This is a contribution from Primenjena lingvistika, No.10, pp. 103-113, 2009

EXTRACTING INFORMATION FROM LARGE DIGITAL CORPORA - *A CASE STUDY IN QUANTITATIVE METHODS IN LINGUISTICS*

Nikola Dobrić

Alpen-Adria Universität Klagenfurt

Abstract:

Every empirical research into language should not only be based on concrete data but it also has to be verified as scientifically relevant. In all scientific areas, including linguistics, such verification is performed by implementing a selection of statistical test in order to examine the significance, distribution and variation of the obtained research results. The first part of the paper presents the necessary procedures when extracting linguistic data from corpora. The corpus used is of exemplary nature and it is created from Dostoyevsky's *Crime and Punishment* (in Russian, English, Serbian and German), in electronic form. The paper further displays the most common quantitative methods used in language analysis, which are all illustrated by clear examples from the small study in this paper, in order to make the complicated statistical calculi understandable and readily useable. In the end, the importance of linguistic statistics in modern language research is further emphasized.

Key words: corpus, statistics, distribution, variation, significance

1. IMPORTANCE OF STATISTICS IN LINGUISTICS

It is a self-evident fact that most scientific disciplines use statistic methods in data processing in order to validate and substantiate their respective research results. Since it also deals in empirical data linguistics should be no different. As is well known, some linguistic patterns are regular and independent of the speaker, writer, or subject matter. Hence, linguistic behavior conforms closely to certain expectations: namely, quantitative or statistical patterns. However, the case of their active application often proves quite to the contrary. Among the linguists, any kind of statistical processing, sometimes even the most rudimentary one, is ignored or avoided, whether due to the lack of training, fear and/or dislike of the procedures or simple unawareness of their importance. The truth is that such statistical processing is of outmost importance in accurately presenting results of any empirical insight into language data. Pragmatics, semantics and even orthodox syntacticians, morphologists and phonologists alike can benefit greatly from utilizing such procedures.

One thing that should encourage linguists to use them more is the fact that for us it is supposed to be only an issue of methodology. A linguist needs not know why a particular statistical procedure is applied or what the mathematical reasons are for it representing what the statisticians say it does. For a linguist, the only knowledge necessary is to know when to implement a particular procedure. Most of such most applicable ones deal with frequency counts and percentages and range from the most basic univariate analysis (descriptive statistics); through bivariate analysis and correlation productmoment; to the complex multivariate analysis (such as regression, cluster analysis, principle component analysis, etc.). It becomes even simpler given that all of the necessary calculations can be done by specialized statistical software or by, if need be, hiring an expert in statistic to aid you. So, to reiterate, the only knowledge necessary to a linguist is which statistical method is used with what kind of data and under what kind of circumstances. And that is not so difficult to find out, since many linguists and statisticians alike had bravely paved a way investigating and standardizing certain fixed quantitative methods to be applied in linguistic analyses.

That is why the emphasis of this paper remains on methodology alone and its aim is to illustrate some potential applications to linguistic research rather than present results of actual in-depth research. All of the empirical data presented in the paper, although being a product of actual analysis, present only example data and should not be interpreted as finalized research results.

2. EXAMPLE CORPORA

A corpus is a collection of pieces of text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research (Sinclair, 2005). There are many kinds of corpora (Dobrić, 2009a) available online, but for the purposes of the paper a small corpus was manually comprised in order to facilitate greater simplicity and ease of comprehension. Parts of the first three chapters of Fyodor Dostoyevsky's *Crime and Punishment* in Russian, in electronic format, were comprised together to serve as a small example of a general corpus. The corpus was annotated automatically for sentence endings and number of words per sentence using a corpus tool called Unitex[1]. This corpus was used as a platform for presenting the most basic forms of statistical procedures and it is the case of descriptive statistics and univariate analysis of data.

For the purposes of demonstrating a bivariate analysis (Pearson's correlation in this case) a parallel corpus was also compiled out of parts of the first three chapters of *Crime and Punishment* in Russian, Serbian, English and German respectively. All the sentences have been paired up using a

[1] For more about Unitex see Vitas et al (2007).

combination of a manual and automatic approach to text alignment. The corpus was also annotated for sentence ending and number of words per sentence using Unitex. Then the texts were automatically aligned two by two (each of the translated texts with the original independently), using Unitex again. Thusly obtained parallel texts were then manually aligned to create four aligned texts (the Russian original plus three translations, as defined previously). One of the problems that occur in translation aligning is how to deal with sentences translated from one original sentence into two in some language or vice versa. A notion of the *translation unit* was used to signify the original sentence as transferred through different translations. The translation unit is referred to here as used in translation software terminology signifying the basic unit of cognition in the original thought spanning usually between two points of interpunction. This was important to define due to issues when one has to count frequency based on sentence numbers.

RUSSIAN	SERBIAN	ENGLISH	GERMAN
Послышавшийся за дверью шум вдруг быстро увеличился, и дверь немного приотворилась.	Šum koji je bio čuo iza vrata naglo se pojačao tarker se malko odškrinula.	The noise behind the door increased, and suddenly the door was opened a little.	Das Geräusch, das hinter der Tür vernehmbar geworden war, wurde schnell tarker, unddie Tür wurde ein wenig geöffnet.
Что такое? крикнул с досадой Порфирий Петрович.	Što je to? srdito je viknuo Porfirij Petrovič.	What is it? cried Porfiry Petrovitch, annoyed.	Was gibt es? rief Porfirij Petrowitsch ärgerlich.

Table 1. Example of the constructed parallel corpus.

3. UNIVARIATE STATISTICS

As the term suggest univariate analyses are designed to deal with a single string of data (as opposed to bivariate for instance which are, as will be shown, designed to compare two strings of data) and such methods are also called descriptive statistics. It is concerned with simple processing of basic frequency counts, but can still be immensely useful in sorting through one's empirical data and arriving at useful results. The described general corpus designed as an example for this paper was used to demonstrate one such application of descriptive statistics.

As with most empirical research, one usually starts with a hypothesis, which is to be tested by examining certain data. An example hypothesis posed here, for the purposes of the presentation only, is that the original Russian version of *Crime and Punishment* will display uniformity in the number of sentences in a chapter, each sentence in turn containing a similar number of words, thusly showing (or not, if proven differently), a stylistic device consciously employed by the writer aimed at maintaining

a steady pace in his storytelling. The corpus, containing parts of the first three chapter, was analyzed[2] for the number of sentences containing between 5 and 50+ words. The result was further divided in sections of five words (0-4, 5-9, 10-14, …, 50+) interval sentences. The results are shown in Figure 1.

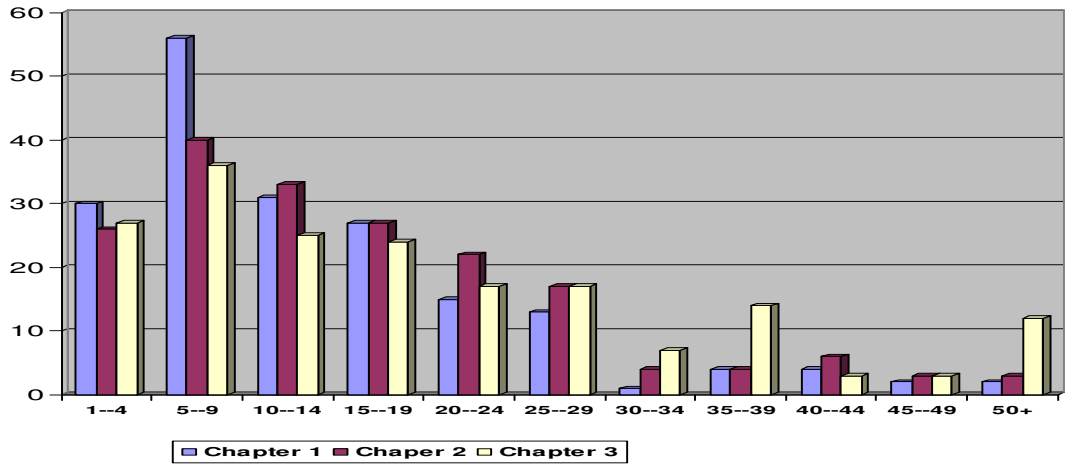


Figure 1. Numbers and sections per chapters.

The results in the first three chapters show that, apart from the sections 5–9, 35–39 and 50+ , all other sentences show relative uniformity in the number of words they contain across the three analyzed chapters. Such a distribution may be pointing out a feature of the writer's style and his manner in which he maintains, as it was stipulated, the pace of his storytelling throughout the book.

However, to test the hypothesis further, extending to the whole of the chapters, and to acquire stronger evidence for laying such claims we have to apply a simple statistical method called the *coefficient of variation* (CV). It will show us, with statistical certainty, how lengths of sentences continue to behave in the course of the whole first three chapters (the hypothesis being that chapters will display a similar number of sentences containing a similar number of words). As was stated before, this is one of the most basic procedures, and it is based on comparisons of *standard deviations*[3] expressed in percentages. To arrive at a coefficient of variation different software is available. Microsoft Excel is the simplest choice, used also in this paper, (although one can use other more specialized software such as R or SPSS) where one needs employ three functions: the AVERAGE

[2] Various software solutions exist for such analyses. Since the annotation of the example corpus was done in Unitex it was consequently used. Apart from that, one can also use concordancers (for more on concordancers see Dobrić, 2009), or even Microsoft Excel, in this case, since there is a specially programmed function (LEN=LEN(TRIM(A1))-LEN(SUBSTITUTE(A1," ",""))+1) designed to facilitate counting numbers of words in corresponding sentences in a cell.

[3] Three main measures of variability are the *range*, the *variance* and the *standard deviation*. Range is the highest value minus the lowest value; variance is the distance of every data item from the mean (squared subtracted mean value from the given score); while the square root of variance is called the standard deviation. However, as is stated in the paper it is irrelevant for a linguist to know that only to know how to use statistical aids to arrive at the values of these notions.

function; the VAR function and the STDEV function. Standard deviation (STDEV) is then divided by mean and multiplied by 100. When completed, the procedures yield a numerical representation as seen in Table 2 bellow.

CHAPTER 1		CHAPTER 2		CHAPTER 3	
Mean	14	Mean	15.91	Mean	19.87
sample variance	174.29	sample variance	139.87	sample variance	260.79
Sample standard deviation	13.2	sample standard deviation	11.83	Sample standard deviation	16.15
coefficient of variation	94.26%	coefficient of variation	74.34%	coefficient of variation	81.27%

Table 2. Coefficients of variation

The results seen in the light of statistical processing now present us with a slightly different picture. The results vary more than the initial sample results have indicated. Regarding the variation within individual chapters, Chapter 1 displays most uniformity while Chapter 2 shows the biggest difference in the sentence distribution. As for the variation across chapters, difference is again significant between all three analyzed chapters. This new data seems not to support the previous supposition about the writer's style and conscious effort in maintaining pace.

4. BIVARIATE STATISTICS

The hypothesis (that chapters will display a similar number of sentences containing a similar number of words) investigated (and disproven) in the previous section was modified and extended to serve as an example of one common statistical procedure used when dealing with two strings of interdependent, but different data (in this case it being the different language texts aligned in pairs). The example hypothesis here was that chapters in the translated texts will display a similar number of sentences of similar length to the original Russian text. In this case one chapter (Chapter 24) was selected as a sample using tables of random selection (Mann, 1995)[4]. The chapter was aligned as described, in four languages, and was analyzed in the same manner as the small corpus described in the previous part of the paper for number of words per sentence for each of the languages, and every language was then compared to the original Russian one. The similarity of the sentence lengths in

[4] These are statistical tables designed to provide the selection of data with as much randomness as possible.

translation and the manner of their relationship is represented by the *Pearson productmoment correlation coefficient*.

Correlation is one of the most basic and the most often used bivariate statistical methods and it represents the degree to which two variables vary together and reflect each other and the link between themselves. To properly calculate the correlation coefficient for two variables the following quantities must first be found: the sum of all values of the first variable X; the sum of all values of the second variable Y; the sum of all squares of X; the sum of all squares of Y; and the sum of the products XY over all data pairs. The correlation coefficient *r* is then:

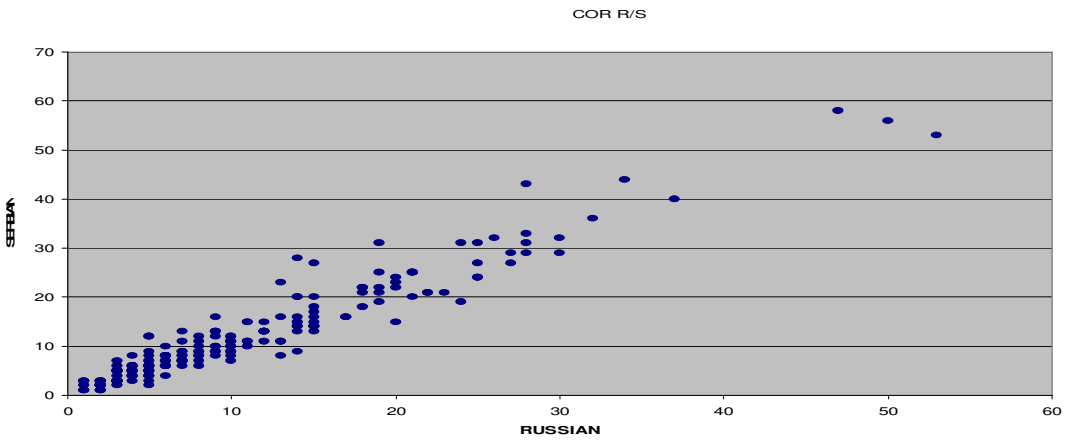
$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}$$

Again, not to be intimidated by the formula, we may call on mentioned software for aid. Using any of the commercially available statistical packages does simplify the procedure a lot. Microsoft Excel is yet again sufficient for such basics: the CORREL function calculates correlation automatically for any two selected columns of data. The correlation coefficient is expressed by a numerical value between -1 and + 1 and is interpreted as follows: +1 and its vicinity is called *positive correlation* and it is arrived at in cases when two variables vary together exactly (in the case of +1) or show a great degree of similarity; -1 its vicinity is referred to as *negative correlation* and is present when two variables vary in reverse proportion and show no or very little similarity in their occurrence; and a result 0 and around it is obtained when there is no correlation and no connection, neither negative nor positive, can be found between the occurrences of the two variables.

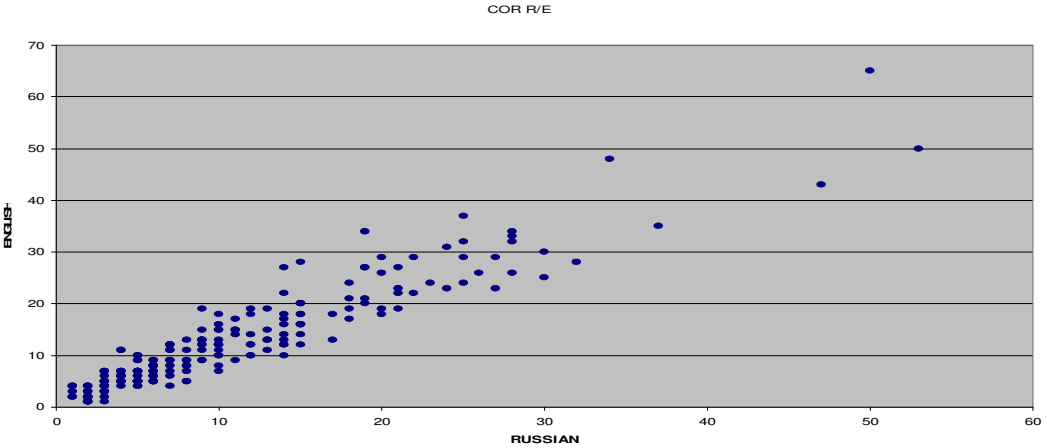
Correlation coefficient is usually presented using a *scatter plot* (Insert/Chart/Scatter Plot provides such graphic representation in Microsoft Excel). For its successful interpretation we have to know that the plot of variables moves up from bottom left to top right if there is a positive relationship and moves down from top left to bottom right if there is a negative relationship and has no uniform scatter pattern if there is no relationship. The tighter the points cluster around the perceived straight line, the stronger the relationship between the two variables is.

4.1 Positive correlation

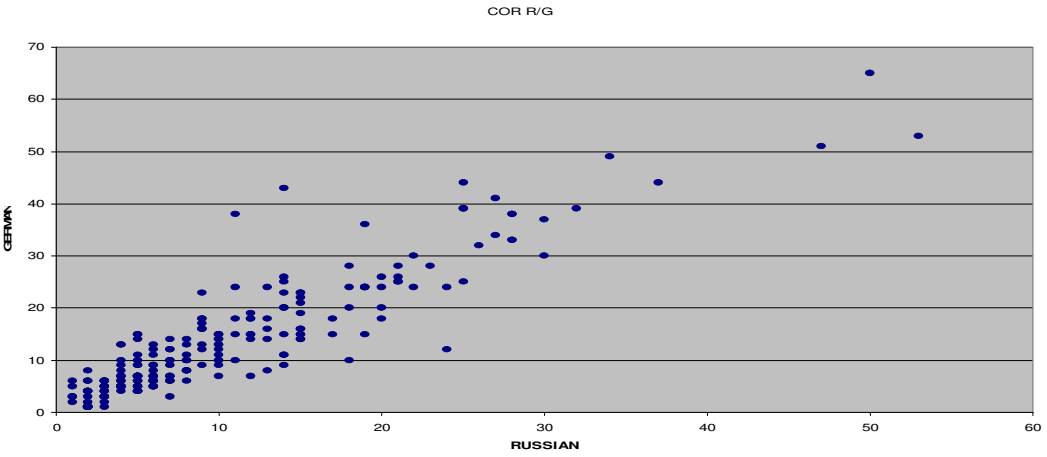
The following three graphs show three cases of positive correlation between the lengths og the sentences in the Russian–Serbian pair (Graph 1); Russian–English pair (Graph 2); and Russian–German pair (Graph 3):



Graph 1. Chapter 24 Russian – Serbian correlation 0.96.



Graph 2. Chapter 24 Russian – English correlation 0.94.



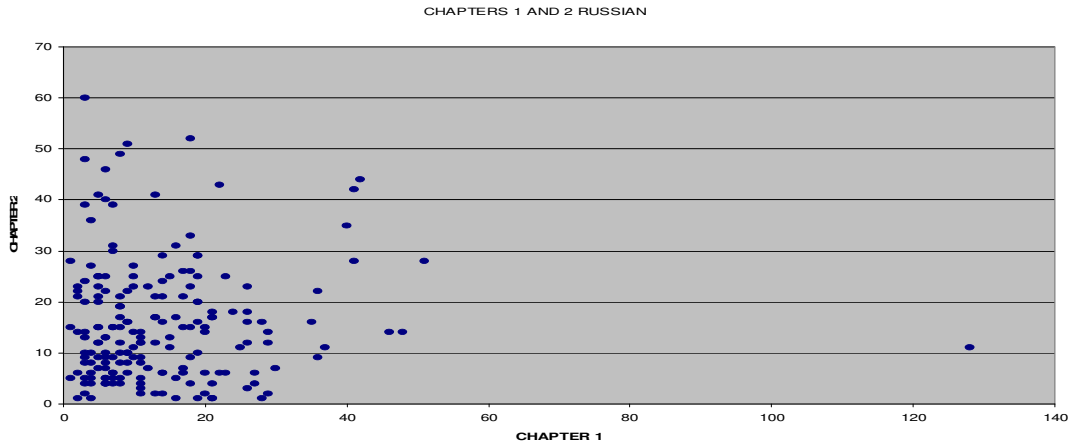
Graph 3. Chapter 24 Russian – German correlation 0.9.

Returning now to the example hypothesis in this part, all of the correlation coefficients seem to show a high degree of connectedness, indicating in such a way that the hypothesis might be proven as true. It was apparently possible for the translators to keep the original lengths and number of sentences, and thusly maintain the style and pace the writer intended to create. It is also interesting to note that the original Russian and the Serbian texts have the highest degree of similarity, perhaps allowing for a speculation that it is due to both of them being Slavic languages.

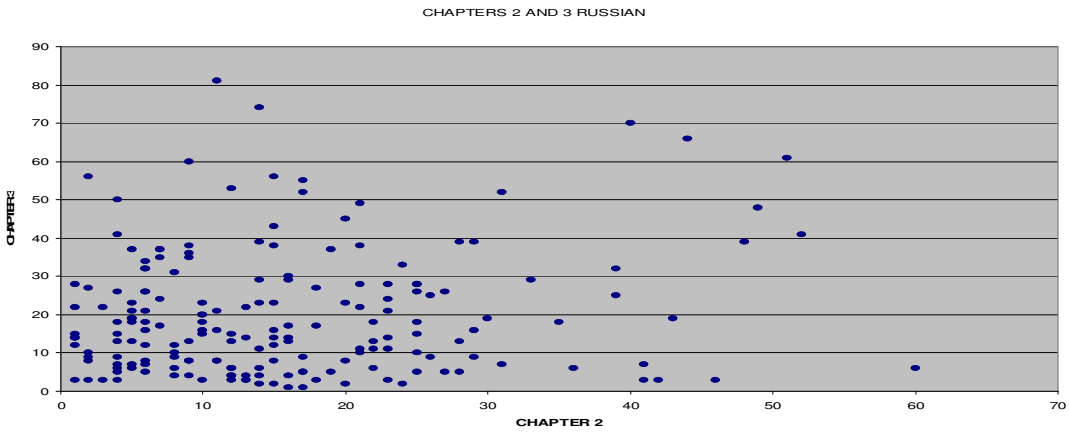
It is important to reiterate that the obtained and their elaborated analysis results regarding the actual correlation and proving the stated hypothesis are not conclusive and definite and should here only be taken as examples.

4.2 No correlation

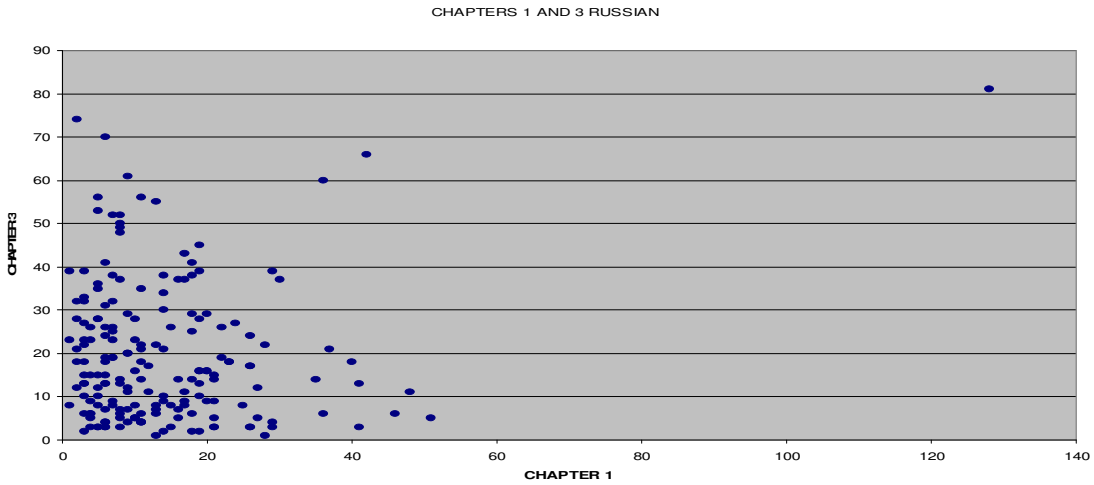
As was explained, when scores are not connected in any way and the result is 0 (and around zero) it is a case of no correlation and the data are not related in any way. To demonstrate this, an example hypothesis has been drafted for this purpose: that every chapter in the given corpus will display similar sentence lengths. The sample consisted of 185 first sentences in the novel, in Russian only, which were analyzed for the number of words per sentence. The lack of correlation is obvious in the Graph 4, Graph 5 and Graph 6.



Graph 4. Chapter 1 – Chapter 2 correlation 0.02.



Graph 5. Chapter 2 – Chapter 3 correlation 0.11.

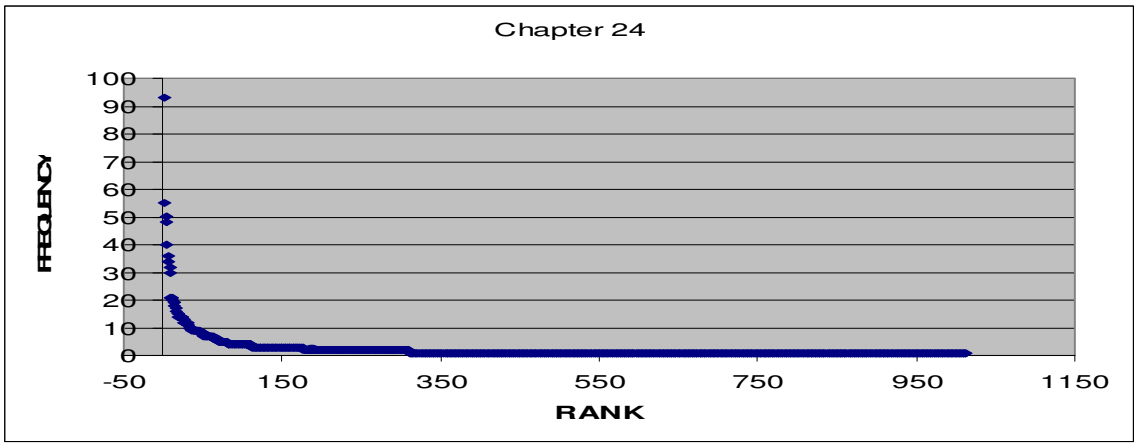


Graph 6. Chapter 1 – Chapter 3 correlation 0.15.

The hypothesis stated in this section seems to be obviously disproven by arriving at such a scatter plot. There is absolutely no link between the given data and as such, the hypothesis must be regarded as false.

4.3 Negative correlation

In cases where the connection between the variables exists, only it is reversely proportional (the larger one score the smaller the connected other) negative correlation occurs. As an example hypothesis here, it was conjectured that the number representing the rank of the word in a chapter is smaller as the number standing for frequency of the word increases. Rank and frequency of words in Chapter 24 in Russian were extracted using Unitex and the Graph 7 represent how the influence each other in a reversely proportional manner.



Graph 7. Chapter 24 rank – frequency correlation -0.89.

The high negative correlation coefficient indicates that the given hypothesis is true. It is actually a well known linguistic and statistical law known as Zipf's Law.

5. CONCLUDING REMARKS

With the increasing accessibility of linguistic corpora and the renewed belief in the linguistic paradigm that theory must be based on language *as it is*, that arose from the emergence and establishment of empirical linguistics once again at the foreground of linguistics, one of the major implication is that linguists will increasingly demand the use of statistics in research. The purpose of this paper was to demonstrate that statistics is not as alien as it might seem and to emphasize that for linguists methodology is only important; it is *what* and *when* rather than the *why*. The paper also presented a small example of the use of most basic statistical methods as applicable in corpus research. There are of course many more procedures to utilize, and it is the matter of the given research and the proficiency of the researcher, which to appropriate. The main thing to remember is that in all empirical language research at least one needs to be used.

6. REFERENCES

Dekking, F.M., Kraaikamp, C., Lopuhaa, H.P., Meester, L.E. (2005). *A Modern Introduction to Probability and Statistics – Understanding Why and How*. London, Springer-Verlag.

Dobrić, N. (2009a). Korpusni pristup kao nova paradigma istraživanja jezika. *Philologia*, No. 6, pp. 31–41.

Dobrić, N. (2009b). Korpusna statistika u lingvistici – Mogucnost primene u izucavanju jezika. *Lecture given at the Statistical Society of Serbia, May 2009,printed content of the lecture. Novi Sad.*

Mann. S. P. (1995). *Statistics for business and economics*. Wiley, pp. 823-826.

Vitas, D, Krstev, C,. Koeva, S. (2007). Towards a Complex Model for Morpho-Syntactic Annotation , in E. Paskaleva and Slavcheva, M. (eds.) *Proceedings of the Workshop Workshop on a Common Natural Language Processing Paradigm for Balkan Languages*, RANLP, Borovets, Bulgaria, pp. 65-71.