Lexical Diversity Parameters Analysis for Author's Styles in Scientific and Technical Publications

Volodymyr Motyka[1], Yaroslav Stepaniak[1], Mariia Nasalska[1] and Victoria Vysotska[1,2]

[1] *Lviv Polytechnic National University, S. Bandera Street, 12, Lviv, 79013, Ukraine*
[2] *Osnabrück University, Friedrich-Janssen-Str. 1, Osnabrück, 49076, Germany*

Abstract

A self-developed dataset based on the analysis of more than 300 Ukrainian-language scientific and technical publications from the specialized Bulletin of the Lviv Polytechnic National University of the Information Systems and Networks series for 2001-2016 is selected. It contains information about the lexical and syntactic development of the author's styles of scientific publications -technical direction. Namely are the total number of words in this text, the number of words in a certain text (without repetitions), the number of words with a frequency of 1, the number of words with a frequency of 10 or more, the number of separate sentences, the number of prepositions, the number of conjunctions, Lexical diversity, Syntactic complexity, coefficient of speech coherence, exclusivity index, concentration index. The purpose of the research is to find the differences and dependence of the given data. For this, various methods of visualization and data processing, smoothing methods and correlation analysis are used.

Keywords

Стиль автора, Лексична різноманітність, Синтаксична складність, Коефіцієнт зв’язності мовлення, Індекс винятковості, Індекс концентрації, кореляційний аналіз, згладжування

1. Introduction

In this work, a rather interesting dataset was investigated, which can be described as statistics of the lexical and syntactic development of literature. The self-developed dataset is based on the results of the research of more than 300 Ukrainian-language scientific and technical publications from the specialized Bulletin of the National University "Lviv Polytechnic" of the "Information Systems and Networks" series for 2001-2016. This dataset reminded us of the popular application Grammarly, the essence of which is to increase the quality of written communication, offering guidance on correctness, clarity, appeal and tone of message.

Such variables as Lexical Diversity, Syntactic Complexity, Cohesion of Speech, Index of Exclusiveness, Index of Concentration on Dependency and Distinction were studied. We will describe the variables to improve the further understanding of the work done: Lexical diversity is the ratio of the number of words to the total number of word forms of the text, Syntactic complexity is the ratio of the number of sentences to the number of words of a certain text, Cohesiveness of speech is the ratio of the number of prepositions and conjunctions to the number of individual sentence Exclusiveness index - the variability of the vocabulary, i.e. the share of the text occupied by words that occurred 1 time, Concentration index - the share of the text occupied by words that occurred 10 times or more.

COLINS-2023: 7th International Conference on Computational Linguistics and Intelligent Systems, April 20–21, 2023, Kharkiv, Ukraine
EMAIL: volodymyr.motyka@lpnu.ua (V. Motyka), yaroslav.stepaniak@lpnu.ua (Y. Stepaniak), mariia.nasalska@lpnu.ua (M. Nasalska), victoria.a.vysotska@lpnu.ua (V. Vysotska)
ORCID 0009-0009-8086-7619 (V. Motyka), 0009-0007-3074-1132 (Y. Stepaniak), 0009-0008-1089-039X (M. Nasalska), 0000-0001-6417-3689 (V. Vysotska)

2. Related works

The scheme of the combination of methods for determining the author of Ukrainian-language textual content of a scientific and technical direction shows that it consists of lexical and syntactic levels [1]. The use of the syntactic level involves the calculation of linguistic relationships in combinations of words [2]. The work [3] proposed a model for building an author's style profile, which consists of a characteristic author's vocabulary and author's syntax [4]. To describe the syntax, it is necessary to use a formalized description of the linguistic relationships between the lexical units of a phrase in a plural-theoretical language [5]. In work [6], a formalized description of any text is put forward, but the formalized description of linguistic relationships between lexical units is not updated. A formalized description of the text is also found in references [7].

In the handbook [8], a formalized textual presentation was compiled for the automation of procedures for the analysis of scientific and educational texts in order to identify semantically significant fragments [9]. The paper [10] sets out a plural-theoretical description of linguistic relations in phrases. Such models can be used to describe images of author's vocabulary and author's syntax, but they do not take into account statistical information about vocabulary frequency and syntax [11]. The formalized description, which was used to analyze the text of the terminological dictionary in order to build a semantic network of its terms, is presented in the reference book [12]. However, the proposed model also does not include accounting for statistical information about the frequency of vocabulary and syntax [13].

Methods of determining the author of Ukrainian-language textual content of a scientific and technical direction are proposed and investigated in works [1–5]. Various algorithms [14], in particular quantitative ones [15], can be used to implement these methods. Therefore, there is a problem of analyzing such algorithms in order to find the most effective one [16].

Authorization of authorship is a technique for determining the author of a text when it is not clear who wrote it [17]. This is useful when several people claim authorship of the same publication [18] or in cases where no one claims authorship of textual content [19], such as so-called trolls in social networks during information warfare [20]. The complexity of the problem of the author's text is obviously exponentially higher, the number of probable authors is greater [21]. The availability of author's text samples is also essential in advancing this problem [22]. Attribution of the author's text includes the following three problems [23]:

- identifying the author of the text author from the group of probable or expected authors, where the author is always in the group of suspects [24];
- not identifying the author of the textual author from the group of probable or expected authors, where the author may not be in the group of suspects [25];
- assessment of the possibility of a given text, written by a given author or not [26].

Therefore, the task of automatically determining the author of scientific and technical textual content is urgent and requires new (more advanced) approaches to its solution [27-36].

3. Methods and materials

We will use the methods of visual presentation of data, smoothing, correlation method to perform the tasks. Methods of visual presentation of data - methods of presenting data in the form of graphs, charts and/or other subtypes of them (histograms, pie charts, etc.), time series, etc. Depending on the specific task, a specific method of data presentation will be used. We will implement these methods using Microsoft Power BI and/or R tools.

Smoothing methods are used to reduce the influence of the random component (random fluctuations) in time series. They make it possible to obtain more "pure" values, which consist only of deterministic components. Some of the methods are aimed at highlighting some components, for example, the trend [37-39]. We will implement these methods using Microsoft Excel, R and/or Microsoft Power BI.

Correlation method (Correlation - analysis) - a method of studying the interdependence of characteristics in the general population, which are random variables with a normal distribution [40-44] for different NLP-talks based on text analysis [45-54].

4. Experiments

Let's open the generated dataset using R Studio:

Author	Year	N	W	W1	W10	P	Z	S	Kl	Ks	Kz	Iwt	Ikt
Author 1	2001	644	366	275	7	52	43	32	0.5683230	0.8579235	0.4807692	0.7513661	0.019125683
Author 2	2001	627	397	304	6	34	42	41	0.6331738	0.9143577	0.8137255	0.7657431	0.015113350
Author 3	2001	659	399	309	8	29	44	53	0.6054628	0.9273183	1.1149425	0.7744361	0.020050125
Author 4	2001	708	419	309	8	36	64	28	0.5918079	0.9140811	0.8518519	0.7374702	0.019093079
Author 5	2001	665	423	318	4	47	63	19	0.6360902	0.8888889	0.5815603	0.7517731	0.009456265
Author 6	2001	689	384	268	8	64	63	35	0.5573295	0.8333333	0.5104167	0.6979167	0.020833333
Author 7	2001	681	385	269	6	53	70	25	0.5653451	0.8623377	0.5974843	0.6987013	0.015584416
Author 8	2001	704	382	272	7	41	67	48	0.5426136	0.8926702	0.9349593	0.7120419	0.018324607
Author 9	2001	769	495	389	6	59	72	38	0.6436931	0.8808081	0.6214689	0.7858586	0.012121212
Author 10	2001	729	380	261	7	62	75	32	0.5212620	0.8368421	0.5752688	0.6868421	0.018421053
Author 11	2001	692	435	323	4	48	73	35	0.6286127	0.8896552	0.7500000	0.7425287	0.009195402
Author 12	2001	657	420	319	6	32	69	44	0.6392694	0.9238095	1.1770833	0.7595238	0.014285714
Author 12	2001	674	378	279	6	39	75	42	0.5608309	0.8968254	1.0000000	0.7380952	0.015873016
Author 13	2001	765	378	252	7	80	70	33	0.4941176	0.7883598	0.4291667	0.6666667	0.018518519
Author 7	2002	678	420	322	6	51	43	55	0.6194690	0.8785714	0.6405229	0.7666667	0.014285714
Author 11	2002	652	390	286	2	35	58	43	0.5981595	0.9102564	0.9619048	0.7333333	0.005128205
Author 14	2002	661	374	267	6	34	47	39	0.5658094	0.9090909	0.8431373	0.7139037	0.016042781
Author 1	2002	683	399	295	5	42	51	46	0.5841874	0.8947368	0.7698413	0.7393484	0.012531328
Author 15	2002	745	439	319	6	45	59	61	0.5892617	0.8974943	0.8888889	0.7266515	0.013667426
Author 3	2002	670	420	331	4	30	60	63	0.6268657	0.9285714	1.3666667	0.7880952	0.009523810
Author 16	2002	769	426	286	6	52	49	47	0.5539662	0.8779343	0.6153846	0.6713615	0.014084507
Author 17	2002	647	422	308	3	62	50	32	0.6522411	0.8530806	0.4408602	0.7298578	0.007109005
Author 5	2002	650	382	269	4	55	66	19	0.5876923	0.8560209	0.5151515	0.7041885	0.010471204
Author 18	2002	659	353	229	7	72	63	41	0.5356601	0.7960340	0.4814815	0.6487252	0.019830028

Figure 1: Displaying the dataset in RStudio

As can be seen from the dataset (Fig. 1), there are 14 columns with data.

- Author – authors of the article;
- Year – the year of writing the text;
- N – the total number of words of this text;
- W – the number of words in a certain text (without repetitions);
- W1 – number of words with a frequency of 1;
- W10 – number of words with a frequency of 10 or more;
- P – the number of individual sentences;
- Z – the number of prepositions;
- S – the number of connectors;
- Kl – Lexical diversity;
- Ks – Syntactic complexity;
- Kz - Coefficient of speech connectivity;
- Iwt - Exclusivity index;
- Ikt - Concentration index.

Let's define the concepts of lexical diversity, syntactic diversity, speech coherence coefficient, exclusivity index and concentration index:

- Lexical diversity is the ratio of the number of words to the total number of word forms of the text. The value of the coefficient lies within [0;1]. Kl=W/N, where Kl is coefficient of lexical diversity; W is the number of words in a certain text; N is the total number of words in this text.
- Syntactic complexity is the ratio of the number of sentences to the number of words of a certain text: Ks=1-P/W, where Ks is coefficient of syntactic complexity; P is the number of sentences, W – the number of words in the entire text.

- Speech coherence coefficient - the ratio of the number of prepositions and conjunctions to the number of separate sentences: Kz=(Z+S)/(3P), where Z is the number of prepositions, S is the number of conjunctions, P is the number of separate sentences.
- Exclusiveness index - the variability of the vocabulary, i.e. the share of the text occupied by words that occurred 1 time, i.e. Iwt =W1/W, where Iwt is the exclusiveness index of the text, W1 is the number of words with a frequency of 1, W is the number of words in the entire text.
- Concentration index - the share of the text occupied by words that occur 10 times or more: Ikt =W10/W, where Ikt is the text concentration index, W10 is the number of words with a frequency of 10 or more, W is the number of words in the entire text.

Let's group the data by the Year field:

```
gr<-dt %>%
  group_by(Year)%>%
  select(Year,W)
new_gr<- gr %>% summarise(avg = mean(W10))
```

Let's calculate the quantitative characteristics by choosing the data column w1, which characterizes the number of words with a frequency of 1, by means of R:

- Sample size – the number of units in the sample: `nrow(new_gr)`
- Sample mean. We find using the built-in method mean(): `median(new_gr$avg, na.rm = FALSE)`
- The median of the sample is the number that "divides" "in half" the ordered set of all the values of the sample, that is, the average value of the changing characteristic, which is contained in the middle of the series, placed in the order of increasing or decreasing of the characteristic. For this, we will use the median() method: `median(new_gr$avg, na.rm = FALSE)`
- Mode - the value that occurs most often in the sample. Since there is no built-in method for finding it in R, we will define our modes function:

```
## modes function
modes <-function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]}##
modes(new_gr$avg)
```

- Sample size – the difference between the maximum and minimum value of the sample. To find the maximum and minimum, use the built-in methods max() and min():

```
max(new_gr$avg)-min(new_gr$avg)
```

- Standard deviation - the amount of spread relative to the arithmetic mean. To find, we will use the built-in method sd(): `sd(new_gr$avg)`
- Coefficient of variation – an indicator that determines the percentage ratio of the average deviation to the average value: `sd(new_gr$avg)*100/mean(new_gr$avg, na.rm = FALSE)`
- Asymmetry reflects the skewness of the distribution relative to the mode. Let's use the built-in skewness() method: `skewness(new_gr$avg)`
- The kurtosis coefficient characterizes the "steepness", that is, the steepness of the rise of the distribution curve compared to the normal curve. Let's use the kurtosis() method: `kurtosis(new_gr$avg)`
- Standard error is the deviation of the sample from the actual mean. To find it, we will use the formula for calculating the standard error and the sd() method for calculating the standard deviation: `sd(new_gr$avg)/sqrt(nrow(new_gr))`

To find the number of intervals, we will use Sturges' formula, and to find the width of the interval - Scott's formula. Cumulative – a continuous curve is displayed graphically, which gives a more accurate result compared to a histogram. For construction, we will use the ecdf() function. Finding the number of intervals and the interval width for the avg attribute:

```
k<-1+log2(nrow(new_gr)) #Number of intervals
h<-3.5*sd(new_gr$avg)*(nrow(new_gr))^(-1/3) #Interval width
```

Construction of a histogram: `hist(new_gr$avg, breaks = k, xlab = "", main = "Histogram of w")`

Construction of cumulata:

```
plot(ecdf(new_gr$avg), main="Cumulate", xlab="", ylab = "Frequency", verticals = FALSE)
```

Smoothing methods are used to reduce the influence of the random component (random fluctuations) in time series. They make it possible to obtain more "pure" values, which consist only of deterministic components. Some of the methods are aimed at highlighting some components, for example, a trend. Smoothing methods can be conventionally divided into two classes based on different approaches: analytical and algorithmic.

The simplest method of forecasting is considered to be an approach that determines the forecast estimate from the actually achieved level using the average level, average growth, average growth rate. Extrapolation based on the average level of the series. The resulting confidence interval takes into account the uncertainty hidden in the estimate of the average value. However, the assumption remains that the predicted indicator is equal to the sample mean, that is, this approach does not take into account the fact that individual values of the indicator have fluctuated around the average in the past, and this will also happen in the future.

Analytical smoothing methods include regression analysis together with the method of least squares and its modifications. To identify the main trend by analytical method means to give the studied process the same development throughout the entire observation period. Therefore, for 4 of these methods, it is important to choose the optimal function of the deterministic trend (growth curve), which smoothes a number of observations.

Forecasting methods based on regression methods are used for short- and medium-term forecasting. They do not allow for adaptation: with the receipt of new data, the forecast construction procedure must be repeated from the beginning. The optimal length of the lead-up period is determined separately for each economic process, taking into account its statistical instability.

The most widely used are the methods of smoothing time series using moving averages. For moving average smoothing, we will use Kendel's formulas to calculate the lost levels at the beginning and end of the smoothed series. Let's prepare the data for using smoothing methods:

```
ma <- new_gr %>% select(Year,avg) %>%
  mutate(ma1 = rollmean(avg, k = 3, fill = NA), ma2 = rollmean(avg, k = 5, fill = NA),
         ma3 = rollmean(avg, k = 7, fill = NA))
```

The method of smoothing according to Kendel's formulas:

```
k_ma1<-matrix(c(5,2,-1,6,3,6),byrow = TRUE,nrow=2)
ma$ma1[1]<-0
ma$ma1[16]<-0
for(i in 1:3){
  ma$ma1[1]<-ma$ma1[1]+k_ma1[1,i]*ma$avg[i]/k_ma1[nrow(k_ma1),1]
  ma$ma1[16]<-ma$ma1[16]+k_ma1[1,i]*ma$avg[17-i]/k_ma1[nrow(k_ma1),3]}
k_ma2<-matrix(c(3,2,1,0,-1,4,3,2,1,0,5,10,5,10,5),byrow = TRUE,nrow=3)
ma$ma2[1]<-0
ma$ma2[2]<-0
ma$ma2[15]<-0
ma$ma2[16]<-0
for(j in 1:2){
  for(i in 1:5)  {
    ma$ma2[j]<-ma$ma2[j]+k_ma2[j,i]*ma$avg[i]/k_ma2[nrow(k_ma2),j]
    ma$ma2[17-j]<-ma$ma2[17-j]+k_ma2[j,i]*ma$avg[17-i]/k_ma2[nrow(k_ma2),j]  }}
k_ma3<-matrix(c(seq(13,-5,by=-3),seq(5,-1,by=-1),seq(7,1,by=-1),28,14,28,7,28,14,28),
              byrow = TRUE,nrow=4)
ma$ma3[1]<-0
ma$ma3[2]<-0
ma$ma3[3]<-0
ma$ma3[14]<-0
ma$ma3[15]<-0
ma$ma3[16]<-0
for(j in 1:3){
  for(i in 1:7)  {
    ma$ma3[j]<-ma$ma3[j]+k_ma3[j,i]*ma$avg[i]/k_ma3[nrow(k_ma3),j]
    ma$ma3[17-j]<-ma$ma3[17-j]+k_ma3[j,i]*ma$avg[17-i]/k_ma3[nrow(k_ma3),j]  }}
```

Data visualization:

```
ma %>% gather(metric, avg, avg:ma3) %>%
  ggplot(aes(Year, avg, color = metric)) + geom_line(size=1)+ labs(title = "Ikt")+
  theme(legend.title = element_blank(),plot.title = element_text(hjust = 0.5))
```

Finding turning points:

```
tp1 <- turnpoints(ma$ma1)
```

summary(tp1)

```
tp2 <- turnpoints(ma$ma2)
```

summary(tp2)

```
tp3 <- turnpoints(ma$ma3)
```

summary(tp3)

Visualization of turning points:

```
plot(ma$ma1, type = "l")
lines(tp1)
```

We are looking for the correlation coefficients of the smoothed values with the original ones, taking into account the fact that with each smoothing we subtract rows:

```
cor(ma$avg,ma$ma1)
cor(ma$avg,ma$ma2)
cor(ma$avg,ma$ma3)
```

Similarly, we do research for w1 and w10.

Exponential smoothing:

```
alpha<-0.1
exp_smooth<-1:16
exp_smooth[1]<-ma$avg[1]
for(i in 2:16){  exp_smooth[i]<-ma$avg[i]*alpha +(1-alpha)*exp_smooth[i-1]}
```

Visualization:

```
ggplot(ma,mapping= aes(x=Year)) + geom_line(mapping= aes(y=avg, col="Real"),lwd=1.5) +
  geom_line(mapping= aes(y=exp_smooth, col="es"),lwd=1.5)+
  scale_color_manual(values= c("Real"="blue","es"="red"))+ labs(x="",y="",title ="alpha = 0.30")+
  theme(legend.title = element_blank(),plot.title = element_text(hjust = 0.5))
```

Median filtering:

```
med_fil<-1:16
med_fil[1]<-(5*ma$avg[1]+2*ma$avg[2]-ma$avg[3])/6
med_fil[16]<-(-ma$avg[14]+2*ma$avg[15]+5*ma$avg[16])/6
for(i in 2:15){
  med_fil[i]<-max(min(ma$avg[i-1],ma$avg[i]),min(ma$avg[i],ma$avg[i+1]),min(ma$avg[i-1],ma$avg[i+1]))}
```

Visualization:

```
ggplot(ma,mapping= aes(x=Year)) + geom_line(mapping= aes(y=avg, col="Real"),lwd=1.5) +
  geom_line(mapping= aes(y=med_fil, col="Median"),lwd=1.5)+
  scale_color_manual(values= c("Real"="blue","Median"="red"))+
  labs(x="",y="Views",title ="Median filter")+
  theme(legend.title = element_blank(),plot.title = element_text(hjust = 0.5))
```

Turning points:

```
tp_mf<-turnpoints(med_fil)
summary(tp_mf)
```

Visualization of turning points:

```
plot(ma$avg, type = "l")
lines(tp_mf)
```

Correlation coefficient: `cor(ma$avg,med_fil)`

In general, correlation can be described as any statistical relationship of data. Correlation allows us to see the trends of changes in the average values of the functions depending on the parameter changes. Correlation can be positive or negative. Negative correlation is a correlation in which an increase in one variable is associated with a decrease in another, and the correlation coefficient is negative. Positive correlation is a correlation in which an increase in one variable is associated with an increase in another, and the correlation coefficient is positive.

Construction of the correlation field (plot)

```
plot(dt$Kl, dt$W, main="Correlation field", xlab="lexical diversity",
     ylab="Word count without duplicates")
plot(dt$Ks, dt$P, main="Correlation field", xlab="Syntax complexity", ylab="Sentance count")
plot(dt$Kz, dt$P, main="Correlation field", xlab="Coefficient of coherent speech",
     ylab="Sentance counts")
plot(dt$Iwt, dt$W1, main="Correlation field", xlab="Coefficient of coherent speech",
     ylab= "Count of words that have only one duplicate")
plot(dt$Ikt, dt$W10, main="Correlation field", xlab="Coefficient of coherent speech",
     ylab="Count of words that have 10 or more duplicates")
```

Finding multiple correlation coefficients:

```
numericData <- cbind(dt$N,dt$W,dt$P,dt$Ks)
chart.Correlation(numericData, histogram=FALSE, pch=19)
numericData <- cbind(dt$P,dt$Z,dt$S,dt$Kz)
chart.Correlation(numericData, histogram=FALSE, pch=19)
numericData <- cbind(dt$N,dt$W, dt$W1,dt$Iwt)
chart.Correlation(numericData, histogram=FALSE, pch=19)
numericData <- cbind(dt$N,dt$W,dt$W10,dt$Ikt)
chart.Correlation(numericData, histogram=FALSE, pch=19)
```

5. Results

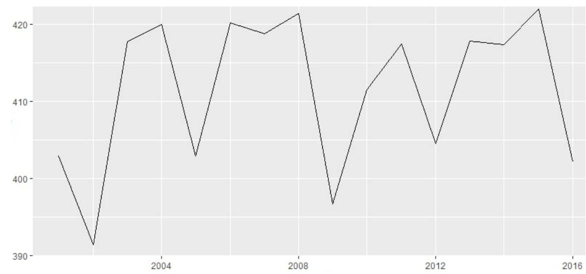
Let's present the dataset in the form of a table and group the data by years:

Author	Year	N	W	W1	W10	P	Z	S	Kl	Ks	Kz	Iwt	Ikt
Author 1	2001	644	366	275	7	52	43	32	0.5683230	0.8579235	0.4807692	0.7513661	0.019125683
Author 2	2001	627	397	304	6	34	42	41	0.6331738	0.9143577	0.8137255	0.7657431	0.015113350
Author 3	2001	659	399	309	8	29	44	53	0.6054628	0.9273183	1.1149425	0.7744361	0.020050125
Author 4	2001	708	419	309	8	36	64	28	0.5918079	0.9140811	0.8518519	0.7374702	0.019093079
Author 5	2001	665	423	318	4	47	63	19	0.6360902	0.8888889	0.5815603	0.7517731	0.009456265
Author 6	2001	689	384	268	8	64	63	35	0.5573295	0.8333333	0.5104167	0.6979167	0.020833333
Author 7	2001	681	385	269	6	53	70	25	0.5653451	0.8623377	0.5974843	0.6987013	0.015584416
Author 8	2001	704	382	272	7	41	67	48	0.5426136	0.8926702	0.9349593	0.7120419	0.018324607
Author 9	2001	769	495	389	6	59	72	38	0.6436931	0.8808081	0.6214689	0.7858586	0.012121212
Author 10	2001	729	380	261	7	62	75	32	0.5212620	0.8368421	0.5752688	0.6868421	0.018421053
Author 11	2001	692	435	323	4	48	73	35	0.6286127	0.8896552	0.7500000	0.7425287	0.009195402
Author 12	2001	657	420	319	6	32	69	44	0.6392694	0.9238095	1.1770833	0.7595238	0.014285714
Author 12	2001	674	378	279	6	39	75	42	0.5608309	0.8968254	1.0000000	0.7380952	0.015873016
Author 13	2001	765	378	252	7	80	70	33	0.4941176	0.7883598	0.4291667	0.6666667	0.018518519
Author 7	2002	678	420	322	6	51	43	55	0.6194690	0.8785714	0.6405229	0.7666667	0.014285714
Author 11	2002	652	390	286	2	35	58	43	0.5981595	0.9102564	0.9619048	0.7333333	0.005128205
Author 14	2002	661	374	267	6	34	47	39	0.5658094	0.9090909	0.8431373	0.7139037	0.016042781
Author 1	2002	683	399	295	5	42	51	46	0.5841874	0.8947368	0.7698413	0.7393484	0.012531328
Author 15	2002	745	439	319	6	45	59	61	0.5892617	0.8974943	0.8888889	0.7266515	0.013667426
Author 3	2002	670	420	331	4	30	60	63	0.6268657	0.9285714	1.3666667	0.7880952	0.009523810
Author 16	2002	769	426	286	6	52	49	47	0.5539662	0.8779343	0.6153846	0.6713615	0.014084507
Author 17	2002	647	422	308	3	62	50	32	0.6522411	0.8530806	0.4408602	0.7298578	0.007109005
Author 5	2002	650	382	269	4	55	66	19	0.5876923	0.8560209	0.5151515	0.7041885	0.010471204
Author 18	2002	659	353	229	7	72	63	41	0.5356601	0.7960340	0.4814815	0.6487252	0.019830028

Year	avg
2001	402.9286
2002	391.4800
2003	417.7727
2004	420.0000
2005	403.0000
2006	420.1667
2007	418.8333
2008	421.4667
2009	396.7143
2010	411.5000
2011	417.5000
2012	404.5556
2013	417.8571
2014	417.3939
2015	422.0000
2016	402.2500

Figure 2: Selected dataset in table form and table view of data grouped by years

The number of words in a certain text without repetitions and the number of words with a frequency of more than 10:



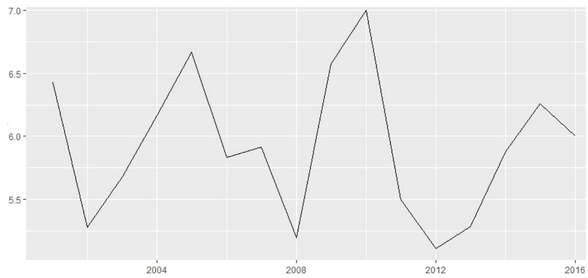


Figure 3: Graph of the number of words in the text without repetitions by year and the number of words with frequency > 10

Let's find the statistical parameters for the attribute (Table 1).

Table 1

Descriptive statistics of attribute w

Name	Value
Sample size	16
Selective average	411.59
Median	417.45
Mode	402.93
Sample size	30.52
Standard deviation	9.87
Coefficient of variation	2.40
Asymmetry coefficient	-0.67
Kurtosis	2.05
Standard error	2.47

After executing the code, we have histograms and corresponding cumulates:

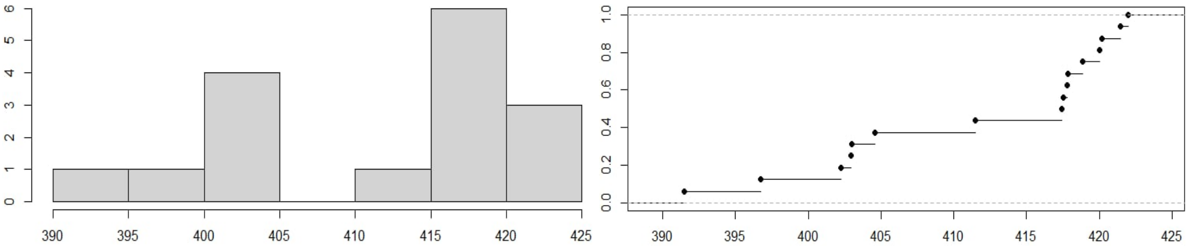


Figure 4: Histogram of data and cumulate about the number of words without repetitions

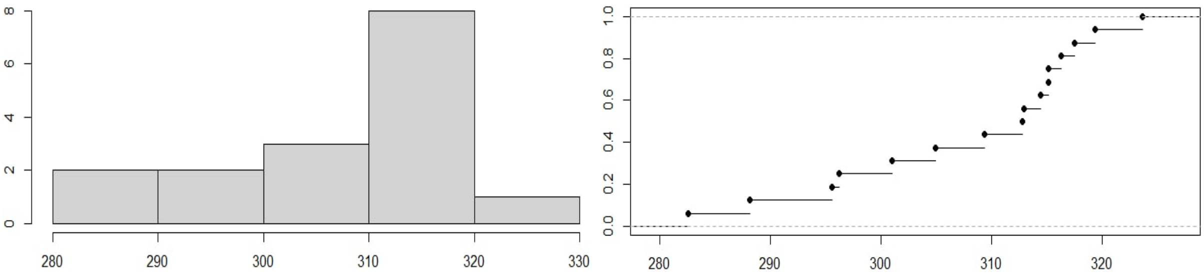


Figure 5: Histogram of data and cumulate about the number of words with a frequency of 1

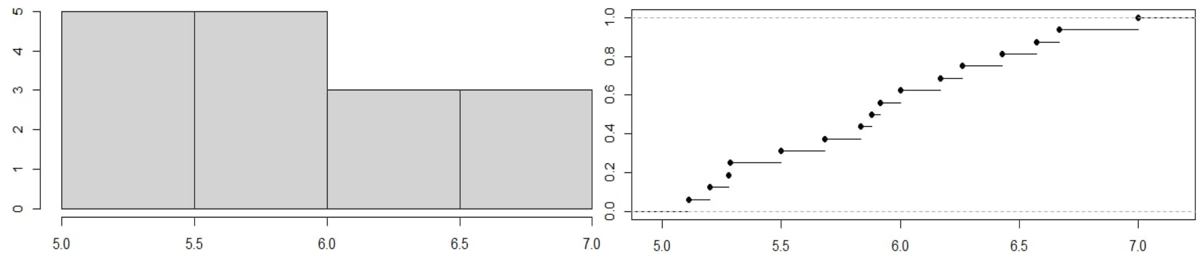


Figure 6: Data histogram and cumulate about the number of words with a frequency of 10 or more

For a general comparison, we will display the graphs of attributes Kl, Ks, Kz, Iwt, Ikt (Fig. 7-8):

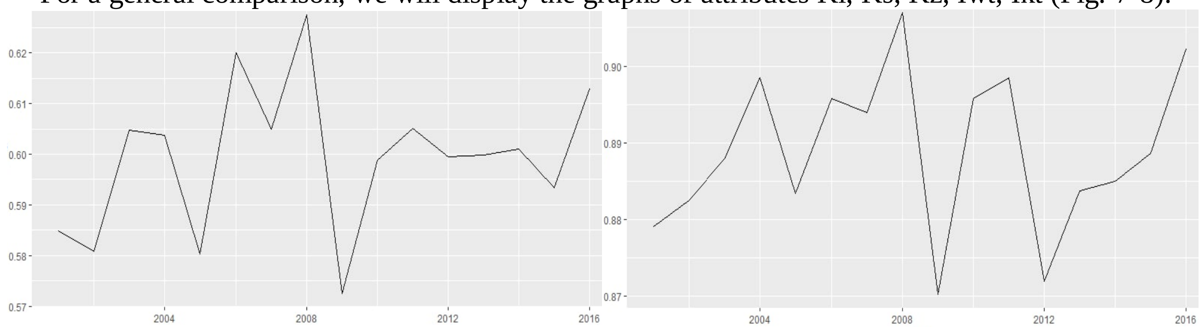


Figure 7: Data graph of attributes Kl and Ks

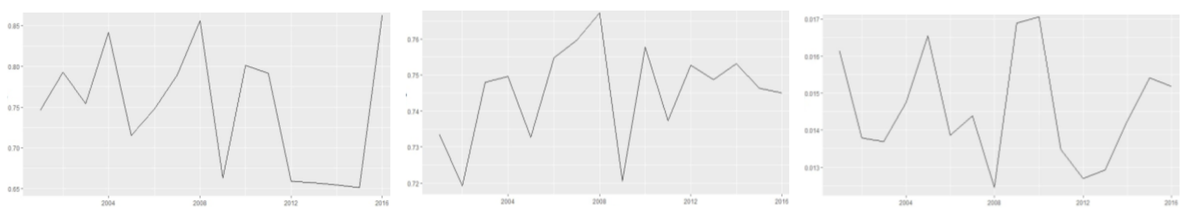


Figure 8: Data plot of Kz, Iwt and Ikt attributes respectively

Let's analyse the change in time series trends using smoothing methods.

Using Kendel's formulas, we obtained the initial and final values that were lost in the calculation of the averages, depending on the average from which we calculate the data.

Year	avg	ma1	ma2	ma3
2001	0.01614256	0.01576454	0.01462981	0.01498111
2002	0.01378499	0.01454102	0.01480630	0.01489967
2003	0.01369551	0.01407577	0.01498279	0.01481823
2004	0.01474683	0.01499547	0.01452668	0.01473679
2005	0.01654408	0.01505097	0.01464600	0.01421032
2006	0.01386202	0.01492922	0.01439835	0.01465302
2007	0.01438157	0.01356695	0.01482576	0.01513361
2008	0.01245726	0.01457424	0.01492887	0.01495283
2009	0.01688389	0.01546692	0.01485275	0.01440338
2010	0.01705961	0.01580830	0.01451601	0.01426937
2011	0.01348141	0.01441297	0.01460935	0.01425064
2012	0.01269789	0.01303442	0.01408267	0.01467275
2013	0.01292396	0.01329078	0.01375315	0.01442930
2014	0.01425047	0.01419548	0.01409281	0.01442123
2015	0.01541201	0.01494740	0.01483799	0.01441317
2016	0.01517973	0.01541203	0.01558316	0.01440510

```
> summary(tp1)
Turning points for: ma$ma1

nbr observations  : 16
nbr ex-aequos     : 0
nbr turning points: 6 (first point is a peak)
E(p) = 9.333333 Var(p) = 2.522222 (theoretical)

  point type       proba     info
1     3 peak 0.250000000 2.000000
2     4  pit 0.066666667 3.906891
3     7 peak 0.007936508 6.977280
4    10  pit 0.027777778 5.169925
5    12 peak 0.100000000 3.321928
6    14  pit 0.100000000 3.321928
> tp2 <- turnpoints(ma$ma2)
> summary(tp2)
Turning points for: ma$ma2

nbr observations  : 16
nbr ex-aequos     : 0
nbr turning points: 6 (first point is a peak)
E(p) = 9.333333 Var(p) = 2.522222 (theoretical)

  point type       proba      info
1     6 peak 0.005952381 7.3923174
2     7  pit 0.666666667 0.5849625
3     8 peak 0.066666667 3.9068906
4    11  pit 0.100000000 3.3219281
5    12 peak 0.666666667 0.5849625
6    13  pit 0.066666667 3.9068906
> tp3 <- turnpoints(ma$ma3)
> summary(tp3)
Turning points for: ma$ma3

nbr observations  : 16
nbr ex-aequos     : 0
nbr turning points: 4 (first point is a peak)
E(p) = 9.333333 Var(p) = 2.522222 (theoretical)

  point type       proba     info
1     5 peak 0.007936508 6.977280
2     7  pit 0.100000000 3.321928
3     9 peak 0.027777778 5.169925
4    12  pit 0.001736111 9.169925
```

Figure 9: Smoothed data according to Kendel's formulas and turning points when smoothing k=3, 5, 7

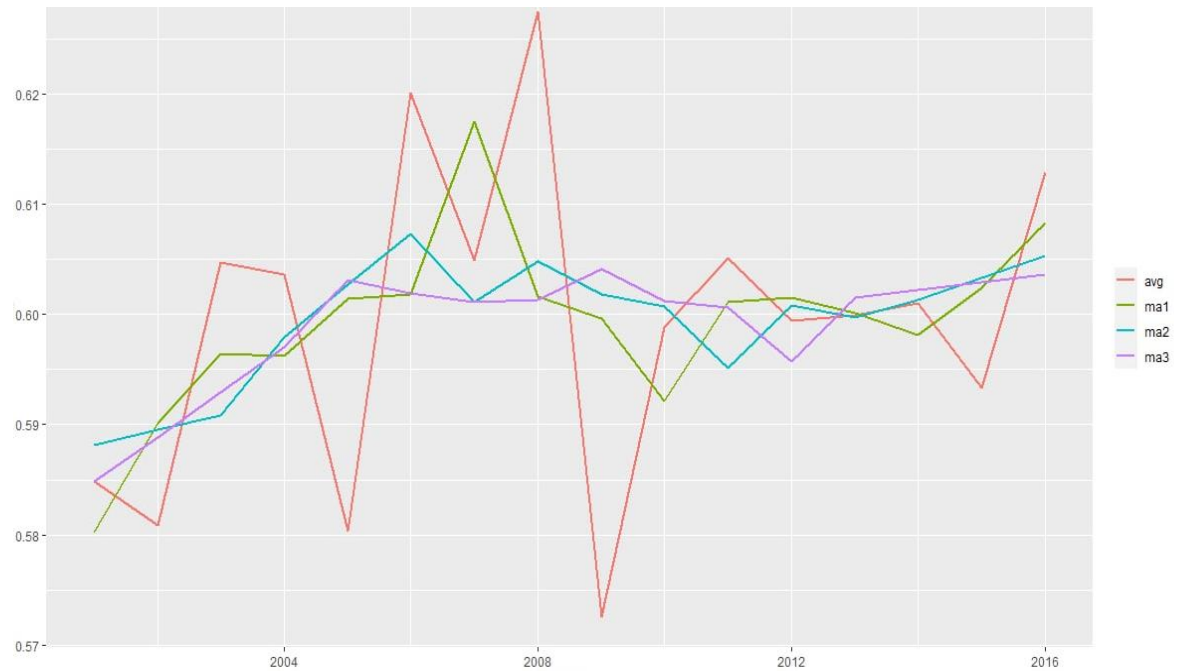


Figure 10: Graphs of smoothed data of attribute Kl at k=3, 5, 7

It can be noted that graphs with k>5 are not very suitable for us to identify trends, since we do not have a large date interval, only 16 years. For more accurate detection of trends, it is desirable to take k=3.5. At k=7 was plotted to show that the data is smoothed too much.

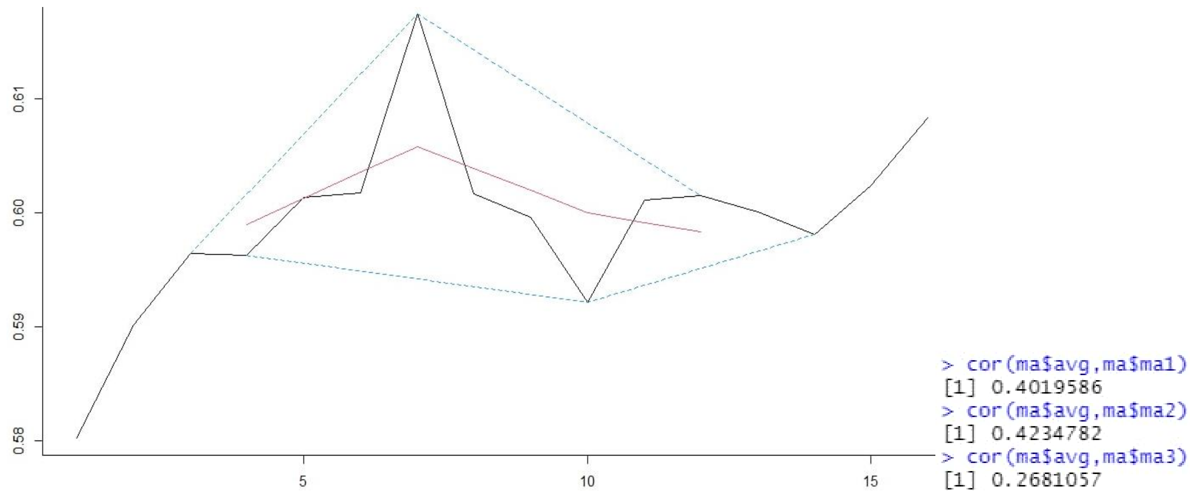


Figure 11: Visualization of turning points and correlation coefficients of smoothed data and real

The correlation coefficients are not large and positive. This is probably because we took annual averages everywhere.

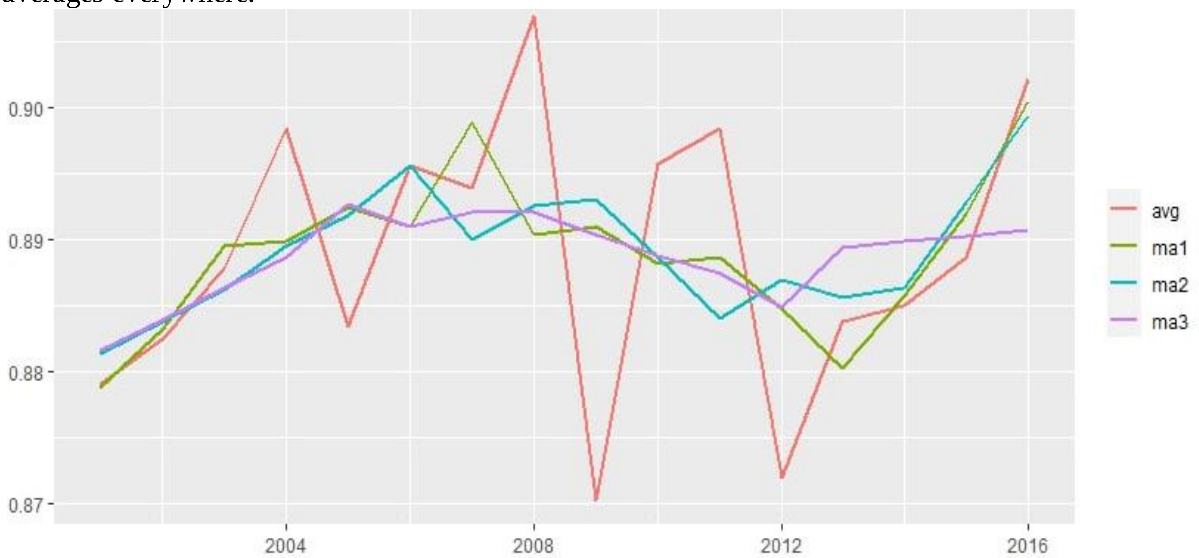


Figure 12: Graphs of the arranged data of the Ks field at k= 3, 5, 7

It can be noted that ma4, ma5, ma6, ma7 are not very suitable for detecting trends, since we do not have a large date interval, only 16 years. For more accurate detection of trends, it is advisable to take ma1, ma2 or ma3.

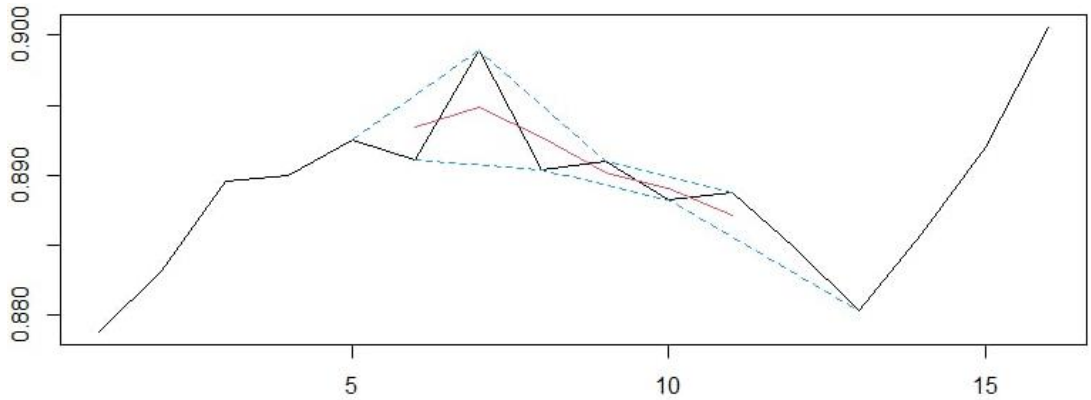


Figure 13: Visualization of turning points

```
> tp1 <- turnpoints(ma$ma1)
> summary(tp1)
Turning points for: ma$ma1

nbr observations  : 16
nbr ex-aequos     : 0
nbr turning points: 8 (first point is a peak)
E(p) = 9.333333 Var(p) = 2.522222 (theoretical)

  point type      proba      info
1     5 peak 0.02777778 5.1699250
2     6  pit 0.66666667 0.5849625
3     7 peak 0.66666667 0.5849625
4     8  pit 0.66666667 0.5849625
5     9 peak 0.66666667 0.5849625
6    10  pit 0.66666667 0.5849625
7    11 peak 0.25000000 2.0000000
8    13  pit 0.02777778 5.1699250
> tp2 <- turnpoints(ma$ma2)
> summary(tp2)
Turning points for: ma$ma2

nbr observations  : 16
nbr ex-aequos     : 0
nbr turning points: 6 (first point is a peak)
E(p) = 9.333333 Var(p) = 2.522222 (theoretical)

  point type       proba      info
1     6 peak 0.005952381 7.3923174
2     7  pit 0.250000000 2.0000000
3     9 peak 0.100000000 3.3219281
4    11  pit 0.250000000 2.0000000
5    12 peak 0.666666667 0.5849625
6    13  pit 0.066666667 3.9068906
> tp3 <- turnpoints(ma$ma3)
> summary(tp3)
Turning points for: ma$ma3

nbr observations  : 16
nbr ex-aequos     : 0
nbr turning points: 4 (first point is a peak)
E(p) = 9.333333 Var(p) = 2.522222 (theoretical)

  point type        proba      info
1     5 peak 0.0277777778  5.169925
2     6  pit 0.2500000000  2.000000
3     8 peak 0.0059523810  7.392317
4    12  pit 0.0003858025 11.339850
```

```
> tp1 <- turnpoints(ma$ma1)
> summary(tp1)
Turning points for: ma$ma1

nbr observations  : 16
nbr ex-aequos     : 0
nbr turning points: 6 (first point is a peak)
E(p) = 9.333333 Var(p) = 2.522222 (theoretical)

  point type       proba      info
1     3 peak 0.027777778 5.1699250
2     6  pit 0.100000000 3.3219281
3     7 peak 0.666666667 0.5849625
4     8  pit 0.666666667 0.5849625
5     9 peak 0.002380952 8.7142455
6    14  pit 0.001736111 9.1699250
> tp2 <- turnpoints(ma$ma2)
> summary(tp2)
Turning points for: ma$ma2

nbr observations  : 16
nbr ex-aequos     : 0
nbr turning points: 7 (first point is a pit)
E(p) = 9.333333 Var(p) = 2.522222 (theoretical)

  point type       proba      info
1     3  pit 0.250000000 2.0000000
2     4 peak 0.666666667 0.5849625
3     5  pit 0.666666667 0.5849625
4     6 peak 0.666666667 0.5849625
5     7  pit 0.250000000 2.0000000
6     9 peak 0.005952381 7.3923174
7    13  pit 0.001736111 9.1699250
> tp3 <- turnpoints(ma$ma3)
> summary(tp3)
Turning points for: ma$ma3

nbr observations  : 16
nbr ex-aequos     : 0
nbr turning points: 5 (first point is a peak)
E(p) = 9.333333 Var(p) = 2.522222 (theoretical)

  point type       proba      info
1     5 peak 0.027777778 5.1699250
2     6  pit 0.666666667 0.5849625
3     7 peak 0.002380952 8.7142455
4    12  pit 0.005952381 7.3923174
5    13 peak 0.066666667 3.9068906
>
```

Figure 14: Pivot points for Ks and Kz attribute data

```
> cor(ma$avg,ma$ma1)
[1] 0.4949916
> cor(ma$avg,ma$ma2)
[1] 0.3824844
> cor(ma$avg,ma$ma3)
[1] 0.4300794
```

```
> cor(viewh$likes,ma$ma1)
[1] 0.9634537
> cor(viewh$likes,ma$ma2)
[1] 0.905811
> cor(viewh$likes,ma$ma3)
[1] 0.8585738
> cor(viewh$likes,ma$ma4)
[1] 0.8479456
> cor(viewh$likes,ma$ma5)
[1] 0.8404666
> cor(viewh$likes,ma$ma6)
[1] 0.823287
> cor(viewh$likes,ma$ma7)
[1] 0.8143748
```

Figure 15: Correlation coefficients of smoothed data and real data

Using Kendel's formulas, we obtained the initial and final values that were lost in the calculation of the averages, depending on the average from which we calculate the data.

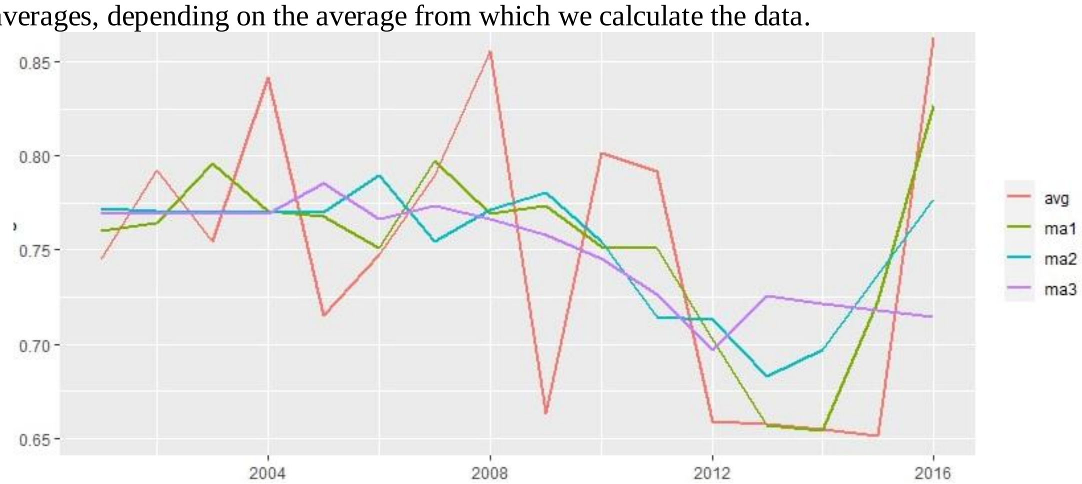


Figure 16: Graphs of the arranged data of the Kz field

It can be noted that ma4, ma5, ma6, ma7 are not very suitable for identifying trends, since we do not have a large date interval, only 40 days. For more accurate detection of trends, it is advisable to take ma1, ma2 or ma3. The number of turning points allows better analysis of trends.

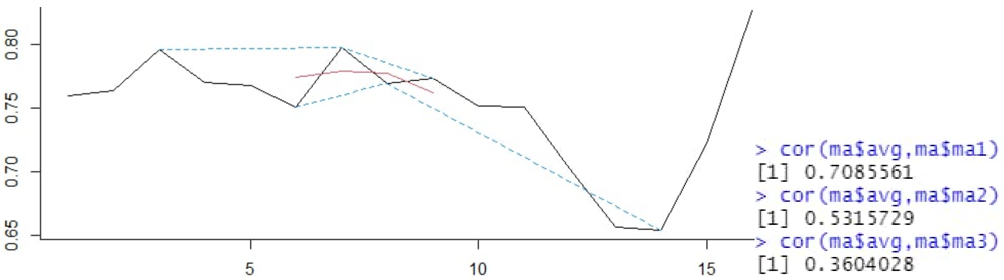


Figure 17: Visualization of turning points and correlation coefficients between smoothed data and actual data

Using Kendel's formulas, we obtained the initial and final values that were lost in the calculation of the averages, depending on the average from which we calculate the data.

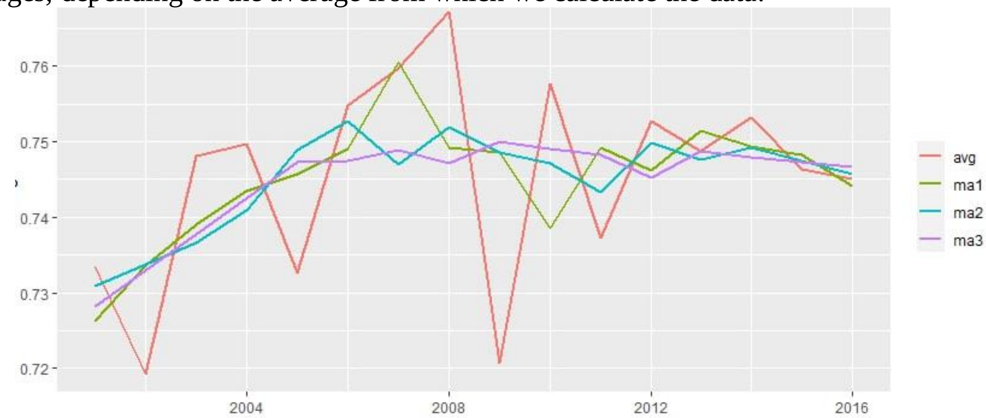


Figure 18: Plots of the lwt field's structured data

It can be noted that ma4, ma5, ma6, ma7 are not very suitable for identifying trends, since we do not have a large date interval, only 40 days. For more accurate detection of trends, it is advisable to take ma1, ma2 or ma3. The number of turning points allows better analysis of trends.

```
> tp1 <- turnpoints(ma$ma1)
> summary(tp1)
Turning points for: ma$ma1

nbr observations  : 16
nbr ex-aequos     : 0
nbr turning points: 5 (first point is a peak)
E(p) = 9.333333 Var(p) = 2.522222 (theoretical)

  point type        proba       info
1     7 peak 6.944444e-05 13.8137812
2    10  pit 1.000000e-01  3.3219281
3    11 peak 6.666667e-01  0.5849625
4    12  pit 6.666667e-01  0.5849625
5    13 peak 6.666667e-02  3.9068906
> tp2 <- turnpoints(ma$ma2)
> summary(tp2)
Turning points for: ma$ma2

nbr observations  : 16
nbr ex-aequos     : 0
nbr turning points: 7 (first point is a peak)
E(p) = 9.333333 Var(p) = 2.522222 (theoretical)

  point type       proba      info
1     6 peak 0.005952381 7.3923174
2     7  pit 0.666666667 0.5849625
3     8 peak 0.066666667 3.9068906
4    11  pit 0.100000000 3.3219281
5    12 peak 0.666666667 0.5849625
6    13  pit 0.666666667 0.5849625
7    14 peak 0.250000000 2.0000000
> tp3 <- turnpoints(ma$ma3)
> summary(tp3)
Turning points for: ma$ma3

nbr observations  : 16
nbr ex-aequos     : 0
nbr turning points: 5 (first point is a peak)
E(p) = 9.333333 Var(p) = 2.522222 (theoretical)

  point type       proba      info
1     7 peak 0.001041667 9.9068906
2     8  pit 0.666666667 0.5849625
3     9 peak 0.066666667 3.9068906
4    12  pit 0.100000000 3.3219281
5    13 peak 0.066666667 3.9068906
```

```
> tp1 <- turnpoints(ma$ma1)
> summary(tp1)
Turning points for: ma$ma1

nbr observations  : 16
nbr ex-aequos     : 0
nbr turning points: 5 (first point is a pit)
E(p) = 9.333333 Var(p) = 2.522222 (theoretical)

  point type       proba     info
1     3  pit 0.100000000 3.321928
2     5 peak 0.100000000 3.321928
3     7  pit 0.027777778 5.169925
4    10 peak 0.027777778 5.169925
5    12  pit 0.005952381 7.392317
> tp2 <- turnpoints(ma$ma2)
> summary(tp2)
Turning points for: ma$ma2

nbr observations  : 16
nbr ex-aequos     : 0
nbr turning points: 8 (first point is a peak)
E(p) = 9.333333 Var(p) = 2.522222 (theoretical)

  point type      proba      info
1     3 peak 0.25000000 2.0000000
2     4  pit 0.66666667 0.5849625
3     5 peak 0.66666667 0.5849625
4     6  pit 0.25000000 2.0000000
5     8 peak 0.10000000 3.3219281
6    10  pit 0.25000000 2.0000000
7    11 peak 0.25000000 2.0000000
8    13  pit 0.02777778 5.1699250
> tp3 <- turnpoints(ma$ma3)
> summary(tp3)
Turning points for: ma$ma3

nbr observations  : 16
nbr ex-aequos     : 0
nbr turning points: 4 (first point is a pit)
E(p) = 9.333333 Var(p) = 2.522222 (theoretical)

  point type       proba     info
1     5  pit 0.007936508 6.977280
2     7 peak 0.005952381 7.392317
3    11  pit 0.027777778 5.169925
4    12 peak 0.013888889 6.169925
```

Figure 19: Pivot points for lwt and lkt attribute data when smoothing

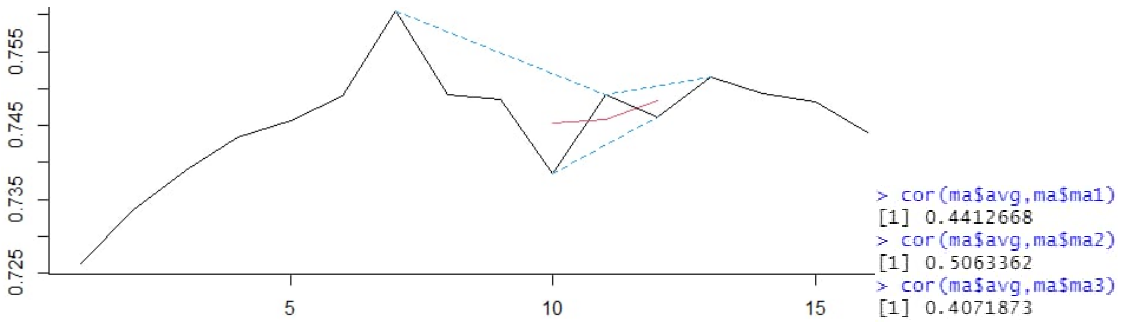


Figure 20: Visualization of turning points and correlation coefficients between smoothed data and actual data

Using Kendel's formulas, we obtained the initial and final values that were lost in the calculation of the averages, depending on the average from which we calculate the data.

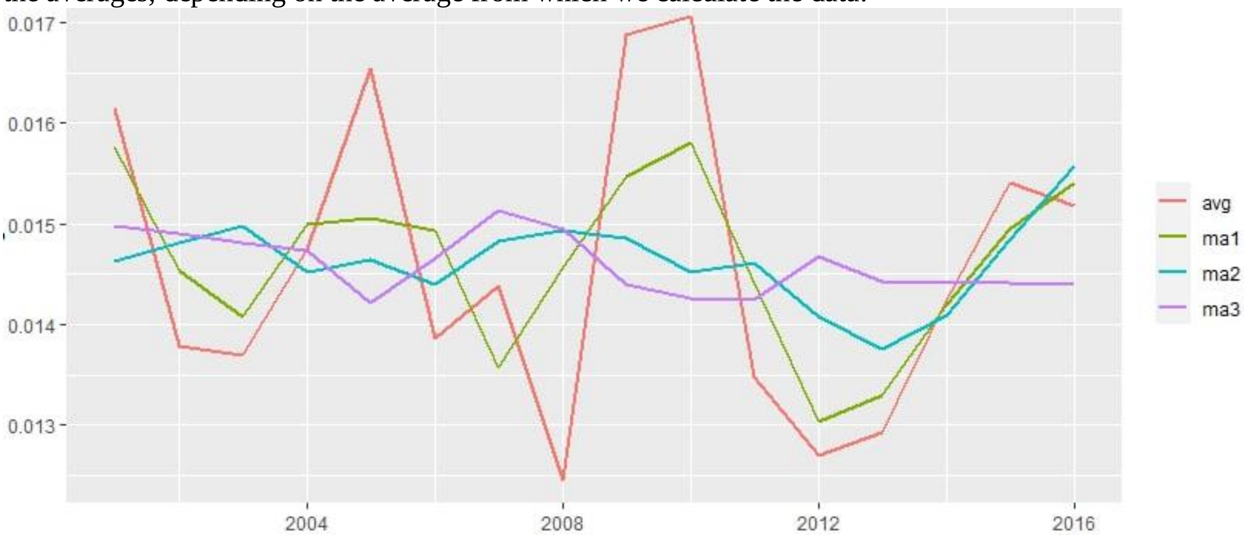


Figure 21: Graphs of organized data of field Ikt

It can be noted that ma4, ma5, ma6, ma7 are not very suitable for detecting trends, since we do not have a large date interval, only 16 years. For more accurate detection of trends, it is advisable to take ma1, ma2 or ma3. The number of turning points allows better analysis of trends.

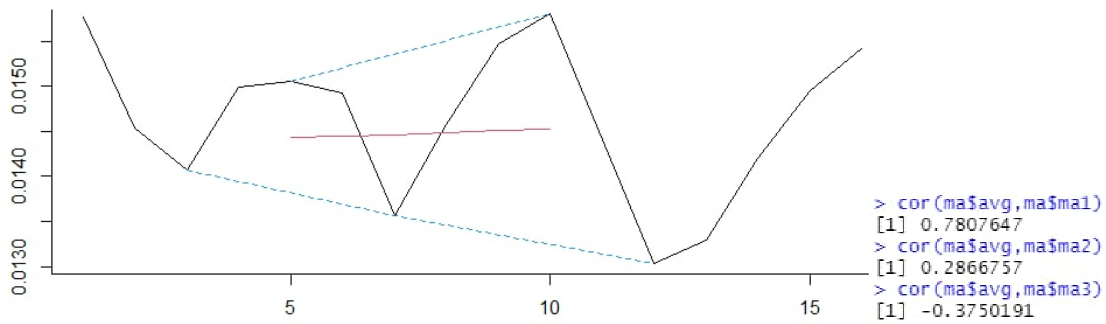


Figure 22: Visualization of turning points and correlation coefficients between smoothed data and actual data

Correlation coefficients approach 1 and decrease as the step increases, as less and less data will influence the average.

Exponential smoothing directly depends on the latest data, i.e. how the weighted average will react quickly to changes.

	dates	views	exp_smooth
1	2021-09-13	239818781	239818781
2	2021-09-14	180704846	222084601
3	2021-09-15	193399025	213478928
4	2021-09-16	194840017	207887255
5	2021-09-17	223707499	212633328
6	2021-09-18	215610185	213526385
7	2021-09-19	214269241	213749242
8	2021-09-20	209934374	212604781
9	2021-09-21	219536480	214684291
10	2021-09-22	242981546	223173468
11	2021-09-23	163013177	205125380
12	2021-09-24	162032174	192197418
13	2021-09-25	187636257	190829070
14	2021-09-26	208819913	196226323
15	2021-09-27	218266110	202838259
16	2021-09-28	196841253	201039157
17	2021-09-29	158367921	188237786
18	2021-09-30	156942080	178849074
19	2021-10-01	156987296	172290541
20	2021-10-02	170109959	171636366
21	2021-10-03	180537851	174306812
22	2021-10-04	179521927	175871346
23	2021-10-05	120340446	159212076
24	2021-10-06	104614544	142832817
25	2021-10-07	113712128	134096610
26	2021-10-08	131682843	133372480
27	2021-10-09	146310488	137253882
28	2021-10-10	168332330	146577417
29	2021-10-11	180602398	156784911
30	2021-10-12	190004425	166750765
31	2021-10-13	169654534	167621896
32	2021-10-14	205967937	179125708
33	2021-10-15	173620876	177474259
34	2021-10-16	204783318	185666976
35	2021-10-17	218558821	195534530
36	2021-10-18	225350001	204479171
37	2021-10-19	193049762	201050348
38	2021-10-20	179647060	194629362
39	2021-10-21	176104475	189071896
40	2021-10-22	206242037	194222938

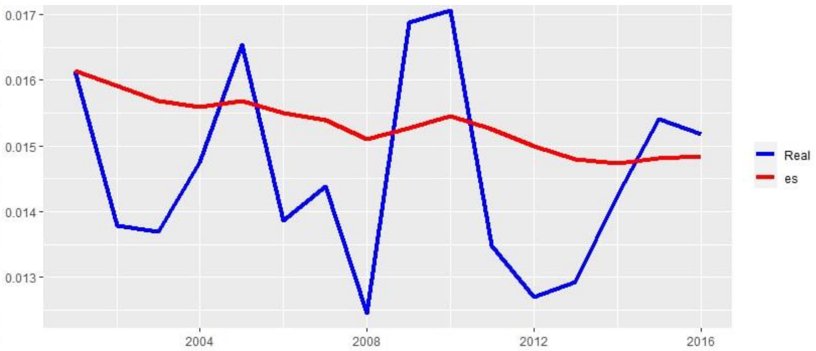


Figure 23: Exponentially smoothed data, alpha=0.1

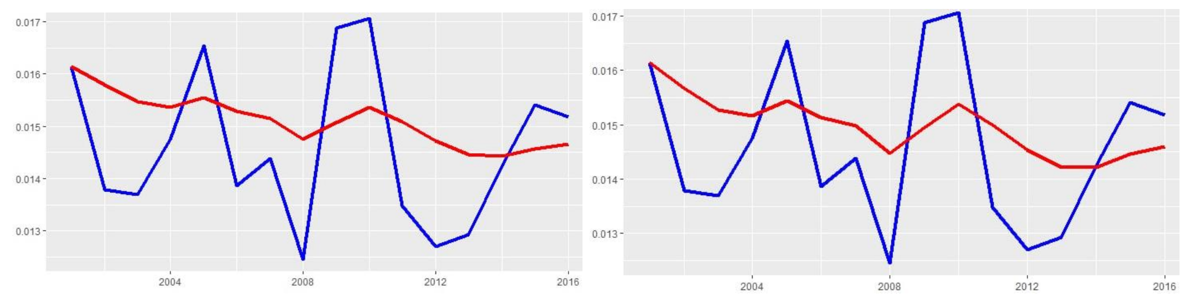


Figure 24: Visualization of smoothed data at alpha = 0.15 and alpha = 0.2

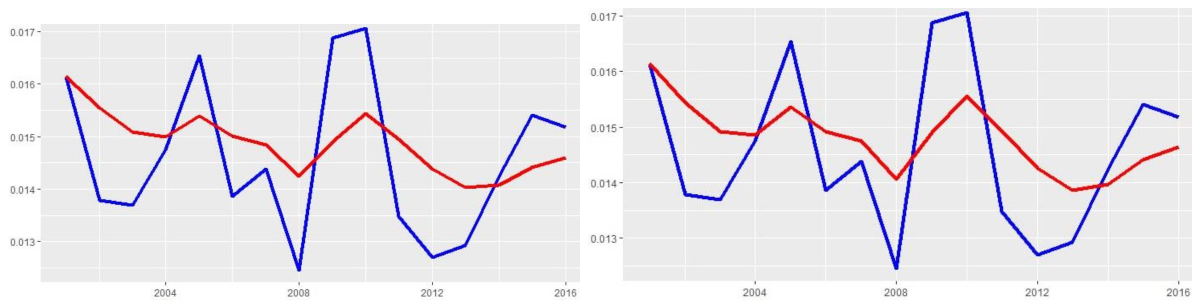


Figure 25: Visualization of smoothed data at alpha = 0.25 and alpha = 0.3

Median smoothing completely removes single extreme or anomalous values of levels that are separated from each other by at least half of the smoothing interval; preserves sharp changes in the trend (moving average and exponential smoothing smooth them); effectively removes single levels with very large or very small values that are random in nature and stand out sharply from other levels.

Year	avg	med_fil
2001	0.01614256	0.01576454
2002	0.01378499	0.01378499
2003	0.01369551	0.01378499
2004	0.01474683	0.01474683
2005	0.01654408	0.01474683
2006	0.01386202	0.01438157
2007	0.01438157	0.01386202
2008	0.01245726	0.01438157
2009	0.01688389	0.01688389
2010	0.01705961	0.01688389
2011	0.01348141	0.01348141
2012	0.01269789	0.01292396
2013	0.01292396	0.01292396
2014	0.01425047	0.01425047
2015	0.01541201	0.01517973
2016	0.01517973	0.01541203

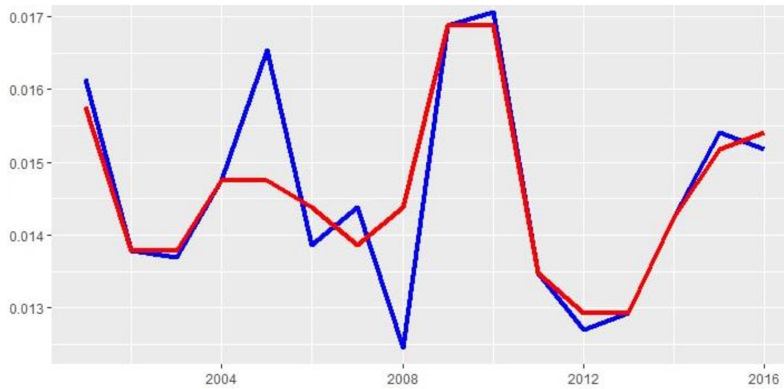


Figure 26: Median filtering

As can be seen from fig. 26, median filtering removed random levels that are random in nature. As a result, we have a more stable schedule.

```
> tp_mf<-turnpoints(med_fil)
> summary(tp_mf)
Turning points for: med_fil

nbr observations  : 16
nbr ex-aequos     : 4
nbr turning points: 5 (first point is a pit)
E(p) = 9.333333 Var(p) = 2.522222 (theoretical)

  point type       proba       info
1     3  pit 0.66666667 0.5849625
2     5 peak 0.25000000 2.0000000
3     7  pit 0.10000000 3.3219281
4    10 peak 0.10000000 3.3219281
5    13  pit 0.02777778 5.1699250
>
> plot(ma$avg, type = "l")
> lines(tp_mf)
>
> cor(ma$avg,med_fil)
[1] 0.8754431
```

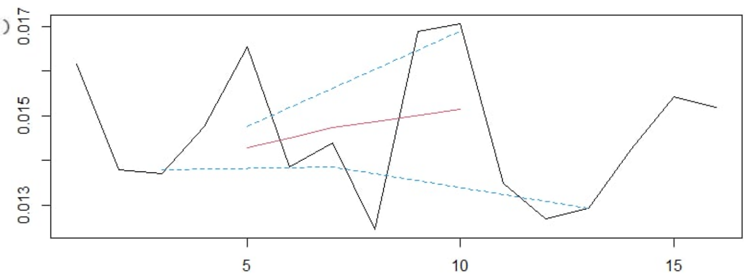


Figure 27: Turning points and visualization of turning points in median filtering

```
> tp_mf<-turnpoints(med_fil)
> summary(tp_mf)
Turning points for: med_fil

nbr observations  : 16
nbr ex-aequos     : 4
nbr turning points: 5 (first point is a pit)
E(p) = 9.333333 Var(p) = 2.522222 (theoretical)

  point type       proba       info
1     3  pit 0.66666667 0.5849625
2     5 peak 0.25000000 2.0000000
3     7  pit 0.10000000 3.3219281
4    10 peak 0.10000000 3.3219281
5    13  pit 0.02777778 5.1699250
>
> plot(ma$avg, type = "l")
> lines(tp_mf)
>
> cor(ma$avg,med_fil)
[1] 0.8754431
```

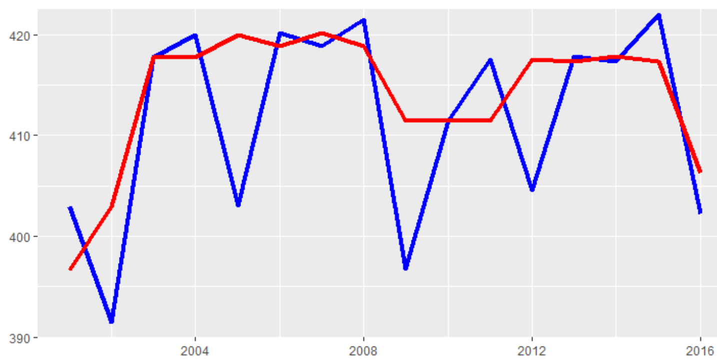


Figure 28: Pivot points, correlation coefficient, and median filtering for the W field

From fig. 30, it can be seen that the average number of words without repetitions remains approximately at the same level. This means that the "jumps" of the graph are not so important, but are only isolated cases and simply related to the texts. Note that the correlation is high, because the median filtering does not calculate, does not generalize, but shows the median on a certain interval. That is why median filtering is very effective when studying time series.

6. Discussion

We will investigate in detail the dependence of attributes on the basis of correlation analysis of time sequences. To do this, we will construct multiple correlation graphs to find the most significant variables by analyzing correlation relations and construct correlation graphs of the most significant variables found.

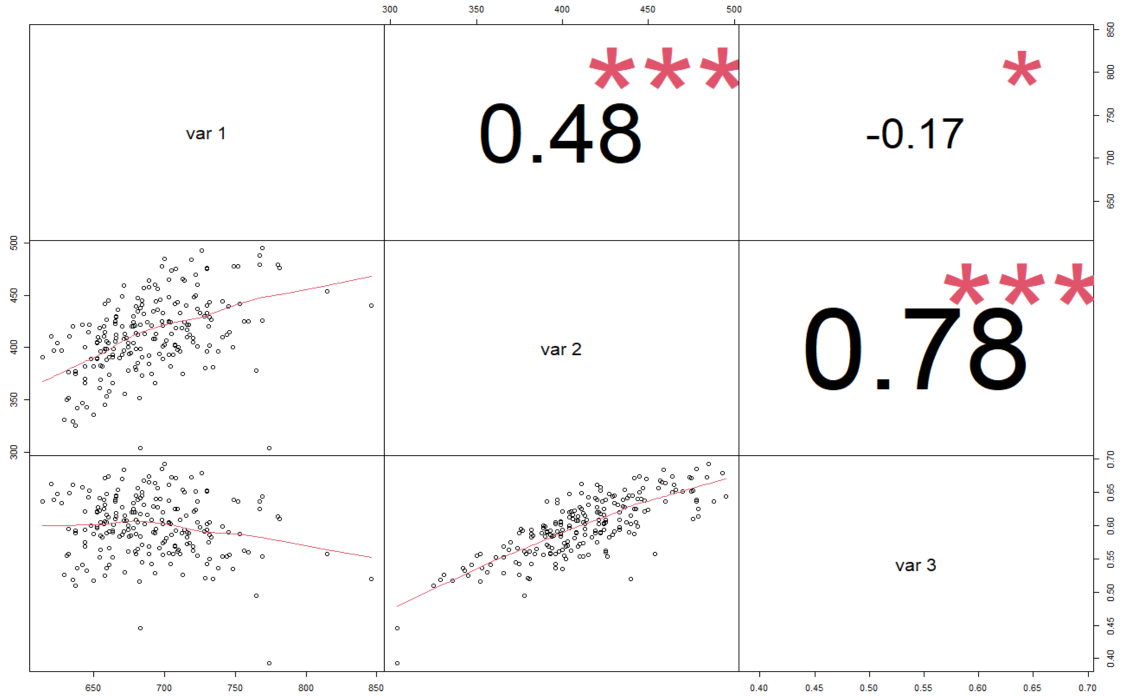


Figure 29: Multiple correlation

This visualization is built on three attributes, namely "total number of words of this text" (var 1), "number of words in a certain text (without repetitions)" (var 2), "Lexical diversity" (var 3). It can be seen from this visualization that the correlation coefficient between the second and third variables is the largest, so let's take a closer look at their correlation graph:

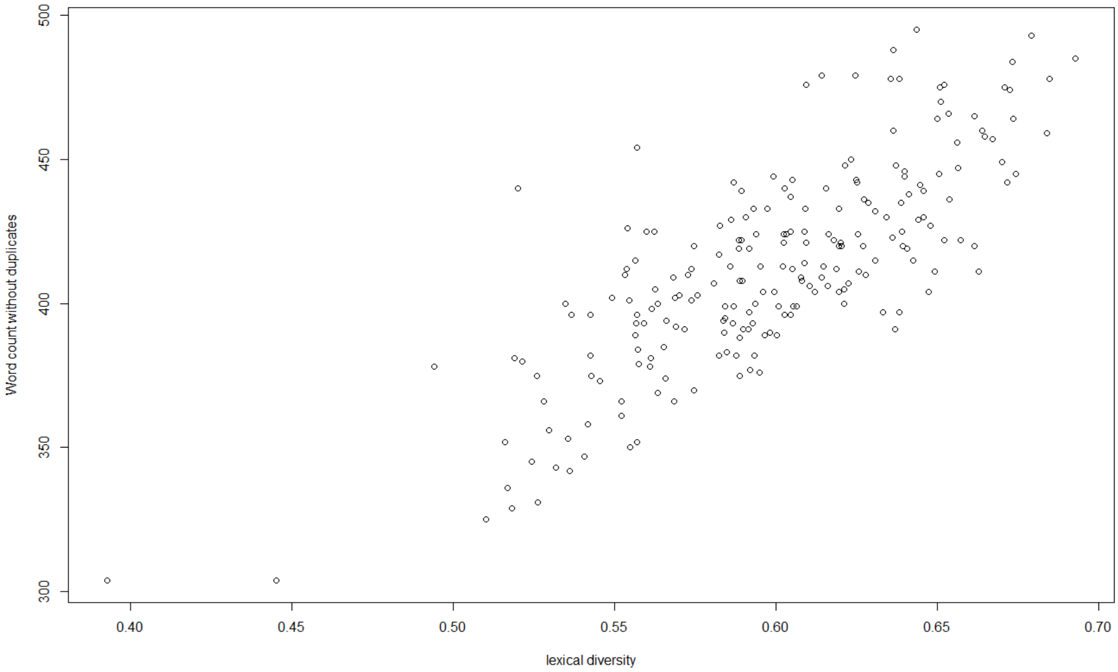


Figure 30: Graph of the correlation relation

The graph shows the linear dependence of the variables - when one variable grows, the other grows accordingly.

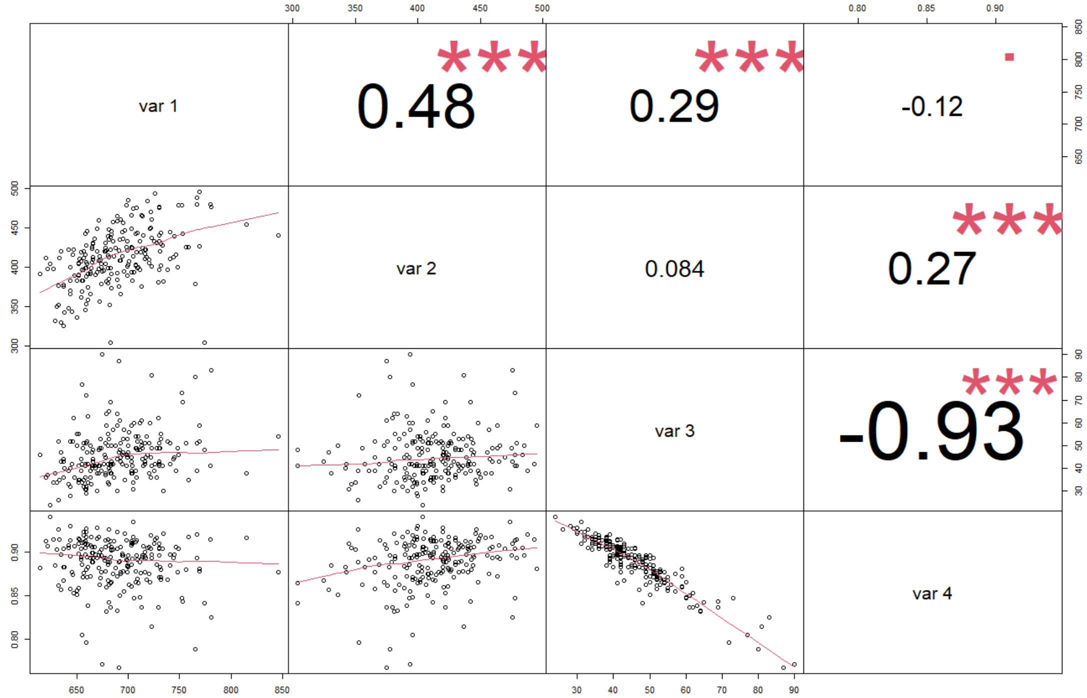


Figure 31: Multiple correlation

This visualization is built on four attributes, namely the total number of words of the text (var 1), the number of words in a certain text (without repetitions) (var 2), the number of separate sentences (var 3) and Syntactic complexity (var 4). It can be seen from this visualization that the correlation coefficient between the third and fourth variables is the most significant, so let's take a closer look at their correlation graph:

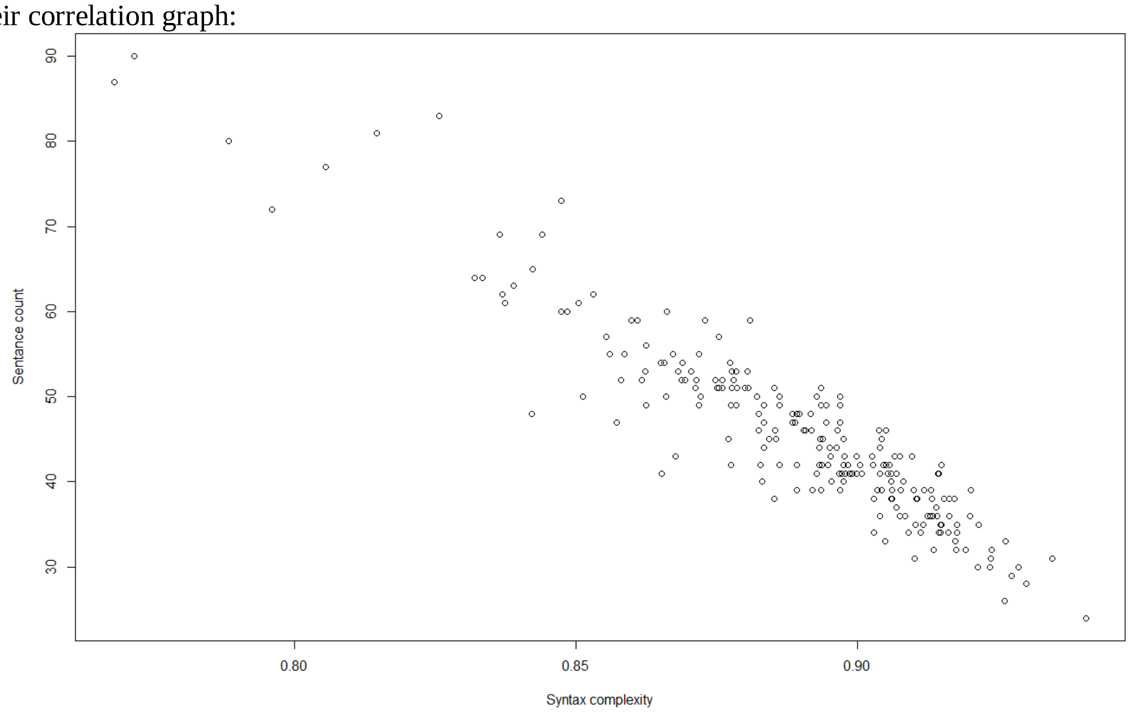


Figure 32: Graph of the correlation relation

Thanks to the graph, you can make sure that when the dependent variable increases, the independent variable drops rapidly, which corresponds to this negative correlation coefficient.

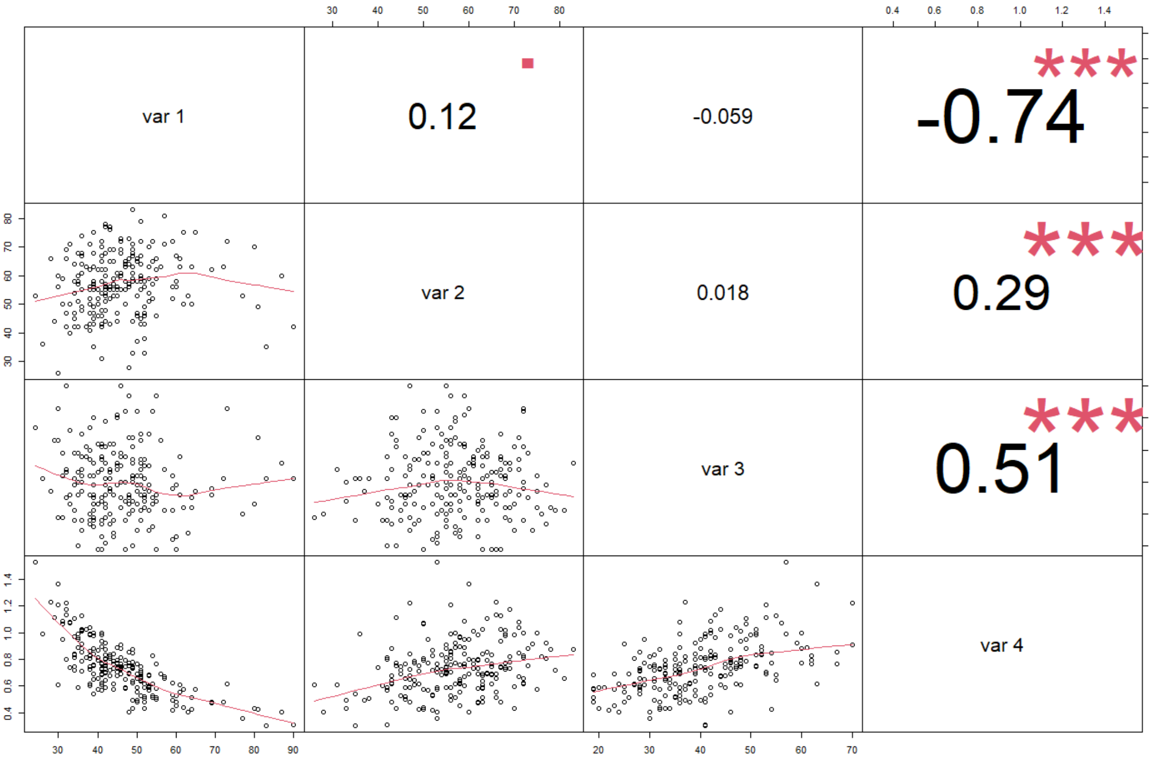


Figure 33: Multiple correlation

This visualization is built on four attributes, namely the number of separate clauses (var 1), the number of prepositions (var 2), the number of conjunctions (var 3) and the Speech Coherence Factor (var 4). It can be seen from this visualization that the correlation coefficient between the first and fourth variables is the most significant, so let's take a closer look at their correlation graph:

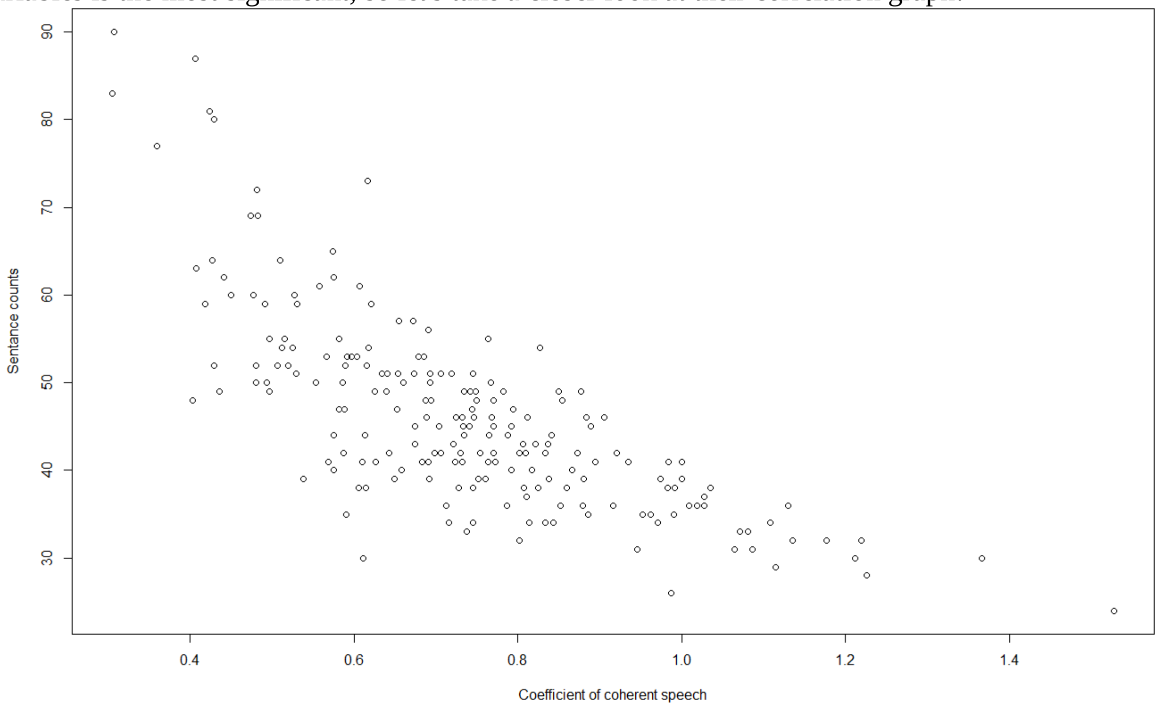


Figure 34: Graph of the correlation relation

This graph visualizes the almost identical logic of dependence as in the previous case.

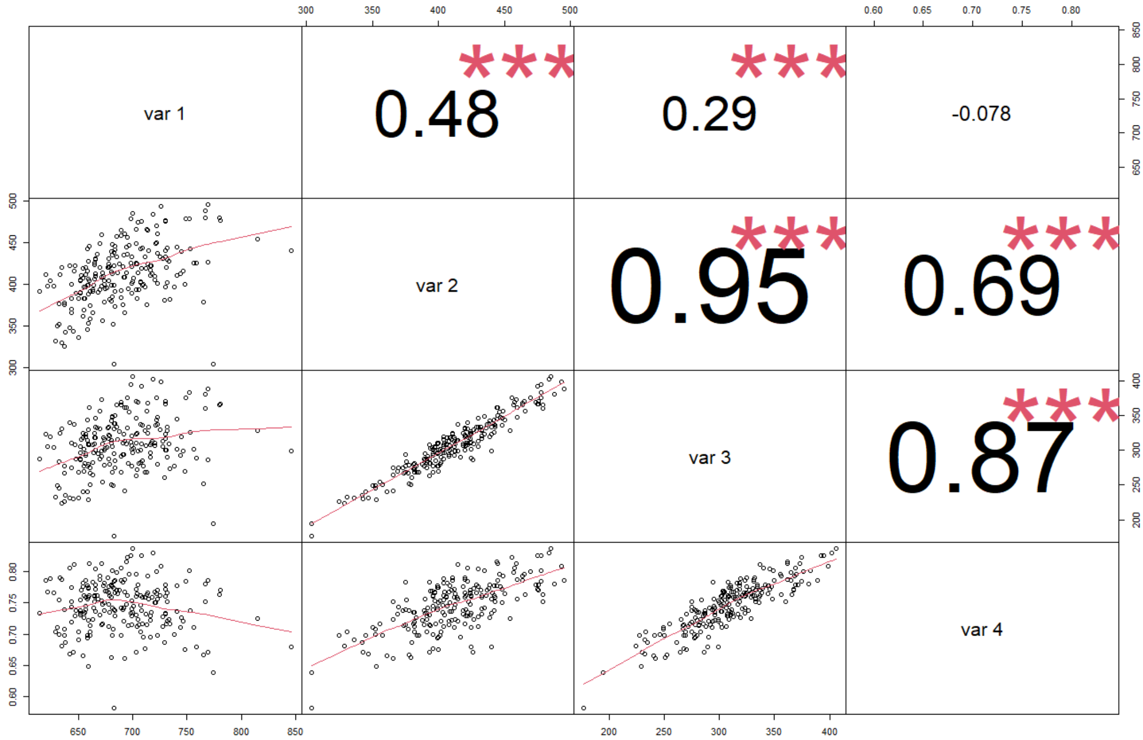


Figure 35: Multiple correlation

This visualization is built on four attributes, namely the total number of words of this text (var 1), the number of words in a specific text (without repetitions) (var 2), the number of words with a frequency of 1 (var 3), and the Uniqueness Index (var 4). It can be seen from this visualization that the correlation coefficient between the third and fourth variables is the most significant, so let's take a closer look at their correlation graph:

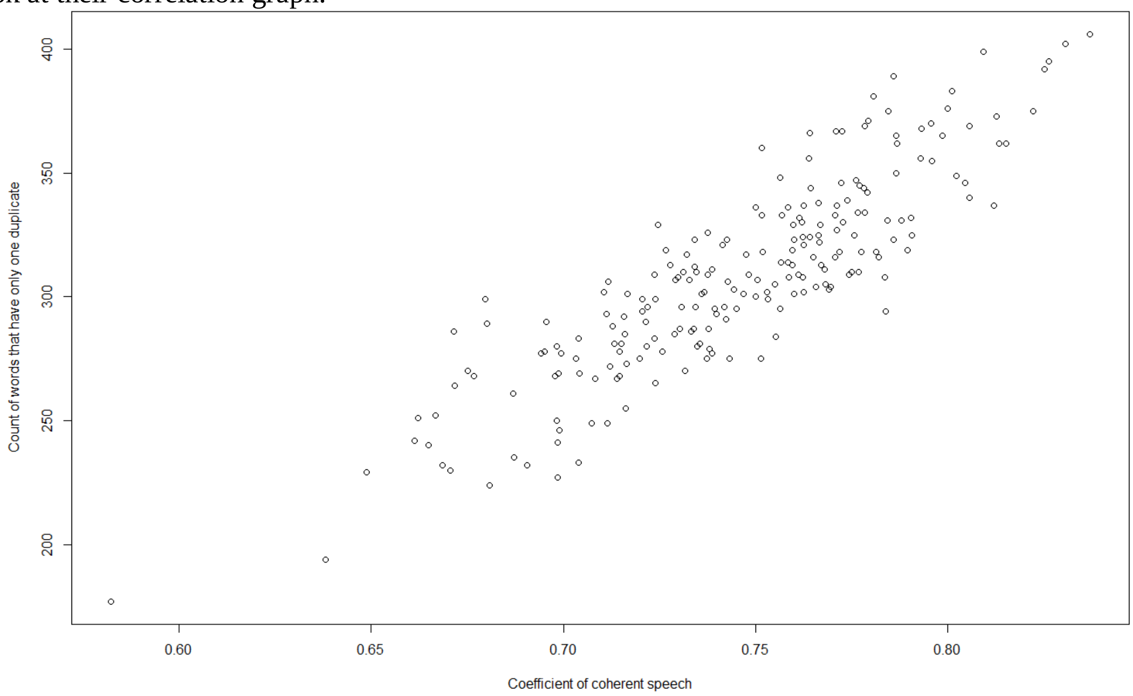


Figure 36: Graph of the correlation relation

The graph shows the linear dependence of the variables - as one variable grows, the other grows, which is why the positive correlation coefficient shows.

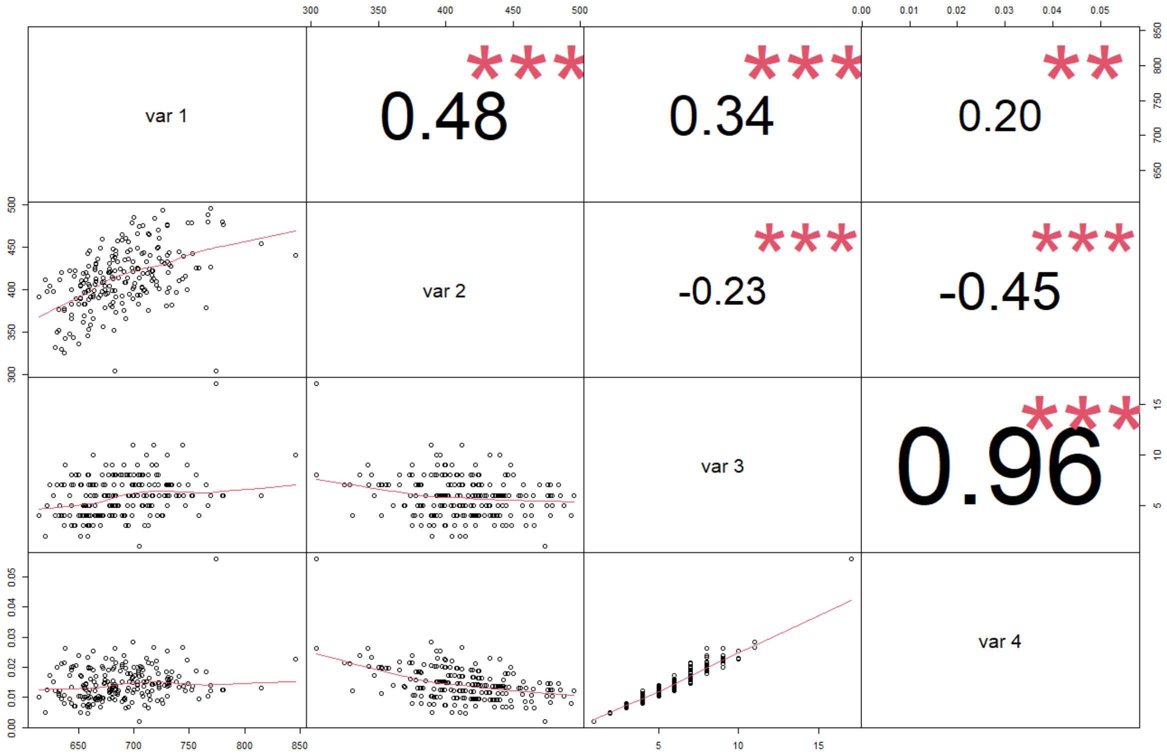


Figure 37: Multiple correlation

This visualization is built on four attributes, namely the total number of words of this text (var 1), the number of words in a certain text (without repetitions) (var 2), the number of words with a frequency of 10 or more (var 3) and the Concentration Index (var 4). It can be seen from this visualization that the correlation coefficient between the third and fourth variables is the most significant, so let's take a closer look at their correlation graph:

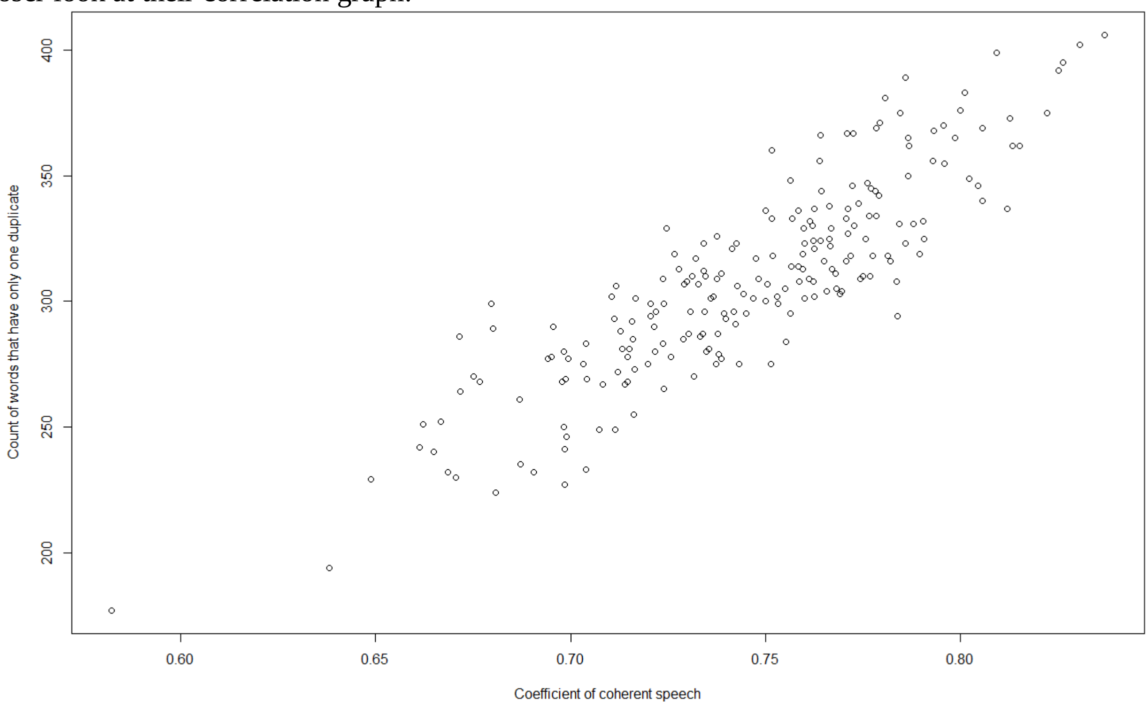


Figure 38: Graph of the correlation relation

The graph shows the linear dependence of the variables - as one variable grows, the other grows, which is why the positive correlation coefficient shows.

7. Conclusions

A simple moving average is suitable for identifying trends in the past, which will help us predict the future with less error. This will allow us to predict how succinct texts will be in the future. To do this, they need methods that quickly respond to the latest data. When performing work on such methods, we used exponential smoothing.

During data analysis, it was found that the larger the text, the fewer words it contains without repetitions, which is logical, since it is difficult to pick up new words every time. Over time, the number of words without repetitions does not increase and does not decrease significantly, although it is not immediately visible on the graph. We reached this conclusion using median filtering

It is also worth noting that the relationship between the number of words, the number of words without repetitions, lexical diversity, syntactic complexity, the coefficient of speech coherence, the exclusivity index and the concentration index was investigated. There is a direct relationship between them, so when one of these attributes increases, the others will also increase.

8. References

[1] V. Lytvyn, V. Vysotska, P. Pukach, Z. Nytrebych, I. Demkiv, A. Senyk, O. Malanchuk, S. Sachenko, R. Kovalchuk, N. Huzyk, Analysis of the developed quantitative method for automatic attribution of scientific and technical text content written in Ukrainian, volume 6(2-96) of Eastern-European Journal of Enterprise Technologies, 2018, pp. 19-31. DOI: 10.15587/1729-4061.2018.149596.

[2] N. Sharonova, I. Kyrychenko, I. Gruzdo, G. Tereshchenko, Generalized Semantic Analysis Algorithm of Natural Language Texts for Various Functional Style Types, CEUR Workshop Proceedings, Vol-3171 (2022) 16-26.

[3] N. Hrytsiv, T. Shestakevych, J. Shyyka, Quantitative Parameters of Lucy Montgomery's Literary Style, CEUR Workshop Proceedings, Vol-2870 (2021) 670-684.

[4] I. Khomytska, V. Teslyuk, Authorship and Style Attribution by Statistical Methods of Style Differentiation on the Phonological Level, Advances in Intelligent Systems and Computing 871 (2019) 105–118. doi: 10.1007/978-3-030-01069-0_8.

[5] I. Khomytska, V. Teslyuk, A. Holovatyy, O. Morushko, Development of Methods, Models and Means for the Author Attribution of a Text. Eastern-European Journal of Enterprise Technologies 3/2 (93) (2018) 41–46.

[6] I. Khomytska, V. Teslyuk, The Method of Statistical Analysis of the Scientific, Colloquial, Belles–Lettres and Newspaper Styles on the Phonological Level. Advances in Intelligent Systems and Computing, 512 (2017) 149–163. doi: 10.1007/978–3–319–45991–2_10.

[7] I. Khomytska, V. Teslyuk, Specifics of Phonostatistical Structure of the Scientific Style in English Style System, in Proceedings of the XIth Scientific and Technical Conference on CSIT, Lviv, 2016, pp. 129–131. doi: 10.1109/stc–csit.2016.7589887.

[8] S. Buk, Osnovy statystychnoi lingvistyky, Lviv, 2008.

[9] A. Berko, Y. Matseliukh, Y. Ivaniv, L. Chyrun, V. Schuchmann, The text classification based on Big Data analysis for keyword definition using stemming, in: proceedings of IEEE 16th International conference on computer science and information technologies, Lviv, Ukraine, 22–25 September, 2021, pp. 184–188.

[10] O. Hladun, A. Berko, M. Bublyk, L. Chyrun, V. Schuchmann, Intelligent system for film script formation based on artbook text and Big Data analysis in: proceedings of IEEE 16th International conference on computer science and information technologies, Lviv, Ukraine, 22–25 September, 2021, pp. 138–146.

[11] V. Perebyinis, Matematychna linhvistyka. Ukrainska mova. Kyiv, 2000, 287–302.

[12] V. Perebyinis, Statystychni metody dlia linhvistiv. Vinnytsia, 176, 2013.

[13] A. Dyriv, V. Andrunyk, Y. Burov, I. Karpov, L. Chyrun, The user's psychological state identification based on Big Data analysis for person's electronic diary, in: proceedings of IEEE 16th International conference on computer science and information technologies, Lviv, Ukraine, 22–25 September, 2021, pp. 101–112.

[14] D. Lande, V. Zhyhalo, Pidkhid do rishennia problem poshuku dvomovnoho plahiatu. Problemy informatyzatsii ta upravlinnia 2 (24), (2008) 125–129.
[15] O. Oborska, V. Andrunyk, L. Chyrun, R. Hasko, A. Vysotskyi, S. Mushasta, O. Petruchenko, I. Shakleina, The Intelligent System Development for Psychological Analysis of the Person's Condition, CEUR Workshop Proceedings, Vol-2870 (2021) 1390-1419.
[16] Victana. URL: http://victana.lviv.ua/nlp/linhvometriia
[17] C. Boyer, L. Dolamic, N. Grabar, Automated Detection of Health Websites' HONcode Conformity: Can N-gram Tokenization Replace Stemming? Studies in Health Technology and Informatics 216 (2015) 1064.
[18] A. Dmytriv, S. Holoshchuk, L. Chyrun, R. Holoshchuk, Comparative Analysis of Using Different Parts of Speech in the Ukrainian Texts Based on Stylistic Approach, CEUR Workshop Proceedings, Vol-3171 (2022) 546-560.
[19] S. Kubinska, R. Holoshchuk, S. Holoshchuk, L. Chyrun, Ukrainian Language Chatbot for Sentiment Analysis and User Interests Recognition based on Data Mining, CEUR Workshop Proceedings, Vol-3171 (2022) 315-327.
[20] N. Kholodna, V. Vysotska, O. Markiv, S. Chyrun, Machine Learning Model for Paraphrases Detection Based on Text Content Pair Binary Classification, CEUR Workshop Proceedings, Vol-3312 (2022) 283-306.
[21] V. Hryhorovych, Analysis of Scientific Texts by Semantic Inverse-Additive Metrics for Ontology Concepts, CEUR Workshop Proceedings, Vol-3171 (2022) 801-816.
[22] S. Albota, Modelling the Impact of the Pandemic on Online Communication: Textual Semantic Analysis, CEUR Workshop Proceedings, Vol-3171 (2022) 471-486.
[23] N. Kunanets, Y. Oliinyk, D. Myhal, K. Shunevych, A. Rzheuskyi, Y. Shcherbyna, Enhanced LSA Method with Ukraine Language Support, CEUR Workshop Proceedings 2870 (2021) 129-140.
[24] B. Mobasher, Data mining for web personalization. The adaptive web, (2007) 90–135. doi: 10.1007/978–3–540–72079–9_3.
[25] C. E. Dinucă, D. Ciobanu, Web Content Mining. Annals of the University of Petroşani. Economics 12 (1) (2012) 85–92.
[26] G. Xu, Y. Zhang, L. Li, Web Content Mining. Web Mining and Social Networking (2010) 71–87. doi: 10.1007/978–1–4419–7735–9_4.
[27] I. Khomytska, V. Teslyuk, Modelling of Phonostatistical Structures of English Backlingual Phoneme Group in Style System, in CADMS : Proceedings of the 14th International Conference. Polyana, 2017, pp. 324–327.
[28] I. Khomytska, V. Teslyuk, Modelling of Phonostatistical Structures of the Colloquial and Newspaper Styles in English Sonorant Phoneme Group, in CSIT : Proceedings of the XIIth Scientific and Technical Conference. Lviv, 2017, pp. 67–70.
[29] I. Khomytska, V. Teslyuk, Authorship Attribution by Differentiation of Phonostatistical Structures of Styles, in CSIT : Proceedings of the XIIIth Scientific and Technical Conference. Lviv, 2018, pp. 5–8.
[30] I. Khomytska, V. Teslyuk, The Software for Authorship and Style Attribution in CADMS : Proceedings of the 15th International Conference. Polyana, 2019,pp. 23–26.
[31] I. Khomytska, V. Teslyuk, Mathematical Methods Applied for Authorship Attribution on the Phonological Level, in CSIT : Proceedings of the XIVth Scientific and Technical Conference. Lviv, 2019, pp. 7–11.
[32] V. Vysotska, O. Markiv, S. Teslia, Y. Romanova, I. Pihulechko, Correlation Analysis of Text Author Identification Results Based on N-Grams Frequency Distribution in Ukrainian Scientific and Technical Articles. In CEUR Workshop Proceedings, Vol-3171 (2022) 277-314.
[33] C. Lu, Y. Bu, J. Wang, Y. Ding, V. Torvik, M. Schnaars, C. Zhang, Examining scientific writing styles from the perspective of linguistic complexity, Journal of the Association for Information Science and Technology 70(5) (2019) 462-475.
[34] B. Chen, D. Deng, Z. Zhong, C. Zhang, Exploring linguistic characteristics of highly browsed and downloaded academic articles. Scientometrics 122 (2020) 1769-1790.
[35] M. Lupei, A. Mitsa, V. Repariuk, V. Sharkan, Identification of authorship of Ukrainian-language texts of journalistic style using neural networks, Eastern-European Journal of Enterprise Technologies 1(2(103)) (2020) 30-36. doi: 10.15587/1729-4061.2020.195041.

[36] C. Luckman, S. A. Wagovich, C. Weber, B. Brown, S. E. Chang, N. E. Hall, N. B. Ratner, Lexical diversity and lexical skills in children who stutter, Journal of Fluency Disorders 63 (2020) 105747.
[37] A. Fronzetti Colladon, C. A. D'Angelo, P. A. Gloor, Predicting the future success of scientific publications through social network and semantic analysis, Scientometrics 124 (2020) 357-377.
[38] S. Kumar, M. Yadava, P. P. Roy, Fusion of EEG response and sentiment analysis of products review to predict customer satisfaction. Information Fusion, 52 (2019) 41-52.
[39] V.V. Hnatushenko, P. I. Kogut, M. V. Uvarov, On Optimal 2-D Domain Segmentation Problem via Piecewise Smooth Approximation of Selective Target Mappings. Journal of Optimization, Differential Equations and Their Applications 27(2) (2019). 60–95. DOI: 10.15421/141908.
[40] Hnatushenko V., Kogut P., Uvarov M. On Satellite Image Segmentation via Piecewise Constant Approximation of Selective Smoothed Target Mapping, Applied Mathematic
[41] L. P. Morency, R. Mihalcea, P. Doshi, Towards multimodal sentiment analysis: Harvesting opinions from the web. In Proceedings of the 13th international conference on multimodal interfaces, 2011, pp. 169-176.
[42] N. Romanyshyn, Algorithm for Disclosing Artistic Concepts in the Correlation of Explicitness and Implicitness of Their Textual Manifestation, CEUR Workshop Proceedings 2870 (2021) 719-730.
[43] Y. Yusyn, T. Zabolotnia, Methods of Acceleration of Term Correlation Matrix Calculation in the Island Text Clustering Method, CEUR workshop proceedings, Vol-2604 (2020) 140-150.
[44] B. Rusyn, V. Ostap, O. Ostap, A correlation method for fingerprint image recognition using spectral features, in: Proceedings of the International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science, TCSET, 2002, pp. 219–220.
[45] S. Voloshyn, O. Markiv, V. Vysotska, I. Dyyak, L. Chyrun, V. Panasyuk, Emotion Recognition System Project of English Newspapers to Regional E-Business Adaptation, in: IEEE 17th International Conference on Computer Sciences and Information Technologies (CSIT), 2022, pp. 392-397, doi: 10.1109/CSIT56902.2022.10000527.
[46] N. Kholodna, V. Vysotska, S. Albota, A Machine Learning Model for Automatic Emotion Detection from Speech, CEUR Workshop Proceedings, Vol-2917 (2021) 699-713.
[47] M. Hryntus, M. Dilai, Translating emotion metaphors from English into Ukrainian: based on the parallel corpus of fiction, CEUR Workshop Proceedings, Vol-3171 (2022) 737-750.
[48] O. Bisikalo, V. Kovenko, I. Bogach, O. Chorna, Explaining Emotional Attitude Through the Task of Image-captioning, CEUR Workshop Proceedings, Vol-3171 (2022) 1056-1065.
[49] K. Smelyakov, O. Bohomolov, M. Kizitskyi, A. Chupryna, Identification of Modern Facial Emotion Recognition Models, CEUR Workshop Proceedings, Vol-3171 (2022) 1267-1281.
[50] D. Nazarenko, I. Afanasieva, N. Golian, V. Golian, Investigation of the Deep Learning Approaches to Classify Emotions in Texts, CEUR Workshop Proceedings, Vol-2870 (2021) 206-224.
[51] I. Bekhta, N. Hrytsiv, Computational Linguistics Tools in Mapping Emotional Dislocation of Translated Fiction, CEUR Workshop Proceedings, Vol-2870 (2021) 685-699.
[52] I. Spivak, S. Krepych, O. Fedorov, S. Spivak, Approach to Recognizing of Visualized Human Emotions for Marketing Decision Making Systems, CEUR Workshop Proceedings, Vol-2870 (2021) 1292-1301.
[53] P.C. Thoumelin, N. Grabar, Subjectivity in the medical discourse: On uncertainty and emotional markers, Revue des Nouvelles Technologies de l'Information, E.26 (2014) 455–466.
[54] N. Grabar, L.O. Dumonet, Automatic computing of global emotional polarity in French health forum messages, Lecture Notes in Computer Science 9105 (2015) 243–248.