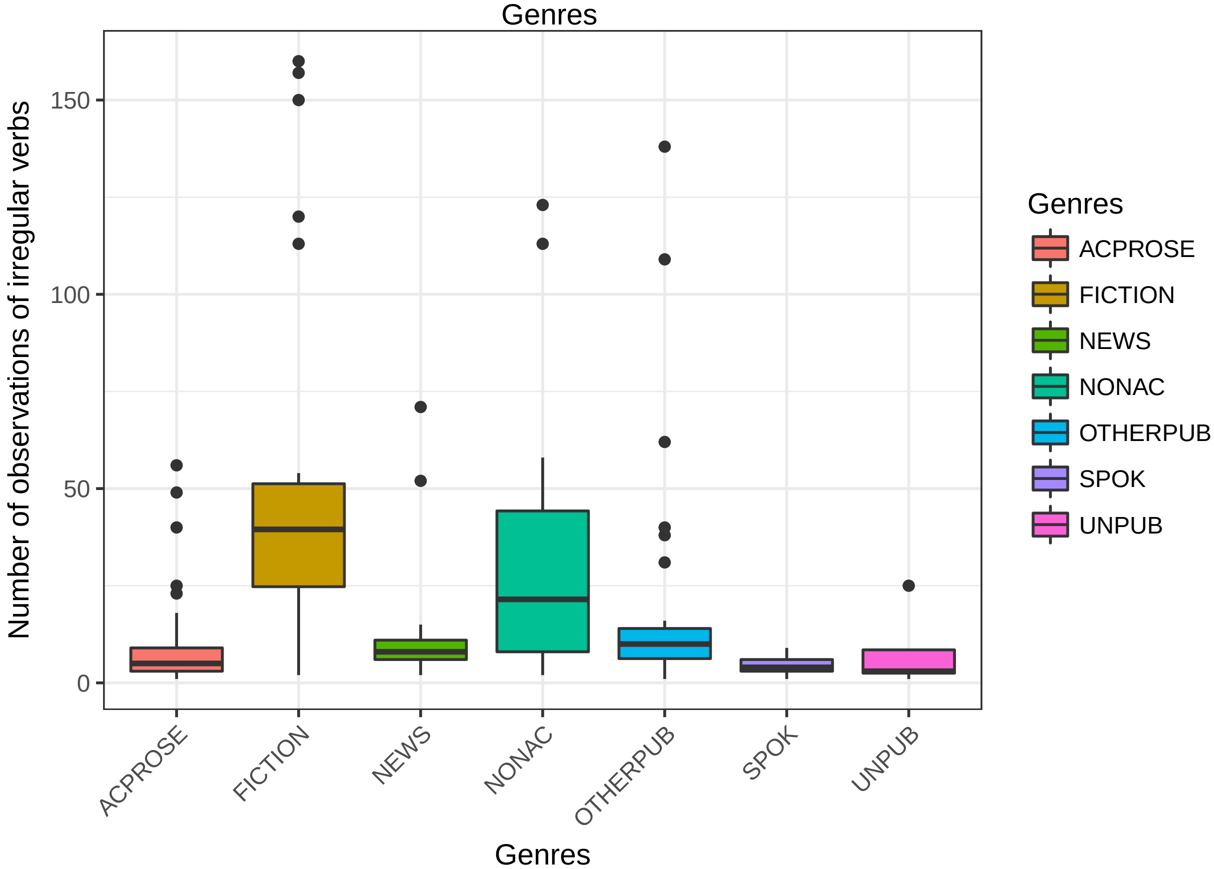

Number of forms used per year in Britain
Number of observations
0
2000
4000
1970
1975
1980
1985
1990
1995
Date
Type
irregular
regular




Let's have a look at the distribution of observations by genres:



Number of observations of regular verbs
0
500
1000
1500
ACPROSE
FICTION
NEWS
NONAC
OTHERPUB
SPOK
UNPUB
Genres
Genres
ACPROSE
FICTION
NEWS
NONAC
OTHERPUB
SPOK
UNPUB

Number of observations of irregular verbs
0
50
100
150
ACPROSE
FICTION
NEWS
NONAC
OTHERPUB
SPOK
UNPUB
Genres
Genres
ACPROSE
FICTION
NEWS
NONAC
OTHERPUB
SPOK
UNPUB

```r
#First, we provide statistics for the data, where date is taken into account:
by_year_br %>% summarise(min=min(n_observations),max=max(n_observations),mean=mean(n_observations),
                         median=median(n_observations),iqr=IQR(n_observations),sd=sd(n_observations))
```

```
## # A tibble: 22 x 7
##     Date   min   max   mean median   iqr     sd
##    <int> <dbl> <dbl>  <dbl>  <dbl> <dbl>  <dbl>
##  1  1969  5.00  47.0   26.0   26.0  21.0   29.7
##  2  1972 10.0   10.0   10.0   10.0   0      NA
##  3  1975  4.00  78.0   41.0   41.0  37.0   52.3
##  4  1976  2.00  62.0   32.0   32.0  30.0   42.4
##  5  1977  2.00  11.0    6.50   6.50  4.50   6.36
##  6  1978  5.00  51.0   28.0   28.0  23.0   32.5
##  7  1979 11.0  147     79.0   79.0  68.0   96.2
##  8  1980 10.0  178     94.0   94.0  84.0  119
##  9  1981 21.0  141     81.0   81.0  60.0   84.9
## 10  1982 16.0  191    104    104    87.5  124
## # ... with 12 more rows
```

```r
by_year_am %>% summarise(min=min(n_observations),max=max(n_observations),mean=mean(n_observations),
                         median=median(n_observations),iqr=IQR(n_observations),sd=sd(n_observations))
```

```
## # A tibble: 28 x 7
##     Date   min   max  mean median   iqr    sd
##    <int> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
##  1  1990  89.0   713   401    401   312   441
##  2  1991  63.0   670   366    366   304   429
##  3  1992  70.0   680   375    375   305   431
##  4  1993  57.0   720   388    388   332   469
##  5  1994  74.0   719   396    396   322   456
##  6  1995  56.0   761   408    408   352   499
##  7  1996  65.0   630   348    348   282   400
##  8  1997  46.0   669   358    358   312   441
##  9  1998  78.0   724   401    401   323   457
## 10  1999  76.0   705   390    390   314   445
## # ... with 18 more rows
```

```r
#Next, we do the same thing for genres:
genres %>% summarise(min=min(n_observations),max=max(n_observations),mean=mean(n_observations),
                         median=median(n_observations),iqr=IQR(n_observations),sd=sd(n_observations))
```

```
## # A tibble: 44 x 7
##     Date   min   max  mean median   iqr    sd
##    <int> <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl>
##  1  1969  5.00  5.00  5.00   5.00  0    NA
##  2  1975  1.00  3.00  2.00   2.00  1.00  1.41
##  3  1976  2.00  2.00  2.00   2.00  0    NA
##  4  1977  2.00  2.00  2.00   2.00  0    NA
##  5  1978  1.00  4.00  2.50   2.50  1.50  2.12
##  6  1979  2.00  9.00  5.50   5.50  3.50  4.95
##  7  1980  1.00  7.00  3.33   2.00  3.00  3.21
##  8  1981  4.00 12.0   7.00   5.00  4.00  4.36
##  9  1982  1.00 11.0   5.33   4.00  5.00  5.13
## 10  1983 16.0  16.0  16.0   16.0   0    NA
## # ... with 34 more rows
```

Let's see if there is a correlation between various periods of times of the relative value of irregular form usage. We convert our data with respect to time into a new data with relative values of irregular forms usage:

```
by_year_df=data.frame(by_year)
genres=data.frame(genres)

time_dist=data.frame("Year"=c(),"Relative value"=c())
for (i in 1:44){
  time_dist[i,'Year']=unique(by_year_df[by_year$Type=='irregular','Date'])[i]
  time_dist[i,'Relative value']=by_year[(by_year$Date==time_dist[i,"Year"]) & (by_year$Type=='irregular
}
```

And then we want to see if there's a correlation between time periods and percentage of irregular forms used in these periods (as our hypothesis implies hiearchy of time periods,we use Kendall's correlation):

```
cor(time_dist,method='kendall')
```

```
##                      Year Relative value
## Year           1.0000000      0.1902748
## Relative value 0.1902748      1.0000000
```

Since the correlation is not equal to zero, we assume, that there is a connection between these two variables. Now, we want to investigate if the dependence between time and relative value can be approximated by linear regression model:

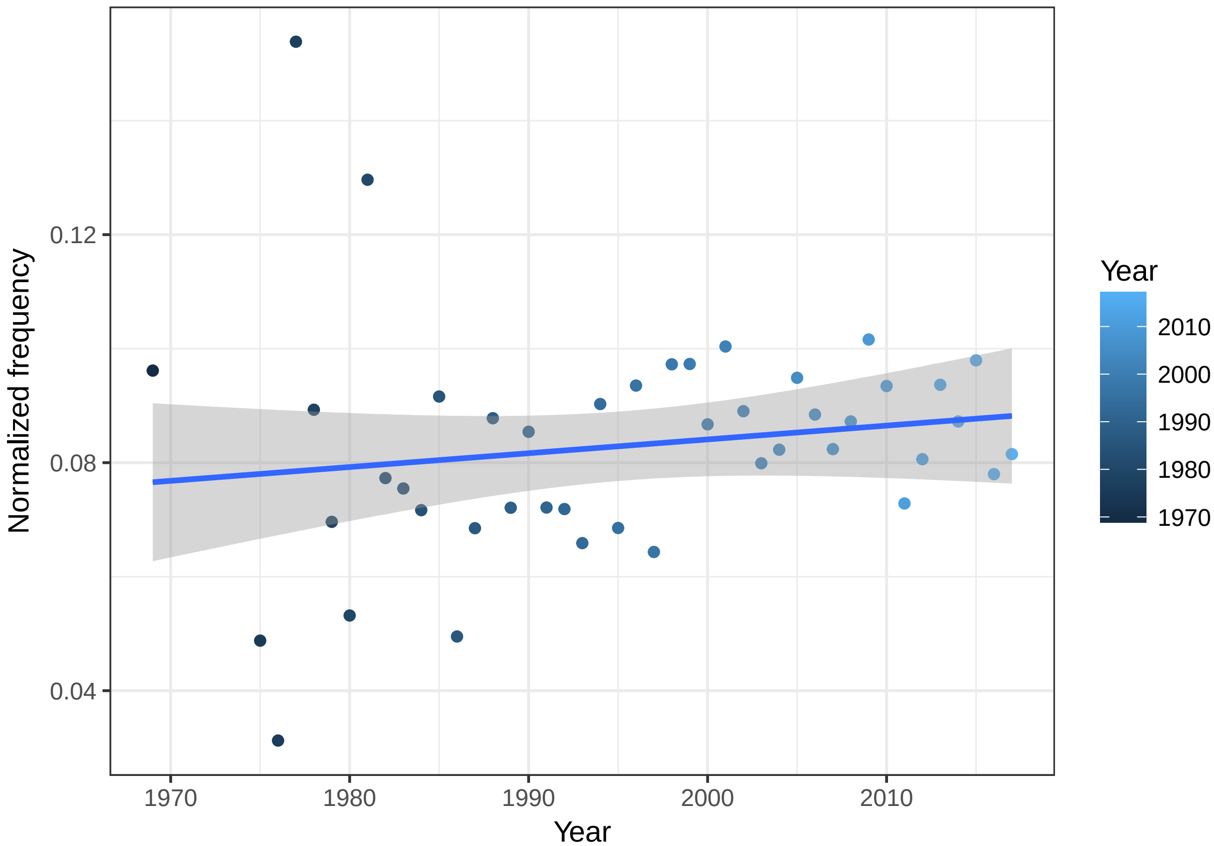
```
fit=lm(time_dist$`Relative value`~time_dist$Year,data=time_dist)
summary(fit)
```

```
##
## Call:
## lm(formula = time_dist$`Relative value` ~ time_dist$Year, data = time_dist)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.047016 -0.009820 -0.001339  0.009758  0.075338
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.4004062  0.4663608  -0.859    0.395
## time_dist$Year  0.0002422  0.0002337   1.036    0.306
##
## Residual standard error: 0.02002 on 42 degrees of freedom
## Multiple R-squared:  0.02494,    Adjusted R-squared:  0.001725
## F-statistic: 1.074 on 1 and 42 DF,  p-value: 0.3059
```

```
time_dist$model=predict(fit)
```

Let's visualize it:

```
time_dist %>% ggplot(aes(Year,`Relative value`))+geom_point(aes(color=Year
                                                              ))+geom_line(aes(Year,model))+geom_smoo
```



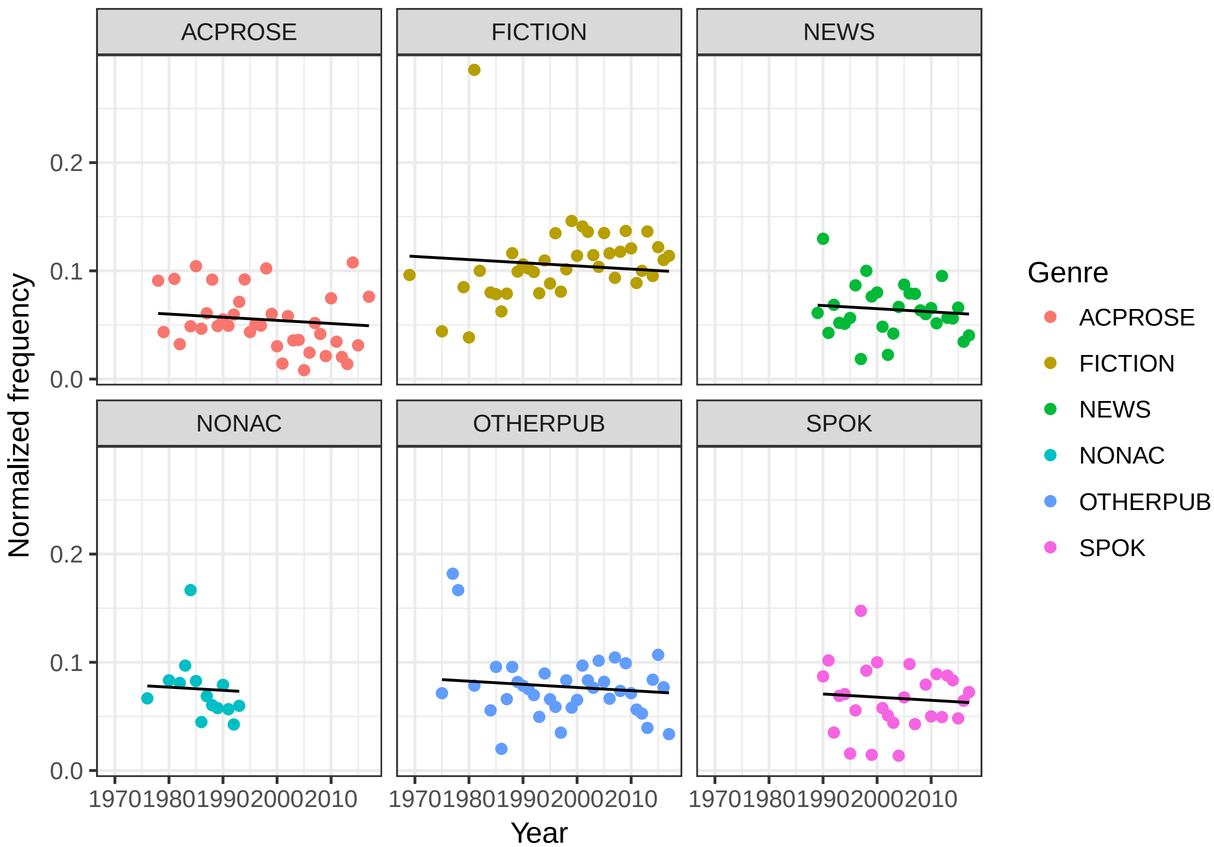

The visualisation here shows us, that the linear approximation for the two variables under consideration reveals an increase in the use of irregular forms over the given time periods. Now, let's try to use some advanced models. Namely, we use mixed-effect models with genre being a random effect:

```
genre_dist=read.csv("/Users/juliakolomenskaya/Downloads/time_genre.csv")
genre_dist=genre_dist[-1]
genre_dist=genre_dist[genre_dist$Genre!='UNPUB',]

fit2=lmer(Rel.value~Year+(1|Genre),data=genre_dist)
summary(fit2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Rel.value ~ Year + (1 | Genre)
##    Data: genre_dist
##
## REML criterion at convergence: -728.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.3453 -0.5519 -0.0787  0.4913  5.7274
##
## Random effects:
##  Groups   Name        Variance  Std.Dev.
##  Genre    (Intercept) 0.0003233 0.01798
##  Residual             0.0009403 0.03066
## Number of obs: 185, groups:  Genre, 6
##
## Fixed effects:
```

```
##                 Estimate Std. Error t value
## (Intercept)  0.6557985  0.4302095   1.524
## Year        -0.0002913  0.0002153  -1.353
##
## Correlation of Fixed Effects:
##      (Intr)
## Year -1.000
```

```
genre_dist$model=predict(fit2) #We use Genre as intercept term

genre_dist %>% ggplot(aes(Year,Rel.value))+geom_point(aes(color=Genre))+geom_line(aes(Year,model))+ylab
```




As opposed to linear model, the approximation by mixed-effect model produces the results, contradicting the linear model: as we see, there's a negative dynamics in the use of irregular forms.

## Multi-factor analysis

### PCA:

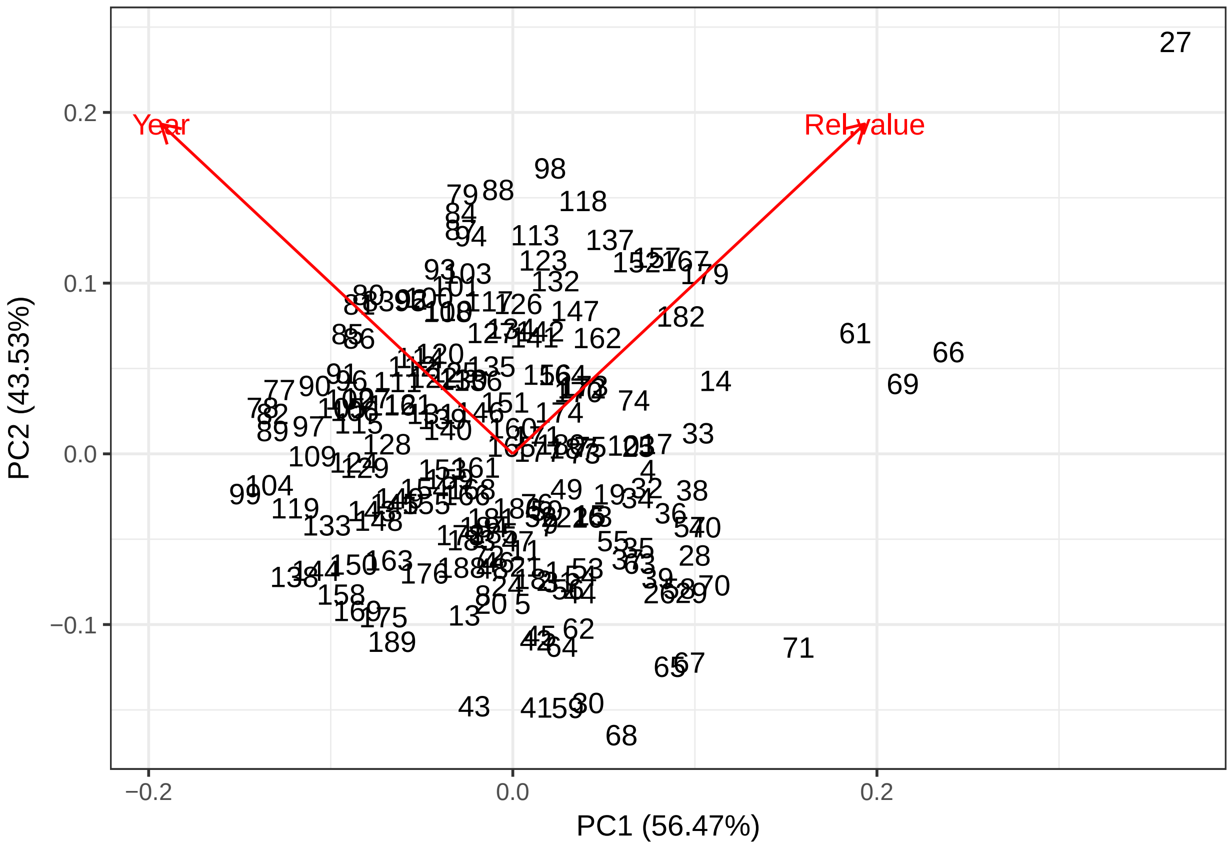
We try to produce 2 dimensions that explain as much variance as possible, using our numeric data:

```
PCA=prcomp(genre_dist[,2:3], center = TRUE, scale. = TRUE)
summary(PCA)
```

```
## Importance of components:
##                           PC1    PC2
## Standard deviation     1.0628 0.9330
## Proportion of Variance 0.5647 0.4353
```

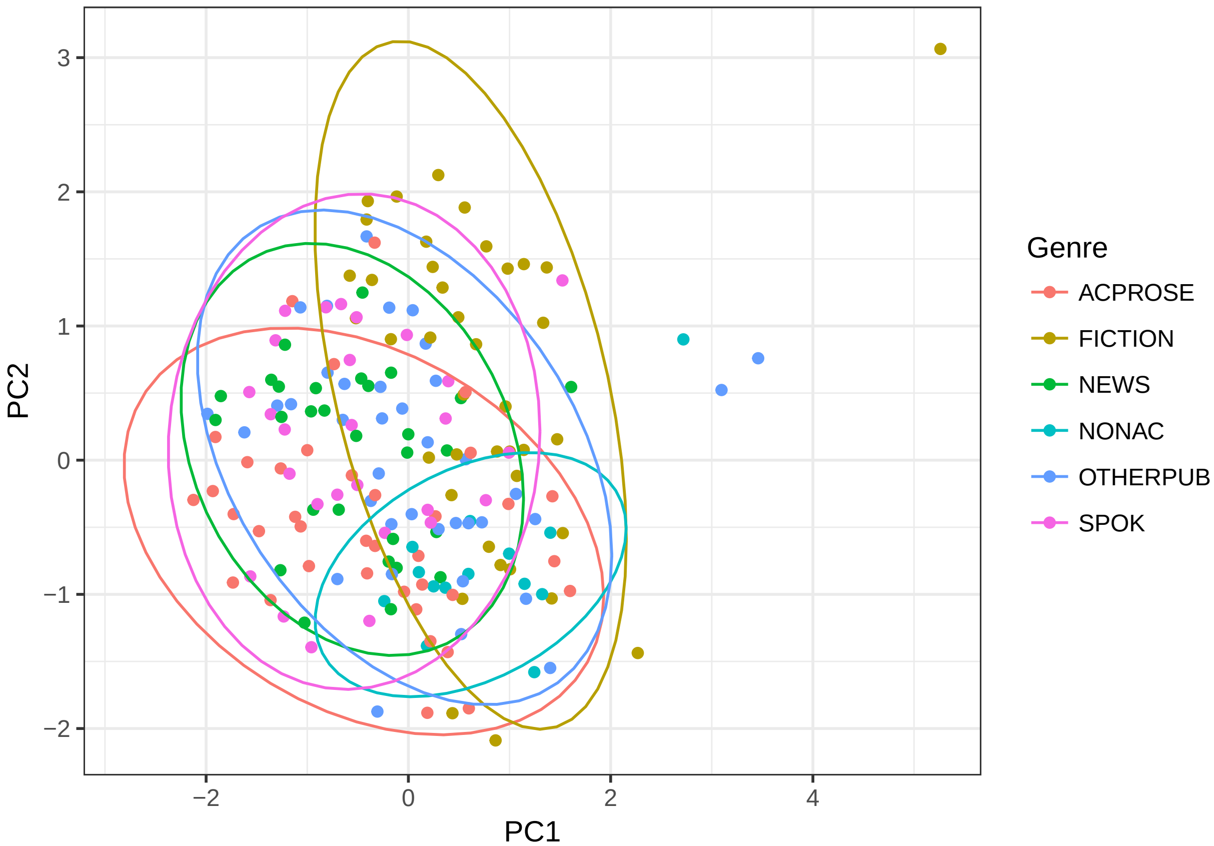
```
## Cumulative Proportion  0.5647 1.0000
```

```r
autoplot(PCA,
         shape = FALSE,
         loadings = TRUE,
         label = TRUE,
         loadings.label = TRUE)+
  theme_bw()
```




Next, we try to produce visualisation to reveal hidden cluster structures in our data:

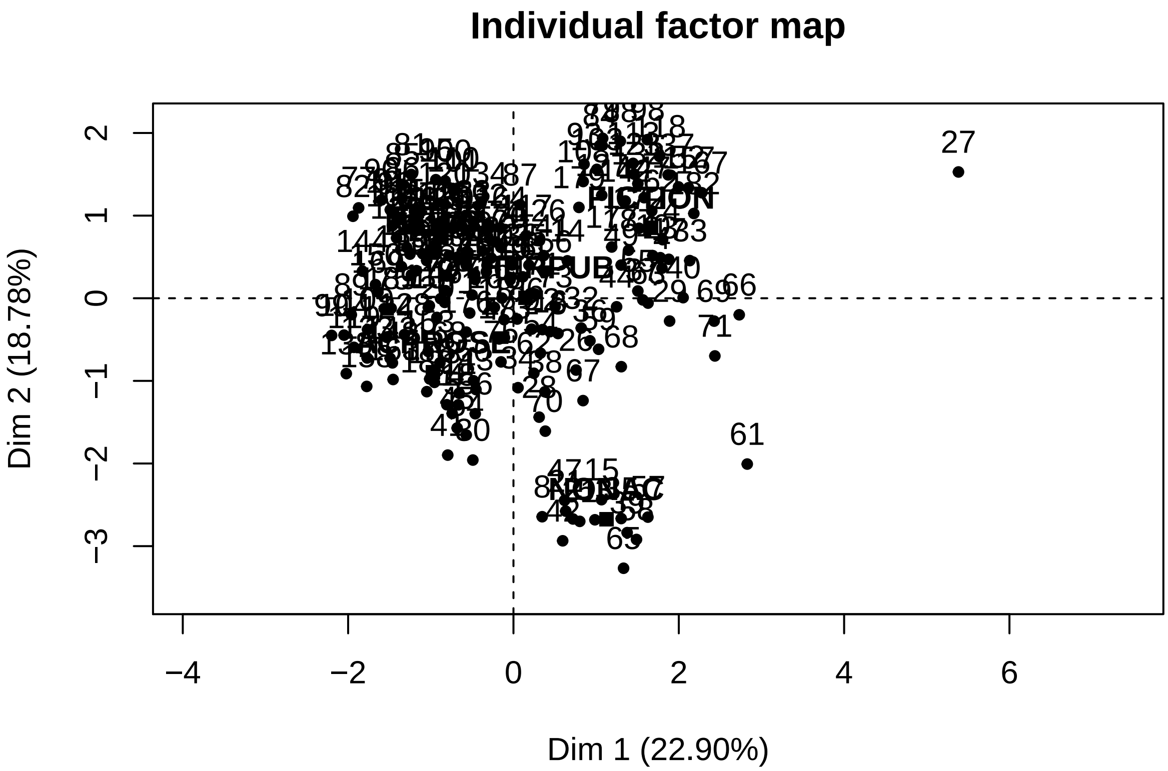
```r
genre_dist=cbind(genre_dist, PCA$x)

genre_dist %>%
  ggplot(aes(PC1, PC2, color = Genre))+
  geom_point()+
  stat_ellipse()+
  theme_bw()
```

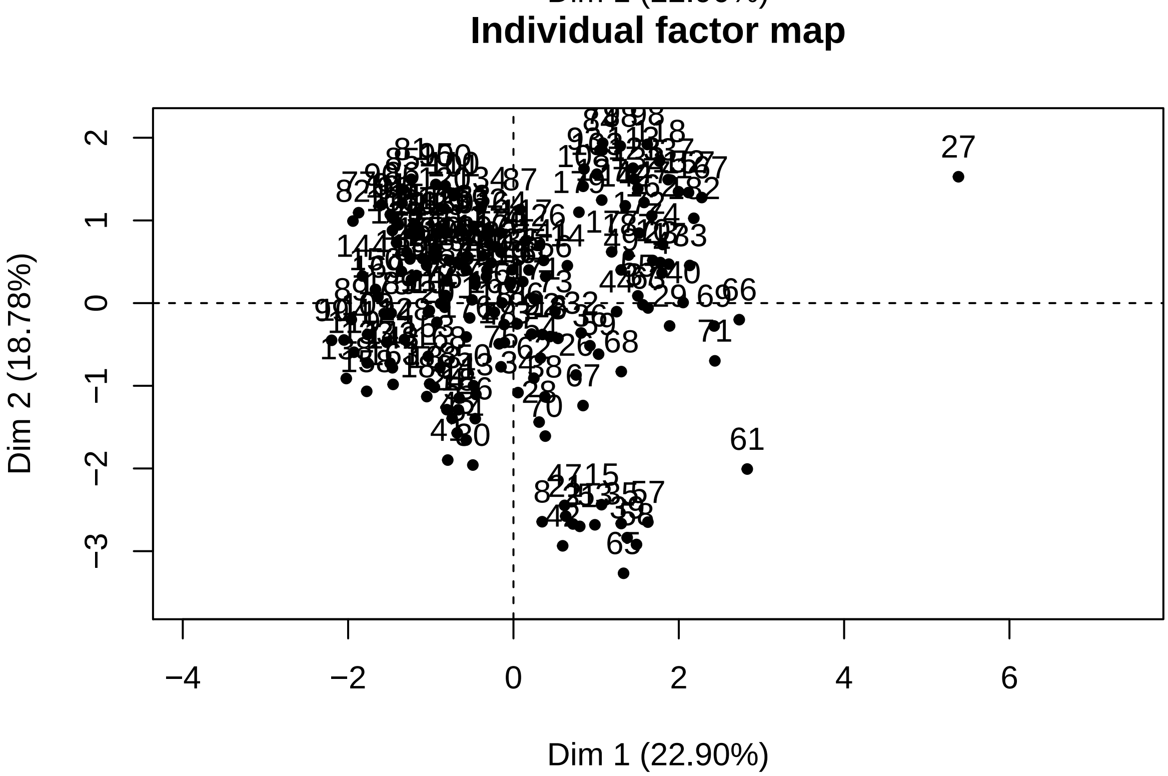


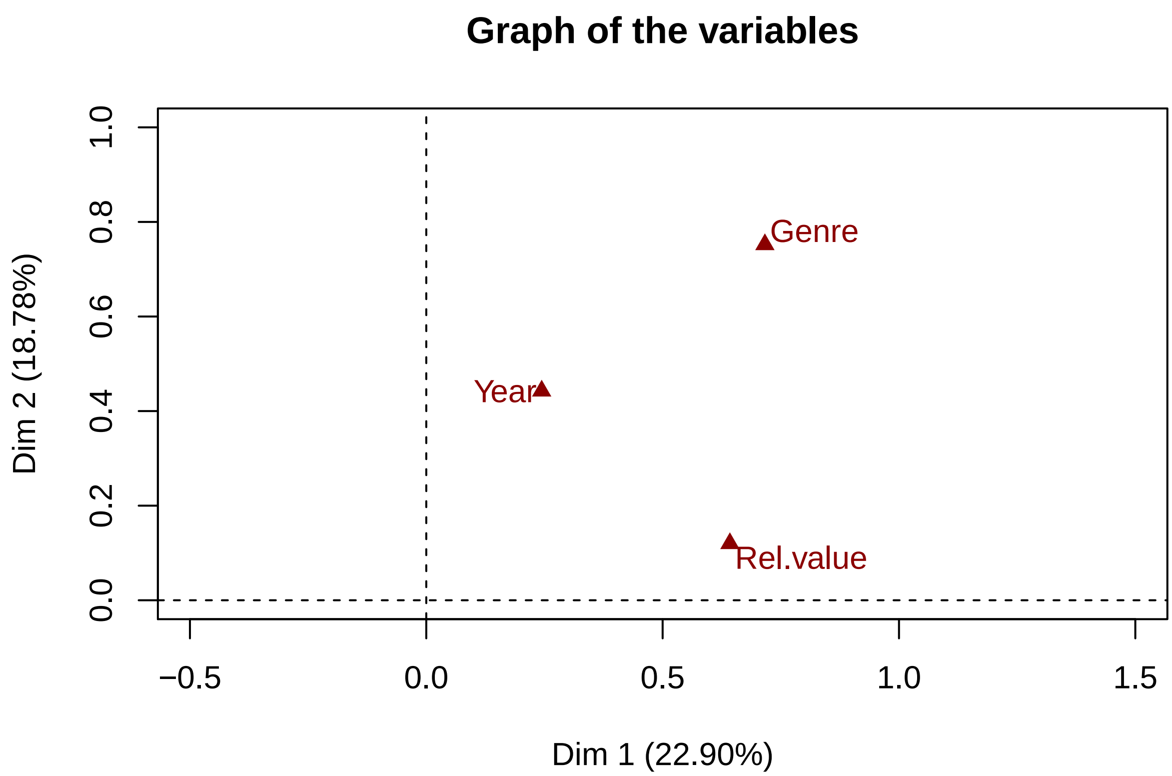

**FAMD:**

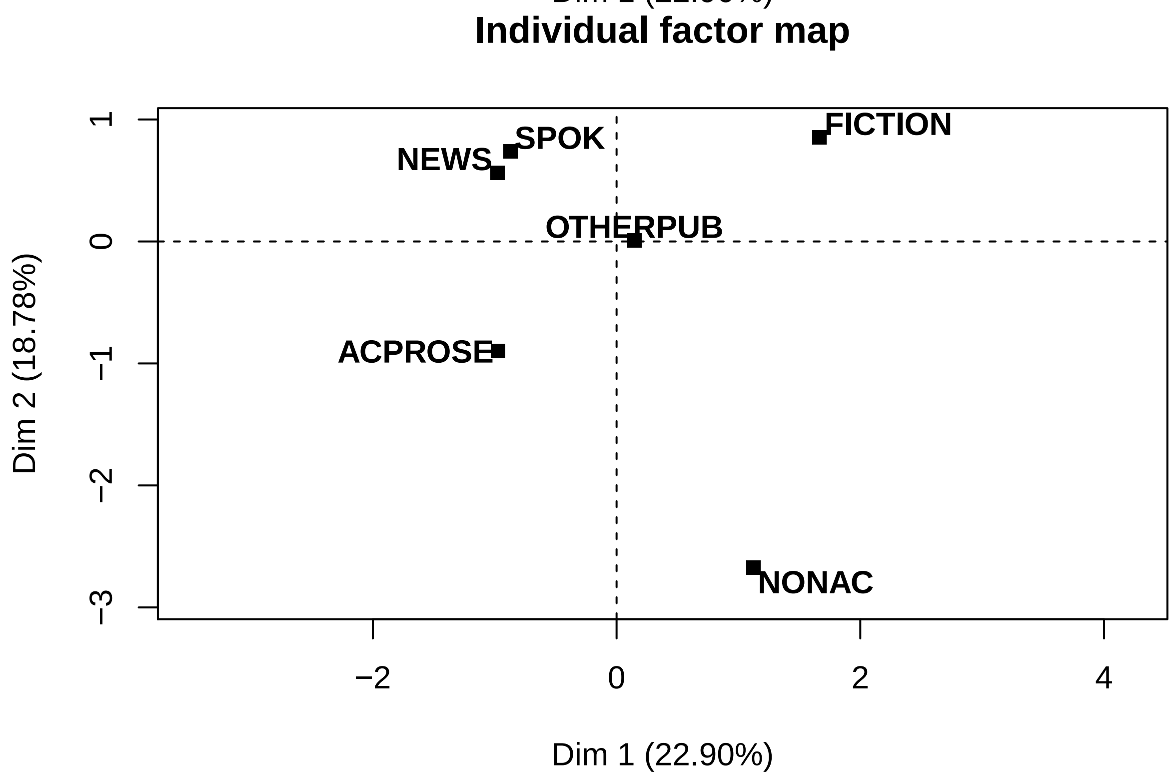
Finally, since our data consists of both the categorical and numeric variables, we use Factor Analysis of Mixed Data.

```
famd=FAMD(genre_dist[,1:3])
```

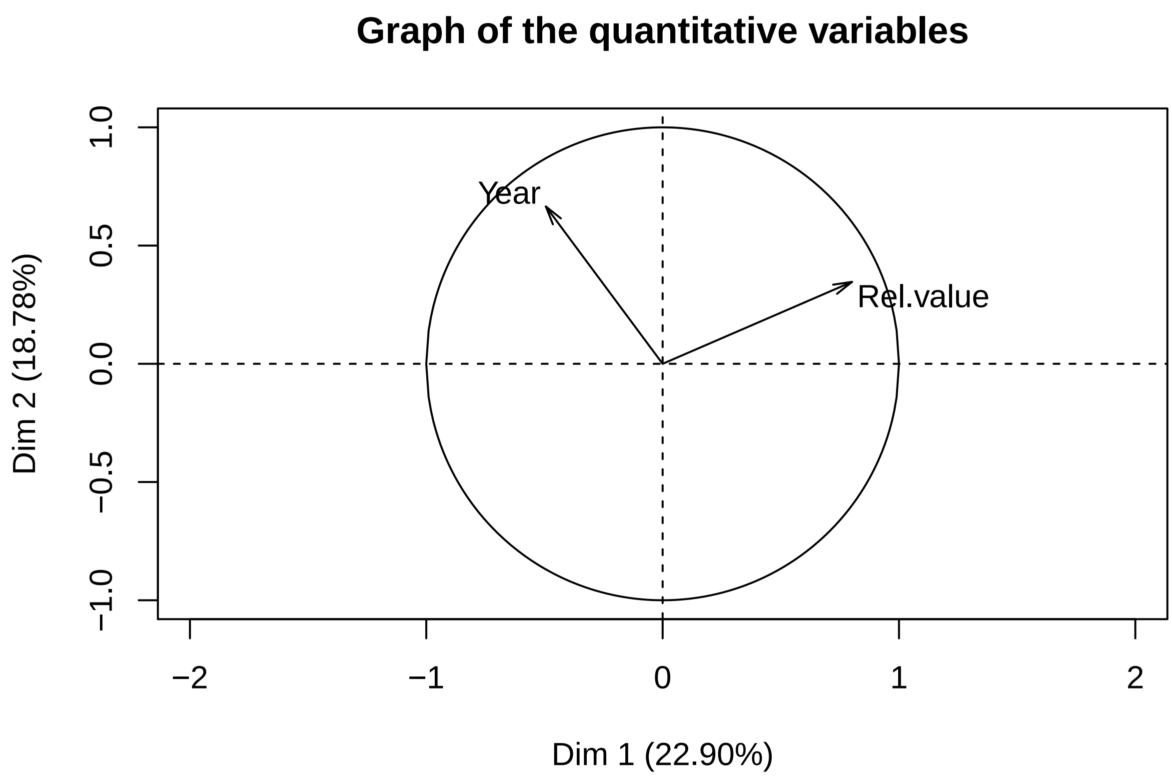

Individual factor map
Dim 1 (22.90%)
Dim 2 (18.78%)


Individual factor map
Dim 1 (22.90%)
Dim 2 (18.78%)




```
plot(famd,choix='quanti')
```




## Linguistic interpretation of the quantitative results

The results are controversial. The results, produced by linear model, contradict the results acquired using mixed-effect model. While liner model shows growth of the irregular verb usage over the time, the mixed-effect model on the contrary shows slow decline for each genre separately. But we can say for sure, that the correlation between time and normalized frequency does exist.

## Discussion on data distribution and quantitative methods in use

Thus our hypothesis was proven by linear model and rejected by the mixed-effect model. Such controversy may result from the corpus being not perfectly balanced, so for further research we can suggest enlarging and balancing our copus thus making it more representative.