

MULTI – WORD EXPRESSION DETECTION FOR TURKISH

**A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED
SCIENCES OF
ÇANKAYA UNIVERSITY**

**BY
NAZLI HÜRMEYDAN**

**IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF
COMPUTER ENGINEERING**

FEBRUARY 2016

Title of the Thesis: **Multi-Word Expression Detection for Turkish.**

Submitted by **Nazlı HÜRMEYDAN**

Approval of the Graduate School of Natural and Applied Sciences, Çankaya University.



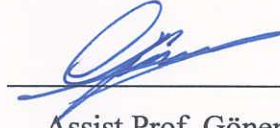
Prof. Dr. Halil Tanyer EYYUBOĞLU
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.



Prof. Dr. Müslim BOZYİĞİT
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.



Assist Prof. Gönenç ERCAN
Supervisor

Examination Date:

Examining Committee Members

Assist Prof. Abdül Kadir GÖRÜR

(Çankaya Uni.)



Assist Prof. Gönenç ERCAN

(Hacettepe Uni.)



Assist Prof. Burcu CAN

(Hacettepe Uni.)



STATEMENT OF NON-PLAGIARISM PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : Nazlı HÜRMEYDAN

Signature

: 

Date

: 19.02.2016

ABSTRACT

MULTI – WORD EXPRESSION DETECTION FOR TURKISH

HÜRMEYDAN, Nazlı

M.Sc., Department of Computer Engineering

Supervisor: Assist Prof. Gönenç ERCAN

February 2016, 38 pages

In this thesis, I performed text analytics on Turkish academic articles about four science subjects and detected collocations according to statistical measures and tried to benchmark the results of each method in these subjects. The main purpose of my thesis is to create a terminology dictionary using only example journal articles. In accordance with this purpose, I applied some machine learning methods on Weka to test if my scores and Weka results can get more reliable results.

Keywords: collocation detection, multiword expression extraction

ÖZ

ÇOK KELİMELİ TÜRKÇE DEYİM BELİRLEME

HÜRMEYDAN, Nazlı

Yüksek Lisans, Bilgisayar Mühendisliği Anabilim Dalı

Tez Yöneticisi: Yrd. Doç. Dr. Gönenç ERCAN

Şubat 2016, 38 sayfa

Bu tezde Türkçe akademik makaleler üzerinden çeşitli metodlarla hesaplama yaparak deyim analizi yapmaya çalıştım. Dört farklı konu başlığı altındaki makalelere uyguladığım metodların sonuçlarını sunarak doğruluk karşılaştırması yaptım. Buradaki temel amacım bir dizi akademik makale verildiğinde klasik istatistiksel analiz yöntemleri ile o araştırma alanı için bir terminoloji sözlüğü çıkarmaktır. Bu amaç doğrultusunda da esas uygulamamın yanı sıra Weka ile makine öğrenmesi yöntemlerini kullandım. Bundaki amacım ise her iki uygulamanın kombinasyonu ile daha iyi sonuç elde edip etmediğimi test etmektir.

Anahtar Kelimeler: deyim belirleme, çok kelimeli söz çıkarma

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Assist Prof. Gönenç ERCAN for his supervision, special guidance, suggestions, and encouragement through the development of this thesis.

It is a pleasure to express my special thanks to my family for their valuable support.

TABLE OF CONTENTS

| | |
|----------------------------------|------|
| STATEMENT OF NON PLAGIARISM..... | i |
| ABSTRACT..... | ii |
| ÖZ..... | iii |
| ACKNOWLEDGEMENTS..... | iv |
| TABLE OF CONTENTS..... | v |
| LIST OF FIGURES..... | vii |
| LIST OF TABLES..... | viii |

CHAPTERS:

| | |
|--|----|
| 1. INTRODUCTION..... | 1 |
| 2. LITERATURE SURVEY | 2 |
| 2.1. Characteristics of Collocations | 3 |
| 2.2. Terminology Extraction Approaches | 4 |
| 2.3. Collocation Descriptions According to Several Researchers | 5 |
| 2.4. Some Past Related Works | 6 |
| 3. METHODOLOGY | 8 |
| 3.1 Frequency | 9 |
| 3.2 Null-Hypothesis | 9 |
| 3.3 T-Test | 10 |
| 3.4 Pearson's Chi-Square | 11 |
| 3.5 Likelihood | 12 |
| 3.6 Mutual Information | 13 |
| 4. APPLICATION AND RESULTS | 14 |
| 4.1. Precision and Recall Values | 18 |
| 4.1.1 Economy | 18 |

| | |
|---|----|
| 4.1.2 Law | 20 |
| 4.1.3 Geography | 22 |
| 4.1.4 Child Development | 24 |
| 4.2 Weka Results | 25 |
| 4.2.1 Economy | 26 |
| 4.2.1.1 Decision Tree on Economy Corpus | 26 |
| 4.2.1.2 Naive Bayes on Economy Corpus | 27 |
| 4.2.2 Law | 28 |
| 4.2.2.1 Decision Tree on Law Corpus | 28 |
| 4.2.2.2 Naive Bayes on Law Corpus | 29 |
| 4.2.3 Geography | 30 |
| 4.2.3.1 Decision Tree on Geography Corpus | 30 |
| 4.2.3.2 Naive Bayes on Geography Corpus | 31 |
| 4.2.4 Child Development..... | 32 |
| 4.2.4.1 Decision Tree on Child Development Corpus | 32 |
| 4.2.4.2 Naive Bayes on Child Development Corpus | 33 |
| 4.3 Summary of Results | 34 |
| 5. CONCLUSION | 36 |

LIST OF FIGURES

| | | |
|-----------------|---|----|
| Figure 1 | Recall Chart on Economy Corpus | 19 |
| Figure 2 | Precision Chart on Economy Corpus | 19 |
| Figure 3 | Recall Chart on Law Corpus | 21 |
| Figure 4 | Precision Chart on Law Corpus | 21 |
| Figure 5 | Recall Chart on Geography Corpus | 22 |
| Figure 6 | Precision Chart on Geography Corpus | 23 |
| Figure 7 | Recall Chart on Child Development Corpus | 24 |
| Figure 8 | Precision Chart on Child Development Corpus | 25 |

LIST OF TABLES

| | | |
|-----------------|---|----|
| Table 1 | Corpus Summary | 9 |
| Table 2 | Example of Bigrams | 14 |
| Table 3 | Example of Bigrams | 15 |
| Table 4 | Example of Terms in Dictionary | 16 |
| Table 5 | Found Collocations on Economy Corpus | 18 |
| Table 6 | Found Collocations on Law Corpus | 20 |
| Table 7 | Found Collocations on Geography Corpus | 22 |
| Table 8 | Found Collocations on Child Development Corpus | 24 |
| Table 9 | Changes in Recall values of Decision Tree on Economy Corpus | 26 |
| Table 10 | Changes in Precision values of Decision Tree on Economy Corpus | 26 |
| Table 11 | Changes in Recall values of Naive Bayes on Economy Corpus | 27 |
| Table 12 | Changes in Precision values of Naive Bayes on Economy Corpus | 27 |
| Table 13 | Changes in Recall values of Decision Tree on Law Corpus | 28 |
| Table 14 | Changes in Precision values of Decision Tree on Law Corpus .. | 28 |
| Table 15 | Changes in Recall values of Naive Bayes Tree on Law Corpus | 29 |
| Table 16 | Changes in Precision values of Naive Bayes Tree on Law Corpus | 29 |
| Table 17 | Changes in Recall values of Decision Tree on Geography Corpus..... | 30 |

| | | |
|-----------------|--|----|
| Table 18 | Changes in Precision values of Decision Tree on Geography Corpus..... | 30 |
| Table 19 | Changes in Recall values of Naive Bayes on Geography Corpus..... | 31 |
| Table 20 | Changes in Precision values of Naive Bayes on Geography Corpus | 31 |
| Table 21 | Changes in Recall values of Decision Tree on Child Development Corpus | 32 |
| Table 22 | Changes in Precision values of Decision Tree on Child Development Corpus | 32 |
| Table 23 | Changes in Recall values of Naive Bayes on Child Development Corpus | 33 |
| Table 24 | Changes in Precision values of Naive Bayes on Child Development Corpus | 33 |
| Table 25 | Best Results of Economy Corpus | 34 |
| Table 26 | Best Results of Law Corpus | 34 |
| Table 27 | Best Results of Geography Corpus | 34 |
| Table 28 | Best Results of Child Development Corpus | 34 |

CHAPTER 1

INTRODUCTION

The importance and need to recognize terms and their relations became more eminent as the current state-of-the-art in Web technologies, Ontology Learning, Machine Translation and especially Artificial Intelligence systems advanced. Progress in information technology requires having more reliable knowledge base systems. Due to this requirement terminology extraction becomes considerably substantial which finds terms salient for a given corpus.

Although this subject has importance in various applications of technological systems, not a lot of works or systems exist in the literature targeting Turkish language and Turkish documents. With this motivation the main purpose of my thesis is create a terminology dictionary using some journal articles by applying classical collocation analysis methods.

This thesis is organized in the following order. In Chapter 2, I will first present a short survey of the literature. Then in Chapter 3, I will present methodologies I used in my research. This problem for Turkish is a challenge as Turkish is an agglutinative language having derivational word structure. Also most idioms cannot be translated literally. In languages similar to Turkish, it is better to find solutions according to the structure of the language but it isn't easy at all. Accordingly, I applied some statistical methods in my thesis like frequency, null hypothesis, t – test, chi – square, likelihood and mutual information. In Chapter 4, results of our experiments in terminology extraction for different research domains is presented. Finally in Chapter 5, I interpreted the results according to the recall and precision values and tried to improve the result using machine learning algorithms.

CHAPTER 2

LITERATURE SURVEY

Natural Language Processing (NLP) is a branch of Artificial Intelligence which emerges from the desire to develop technology for understanding and generating natural language by computers. Since Anthropology science illustrate the relation between developed human intelligence and language, studies about language processing conducted over the Artificial Intelligence have gained importance again thus studies on NLP are increased. Automated terminology extraction problem got its share from this interest as well.

As known multi-word expressions compose an important part of languages, therefore an important body of literature focusses on detecting and extracting expressions or idioms [1][7]. The usage of notions like term, phrase, technical term facilitates the understanding of the problem, but in the context of informational retrieval they are not clear enough. Differences or similarities of these notions must be known as *term* collapses the meaning of both *word* and *phrases* but regarding to pure linguistic view using *collocation* is a proper choice. The definition of collocation is forming meaningful phrases with more than one word that reflects the culture of the language and is understood easily among everyone who knows the language. A more clear one is "it is a *co-occurrence* which encompasses both the observable frequency information and its interpretation as an indicator of statistical association." [1]. To internalize in the best way: "A collocation is a word combination whose semantic and / or syntactic properties cannot be fully predicted from those of its components and, which therefore has to be listed in a lexicon." [1].

It can be classified in two basic ways like positional and relational co-occurrences. Positional one represents the physical distance between words and relational one

refers linguistic or syntactic interpretations of words. It will be mentioned more under the title of terminology extraction approaches.

2.1 Characteristics of Collocations

- Non-compositionality: If a phrase has a totally different meaning than the interpretational sense of its lexical components. For instance, *kick the bucket* (means to die) or *white elephant* phrases. The latter of them means “unnecessary stuff” that both white or elephant terms are not relevant to the meaning of the phrase.
- Non-substitutability: The components of collocations cannot be altered with any other terms or substitutions even if the meaning of substituted word suits. For example *every minute counts* can't be replaced by *hours* or *separate the men from the boys* is a stereotype expression it isn't related with gender.
- Non-modifiability: Some expressions cannot be changed with ease if it is seen there won't be meaning distortions. For instance, *small hours* (means after midnight) cannot be modified to *little hours* or *nine days' wonder* (easily forgotten) can't be modified like *ten days' wonder*.

The resulting of these features of collocations we realize that if word-by-word translations aren't probable, combination is certainly a collocation. In an example, *make a decision* phrase's exact translation in Turkish is *karar vermek*. *karar* refers *decision* whereas *vermek* means *give*. Understanding the notion of collocation in detail is the most significant part of my study because it forms the basis of my thesis. Same with mine the goal of most researcher has been provide the set of collocations from main corpus [1, 2, 3, 5, 6, 7, 8, 16].

2.2 Terminology Extraction Approaches

Basically there are two mainstream approaches identified in most papers [1, 2, 3, 4, 5, 6, 7, 16]. Linguistic approaches and statistical approaches are essential methods. Also there are hybrid approaches which are formed by using both of them has been tried to increase the efficiency each of it [1, 2, 3, 4, 5, 6, 7, 16].

- Linguistic approaches: In this approach term recognition is relied on syntactic properties of terms and linguistic analysis. In a simple way within a linguistic approach main corpus can be parsed via syntactic patterns or linguistic filters, basically pronouns verbs or any other stop words can be eliminated to expose sufficient and useful candidates. Under this subject expressions' variations can be collected as a unique term if there exists synonyms [2, 3, 6, 16].
- Statistical approaches: Statistical approaches try to extract most probable collocations by evaluating ranks of each although statistical measures can lead to a list of phrases that are not linguistically correct – filtered expressions. In every method it is necessary to calculate *frequency* of each single word and pairs because pair frequency is a member of association measures and single words are certainly used in order to evaluate other measures like Dice Factor, Mutual Information, T-Score, Log-Likelihood Ratio, Chi-Square, Null-Hypothesis. In addition to them, Standard Deviation calculation between pairs can indicate whether it is a collocation or not. [1, 2, 3, 4, 5, 6, 7, 16].
- Hybrid approaches: This type is combination or mixture of two basic methods linguistic and statistical approaches. Indeed, application o linguistic method before statistical approach makes the candidate list more reliable and it can be

more meaningful than implementation of solely statistical measures on main primitive and non-eliminated sources [2, 3, 6, 16].

2.3 Collocation Descriptions According to Several Researchers

- Firth describes the collocation that it is a structure which a word forms with other words and it must be addressed separately each meaning and usage of pairs [9].
- Similar to Firth's definition, Halliday qualifies the collocations as syntactic units or groups that a word can establish in conjunction with other words. He indicates the words that is chosen as examples in written and oral expressions, exhibits a relation. For example if we deal with pairs like "tough discussion", "stiffness in the discussion", "hardened discussion", we see the collocational properties in "discussion" and the other words associated with tough or toughness [10].
- Nesselhauf defines the collocation as associations within a certain range of words that provides a specific amount of usage and specifies them usually having some formulaic structure. For him, because of this formulaic nature it is led to difficulties in separating collocations from expressions or idioms [12].
- According to Mitchell's opinion, idioms differ from the other word combinations both formulaic structures and features shown by the mean of them. So Mitchell claims that collocations does not have much restrictions and stereotyping that idioms have [15].
- Cruse describes expressions as " combinations of words that has a single semantic structure", collocations as "constantly combined sequence of lexical elements" and thinks in determining the meaning of a word, examples of this

word's collocations is helpful. Thus Cruse has an opinion that expressions and collocations are totally different from each other[13].

- His work in which he analyzes English collocations, Kjellmer seen that he has a wider approach to collocations. Kjellmer defines the collocations as a number of frequently adjacent elements that are formed in a grammatical rules who accepted expressions as a subgroup of collocations[14].

2.4 Some Past Related Works

- In study [2], term recognition is made on a specific application domain by specifically implemented terminology extraction architecture which consists of many association measures in literature and a new linguistic approach sufficient for subject or structure of this specific domain. The corpus is provided by European Space Agency, relevant to the topic of institution and has about 673000 words. Briefly it is firstly carried out by a modular syntactic parser, eliminated adjective stop-words and then the remaining terms are incurred on statistical step. In this way comparatively “best” list of collocations are recognized.
- The study of Dunning [4] particularly uses the likelihood method which can reveal substantial conclusions even when text corpora is small. It is mentioned and illustrated about comparisons between chi-square and likelihood results intercalarily touching on problems caused by applying normal distribution to find and analyze rare events.
- With similar aim of detecting multiword expressions there is a reference research in Arabic [6]. It proposes three complementary methods that relies on cross lingual correspondence asymmetries, translational-based approach and corpus-based approach. As a result of this study only few candidates are found because of the rich and complex structure of the language.

- A. Dinu, P. Dinu, I. Sorodoc et al. shows that [5] for efficient collocation detection a rank aggregation method is proposed. Application of the research is done by first carrying out some association measures like dice method, z-test, chi-square test, likelihood ratio. Then ranking distance (aggregation) is applied to measure similarity between two lists [5]. With comparing all methods it is recognized that aggregation method is better. After verifying aggregation method's efficiency it is used to calculate the language similarity among English, Italian, French and Spanish.
- The research [3] focused on hybrid method in order to find accurate collocations from English, French, Spanish and Italian corpora with a two-stage process. In first stage, candidates are selected from the base corpora according to syntactic parses. Following this, terms are ranked by the log-likelihood ratio test. It is shown that traditional window method isn't as useful as the hybrid method that consists of both statistical and syntactic analysis of raw text corpora.
- Although there are many studies in different languages like English, Spanish, Chinese, French, Arabic, Russian and Portuguese there is only two corresponding research in Turkish literature on this subject. In one of them, the study includes statistical approaches on both stemmed and surface-formed corpora and it is recognized that chi-square hypothesis test and mutual information method result better accuracy [8]. I used same methods but as different from that study I gathered the corpus from four different domain and did not use stemmed corpora. In second one, two corpora of news text are used which one is large and other one is small. By using syntax based method on large corpora some rules are created and processed on small one. So results are considerably good [16].

CHAPTER3

METHODOLOGY

This thesis tries to detect real multiword expressions in real Turkish academic articles. My aim is extracting collocations as effectively as possible by assessing different statistical collocation methods in Turkish language terminology extraction.

On the first stage I collected raw text corpora from DergiPark which is an electronic platform to publish national academic journals conducted by TÜBİTAK ULAKBİM. I gathered the corpus from four different domains, namely Child Development, Law, Geography, Economic and Administrative Sciences. For each corpus frequency, null hypothesis, t-test, chi-square, likelihood and mutual information scores are calculated for candidate biword phrases.

Before explaining methods that I used in this study, I have to touch upon how did I evaluate and detailed information of the corpus. In the beginning I wrote a Java program and used Lucene Analyzer to count frequencies which is the basic and significant part of these formulations. Apache Lucene is a free open source library for performing full text search. Briefly the main logic is indexing the data and working on it. In the indexing stage I filtered the stopwords to decrease redundant words but I didn't use stemming which reduces words to a root form. The reason of not using stemming is not to eliminate the collocations including words having affixes. For example; "mal geliri" is a collocation in case of stemming this phrase becomes "mal gelir" which will not be considered as accurate. After indexing single words I indexed pairs. These pairs are generated by combining all word pairs. For example, in the original text I have a sentence like "a b c d" then pairs will be "a b", "b c", "c d" will be generated. The corpus has a structure as in Table 1.

| Subject Of Corpus | N(tokens) | number of single unique words | number of unique bigrams |
|--------------------------|-----------|-------------------------------|--------------------------|
| <u>Child Development</u> | 121.493 | 26.209 | 66.596 |
| <u>Law</u> | 324.403 | 43.718 | 159.290 |
| <u>Geography</u> | 198.862 | 31.028 | 103.658 |
| <u>Economics</u> | 198.229 | 40.691 | 113.486 |
| | | | |
| <u>ALL</u> | 842.987 | 141.646 | 443.030 |

Table 1. Corpus Summary

My purpose in here is to score each bigram by using each methods that I mention above and find a terminology of each domain according to these scores. In the following subsection I will be explaining the statistical methods I used in detail.

3.1 Frequency

Frequency is directly produced from the raw corpora, without any statistical methods, thus it is the simplest method for extracting collocations. Simply we count the terms and order them from most frequent bigrams to least. Being simplest method arises a solid difficulty as high frequency can be accidental not related to the importance of the phrase for the research domain, thus there could be many insignificant or irrelevant bigrams.

3.2 Null – Hypothesis

Co-Occurrence by chance is a common problem in statistics therefore we want to know if two words are random occurrences or not. A *null hypothesis* H_0 is formed for modelling randomness stating that this combination occurs just by chance. First we need to identify the null hypothesis of biword phrases where w^1 is the first term and w^2 is the second term of the phrase

$$H_0 : P(w^1w^2) = P(w^1)P(w^2)$$

With this formulation we assume the probability of cooccurrence is independent from each word's occurrence and equal to product of individual probabilities. This independence model is our null hypothesis. Then the probability of occurrence is calculated, depending on this value of probability the null hypothesis is accepted or rejected. If the co-occurrence cannot be explained by chance, it could hint that this phrase is a collocation.

3.3 The T – Test

The t-test looks at the mean and variance of a sample, where the null hypothesis is that the sample is drawn from a distribution with mean μ . The test computes the difference between the observed and expected means, scaled by the variance of the data, and tells us how likely it is to get a sample of that mean and variance (or a more extreme mean and variance) assuming that the sample follows normal distribution.[7]

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

where \bar{x} is the sample mean, s^2 is the sample variance, N is the sample size, μ is the mean of the distribution. To explain in probability form that is used in my calculation part;

$$t = \frac{P(w^1w^2) - P(w^1)P(w^2)}{\sqrt{\frac{P(w^1w^2)}{N}}}$$

If the t is large enough we can reject the null hypothesis' independency and it means word pair forms a collocation.

3.4 Pearson's Chi – Square Test

Chi square test is defined with a similar purpose to t-test. However t-test does not assume a normal distribution, chi-square method simply sums the difference between expected and observed frequencies. A contingency table is formed as given below;

For P(ab);

| | | |
|--------------|-----------|--------------|
| | $W_1 = a$ | $W_1 \neq a$ |
| $W_2 = b$ | f_0 | f_1 |
| $W_2 \neq b$ | f_2 | f_3 |

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where i ranges over rows of the table, j ranges over columns, O_{ij} is the observed value for cell and E_{ij} is the expected value.

Applying on two word evaluation as mine;

$$\chi^2 = \frac{N(f_0f_3 - f_1f_2)^2}{(f_0 + f_1)(f_0+f_2)(f_1 + f_3)(f_2 + f_3)}$$

In formulation, f frequencies are combined according to the existance of pairs in bigrams. In case of calculating the phrase "a b", f_0 is a frequency of bigram that first word is and second word is b. f_1 is a frequency of bigram that first word isn't a but second word is b. f_2 is a frequency of bigram that first word is a but second word isn't b. f_3 is a frequency of bigram that neither first word is a nor second word is b. If

the χ^2 is large enough we can reject the null hypothesis' independency and it means word pair forms a collocation.

3.5 Likelihood Ratios

It is based on ratio between the likelihood of the observed and expected data so the result implies how much likely one hypothesis is than the other. Likelihood ratio test does not depend on a point null hypothesis because it computes the maximal likelihood estimate and checks its consistency with null hypothesis value.

The formula is;

$$\begin{aligned} \log \lambda &= \log \frac{L(H_1)}{L(H_2)} \\ &= \log \frac{b(c_{12}; c_1, p) b(c_2 - c_{12}; N - c_1, p)}{b(c_{12}; c_1, p_1) b(c_2 - c_{12}; N - c_1, p_2)} \\ &= \log L(c_{12}, c_1, p) + \log L(c_2 - c_{12}, N - c_1, p) - \log L(c_{12}, c_1, p_1) - \\ &\quad \log L(c_2 - c_{12}, N - c_1, p_2) \end{aligned}$$

where b is a binomial distribution;

$$b(k; n, x) = \binom{n}{k} x^k (1 - x)^{n-k}$$

$$L(k, n, x) = x^k (1 - x)^{n-k}$$

and

$$p = \frac{c_2}{N} \quad p_1 = \frac{c_{12}}{c_1} \quad p_2 = \frac{c_2 - c_{12}}{N - c_1}$$

c_1 is frequency of w_1

c_2 is frequency of w_2

c_{12} is frequency of $w_1 w_2$

N is frequency of single unique words

3.6 Mutual Information

An information-theoretically motivated measure for discovering interesting collocations is pointwise mutual information and is an estimate for logarithm of the μ -value. By this method it is found out how much information does the entity of one word gives about the other word.

Assuming two words x and y ;

$$I(x, y) = \log \frac{P(xy)}{P(x)P(y)}$$

To interpret results and illustrate that;

$$\begin{aligned} I(x, y) &= \log \frac{P(xy)}{P(x)P(y)} \\ &= \log \frac{P(x)P(y)}{P(x)P(y)} \end{aligned}$$

If the cooccurrence appears by chance and two words are independent from each other we get result as 0. So we try to find collocations in higher values.

CHAPTER4

APPLICATION AND RESULTS

As an application part, I applied all methods on data of each subject. The tables Table 2 and Table 3 shows some examples of extracted phrases for each subject.

| | CHILD DEVELOPMENT | LAW |
|--------------------|--------------------------|----------------------------------|
| FREQUENCY | özel eğitim | söz konusu |
| | okul öncesi | toplu iş |
| | özel gereksinimli | türk borçlar |
| NULL-HYPOTHESIS | çocuk eğitim | iş iş |
| | eğitim çocuk | iş hukuk |
| | eğitim çocukların | sayılı iş |
| T-TEST | özel eğitim | söz konusu |
| | okul öncesi | toplu iş |
| | özel gereksinimli | türk borçlar |
| CHI-SQUARE | cellular subscriptions | uğur yağcı |
| | elektrik çarpması | ggüüvveennlliikk kkoonnssseeyyii |
| | telephone lines | hye jeong |
| LIKELIHOOD | children with | netice sebebiyle |
| | primary school | sözleşmesi yapma |
| | yürüttüğü down | berner kommentar |
| MUTUAL INFORMATION | a.n gutierrez | aannddlla mmaallaarraa |
| | aba eya | aavvruupaa bbiirllii |
| | abd'deki avustralya'daki | abdel nour |

Table 2. Example of Bigrams 1

| | GEOGRAPHY | ECONOMICS |
|--------------------|--------------------------|------------------|
| FREQUENCY | coğrafya dergisi | sosyal bilimler |
| | yer alan | kayıt dışı |
| | sosyal bilgiler | yeni ekonomi |
| NULL-HYPOTHESIS | coğrafya coğrafya | ekonomi ekonomi |
| | arasında coğrafya | ekonomi ekonomik |
| | doğal coğrafya | ekonomi sosyal |
| T-TEST | coğrafya dergisi | sosyal bilimler |
| | yer alan | kayıt dışı |
| | sosyal bilgiler | yeni ekonomi |
| CHI-SQUARE | ayşe çağliyan | vrđ tmt |
| | çamlık mağaraları | asuman oktayer |
| | doğrultu atımlı | yaln zca |
| LIKELIHOOD | mezunu öğretmenlerin | test sonuçları |
| | yayınları no | televizyon yay |
| | erinç kuraklık | döneme ait |
| MUTUAL INFORMATION | a.coğrafya öğretmeninden | a.lđıa'ı sö |
| | abrasion platform | a.o balkanlı |
| | acaglayan fırat.edu.tr | a.vası tasız |

Table 3. Example of Bigrams

Table 2 and Table 3 have results on each domain and each method. Chi-square, likelihood and mutual information methods look partially bad and I think that is because of the methods' "list rare ones first" structure.

After taking results I took them and compare with the dictionary that I have from internet site of TÜBA(Turkish Academy of Science)[17]. TÜBA started "Science Glossary of Turkish Terms" project with support of State Planning Organization of Republic of Turkey on 2002. The purpose is to use and develop Turkish in education, communication and scientific works. Experts form the list of subdomains and after the unity of ensurement is provided these subdomains are collected and glossary is generated. Some examples in this dictionary that I used is as in Table 4.

| Economy | Law | Geography | Child Development |
|-------------------|------------------|--------------------|--------------------|
| cari kur | davanın düşmesi | abrazyon platformu | aile planlaması |
| ticaret kredisi | adli sicil | akarsu havzası | dil gelişimi |
| dolaylı vergi | idari karar | yaz günü | yaratıcı bellek |
| nominal değer | nedensellik bağı | bitki örtüsü | sembolik oyun |
| mal geliri | zimmet suçu | kapalı havza | üstün zeka |
| enflasyon sarmalı | başkanlık divanı | tropikal iklim | zihin körlüğü |
| dışsal etkenler | limitet sirket | richter ölçeği | gelişim aksamaları |
| otonom yatırım | oturma hakkı | salt nem | işitme sınırı |
| arızı işsizlik | gaiplik kararı | toprak kayması | mekanik zeka |
| iş değerlemesi | borç erteleme | hava basıncı | rochester yöntemi |

Table 4. Example of Terms in Dictionary

In the first step for evaluation of terms I created tables that contain bigram numbers, precision and recall values. First tables show how many bigrams are founded in first 50, first 100, the list's first middle and the last parts among the run list. Other two graphics demonstrate the precision recall values according to each method that I applied.

Secondly, I run all of my list on Weka and applied some machine learning methods to test if my scores and Weka results can get more reliable results. To apply Weka step, first I created "arff" files to run this tool. I combined the results of all methods on each domain under specific attributes like this;

```
@relation DOMAIN_NAME
  @attribute frequency numeric
  @attribute null numeric
  @attribute ttest numeric
  @attribute chi numeric
  @attribute like numeric
  @attribute mutual numeric
  @attribute class {var,yok}
```

@data

262,0.000008,16.125187,69135.552724,-0.758994,8.04637,yok
123,0.000005,11.03733,25540.477844,-224.479803,7.703501,yok

Then I used SpreadSubsample class that is a sampling tool and dataset can be spreaded specifically with this filter. I used SpreadSubsample on Weka to spread my each list in a random and balanced manner according to existing or non-existing property of terms(according to attribute *class*) and I got train list with this. Because it is important of have a balanced distribution of classes “var”and “yok” to make reliable training on the dataset. Then rest of the list became test part on which I will run J48tree (decision tree) and Naive Bayes methods.

J48 is a Java implementation in Weka of the algorithm which is known as C4.5 and also ID3. Basic concept of these algorithms are difference in entropy (gain). Formulation is[18];

$$Gain(A) = Info(D) - Info_A(D)$$
$$Info(D) = - \sum_{i=1}^m p_i \log_2(p_i) \quad Info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} x Info(D_j)$$

Attributes with highest gain are choosen for splitting the data while generating subsets of the tree and the tree itself. So J48 is a method which is a tree created using a rule based machine learning model. According to train set a decision tree is generated and in case of new item classification this tree is used.

Naive Bayes classifier predicts an item according to data of train set individually and evaluate the probability of inclusion to the class by simple Bayesian method. The formulation is as below [18];

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

$$P(X|C_i) = \prod_{k=1}^n P(X_k|C_i)$$

Last step is gathering all the results and comparing them with my results that I got from my program.

4.1 PRECISION AND RECALL VALUES

After giving the tables I want to mention that efficiency of method is understood from the location of terms. It is expected from a good method that most of the collocations are found at first 50 but at least in all tables I expect to find collocations in the first parts of the list.

4.1.1 ECONOMY

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|--------|--------------------|------------|
| first 50 | 2 | 4 | 2 | 4 | 0 | 0 |
| first 100 | 7 | 9 | 3 | 9 | 0 | 1 |
| FIRST | 124 | 160 | 185 | 149 | 19 | 49 |
| MID | 151 | 206 | 213 | 160 | 82 | 118 |
| LAST | 230 | 230 | 230 | 230 | 230 | 230 |

Table 5. Found Collocations on Economy Corpus

Table 5 shows the amount of found collocations on Economy corpus by each method. On Economy corpus it can be realized that 230 collocations are found as distribution on Table 5 and recall and precision values are evaluated from these values.

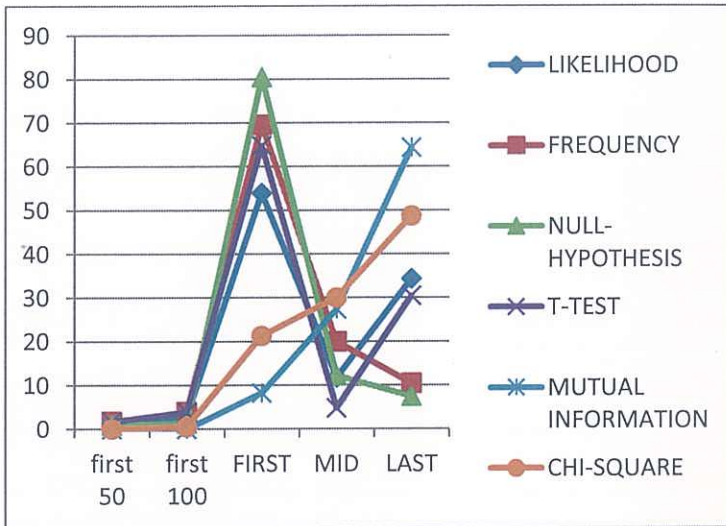


Figure 1. Recall chart on Economy Corpus

In Figure 1, although first 50 and 100 class has less ground truth terms, frequency and t-test are the best methods. According to recall chart on economy most effective method is null-hypothesis on first part of list, chi-square on middle part of list and mutual information on last part of list.

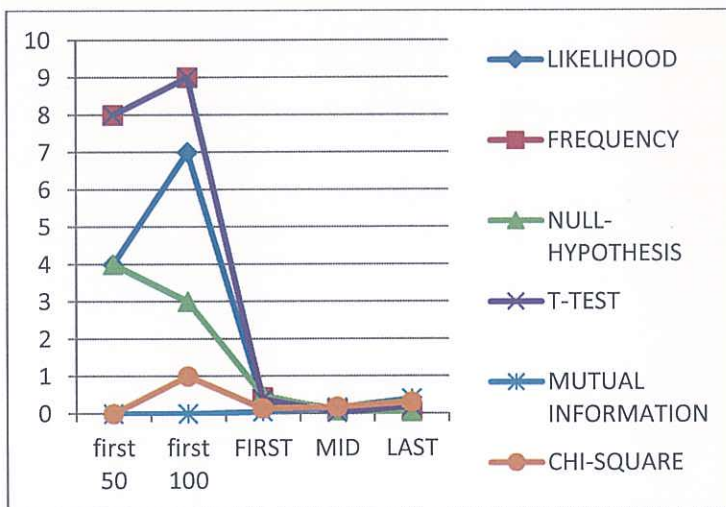


Figure 2. Precision chart on Economy Corpus

In Figure 2, precision values due to the too small ratio (denominator is high) first, mid and last categories performed poorly. According to precision chart on economy most effective methods are frequency and t-test on first 50 and 100 part of list, second good method is likelihood.

4.1.2 LAW

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|--------|--------------------|------------|
| first 50 | 2 | 10 | 2 | 10 | 0 | 0 |
| first 100 | 2 | 12 | 3 | 12 | 0 | 0 |
| FIRST | 101 | 132 | 112 | 126 | 27 | 59 |
| MID | 142 | 153 | 162 | 146 | 93 | 126 |
| LAST | 176 | 176 | 176 | 176 | 176 | 176 |

Table 6. Found Collocations on Law Corpus

Table 6 shows the amount of found collocations on Law corpus by each method. On Law corpus it can be realized that 176 collocations are found as distribution on Table 6 and recall and precision values are evaluated from these values.

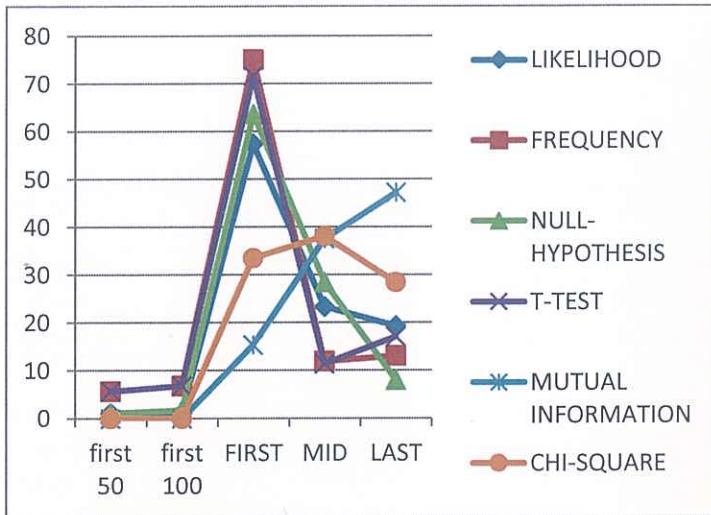


Figure 3. Recall chart on Law Corpus

In Figure 3, although first 50 and 100 class has less bigrams, frequency and t-test are the best methods. According to recall chart on law most effective method is frequency on first part of list, chi-square on middle part of list and mutual information on last part of list.

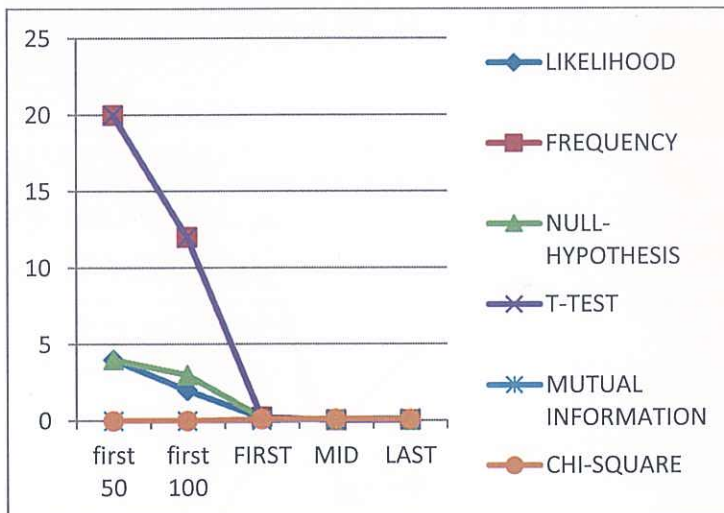


Figure 4. Precision chart on Law Corpus

In Figure 4, precision values due to the too small ratio (denominator is high) first, mid and last categories performed poorly. According to precision chart on law most

effective methods are frequency and t-test on first 50 and 100 part of list, second good method is null-hypothesis.

4.1.3 GEOGRAPHY

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|--------|--------------------|------------|
| first 50 | 0 | 2 | 4 | 2 | 0 | 0 |
| first 100 | 1 | 3 | 4 | 3 | 0 | 0 |
| FIRST | 33 | 43 | 43 | 40 | 7 | 16 |
| MID | 41 | 52 | 57 | 46 | 25 | 34 |
| LAST | 62 | 62 | 62 | 62 | 62 | 62 |

Table 7. Found Collocations on Geography Corpus

Table 7 shows the amount of found collocations on Geography corpus by each method. On Geography corpus it can be realized that 62 collocations are found as distribution on Table 7 and recall and precision values are evaluated from these values.

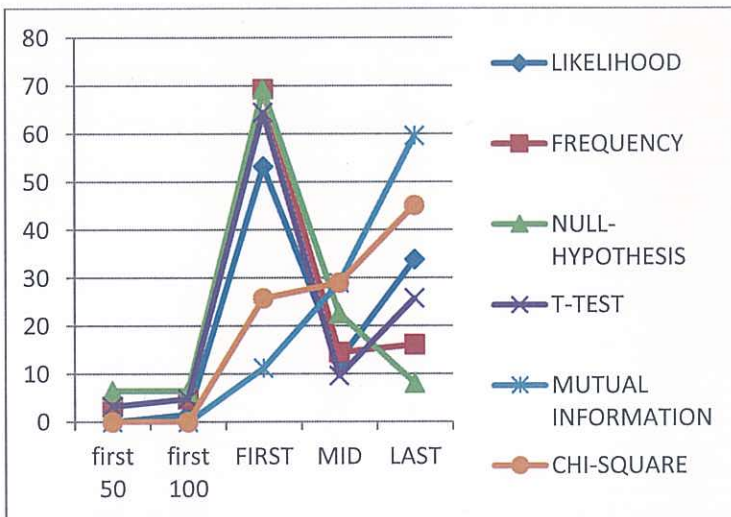


Figure 5. Recall chart on Geography Corpus

In Figure 5, although first 50 and 100 class has less bigrams, null-hypothesis is best method. According to recall chart on geography most effective methods are frequency and null-hypothesis on first part of list, chi-square on middle part of list and mutual information on last part of list.

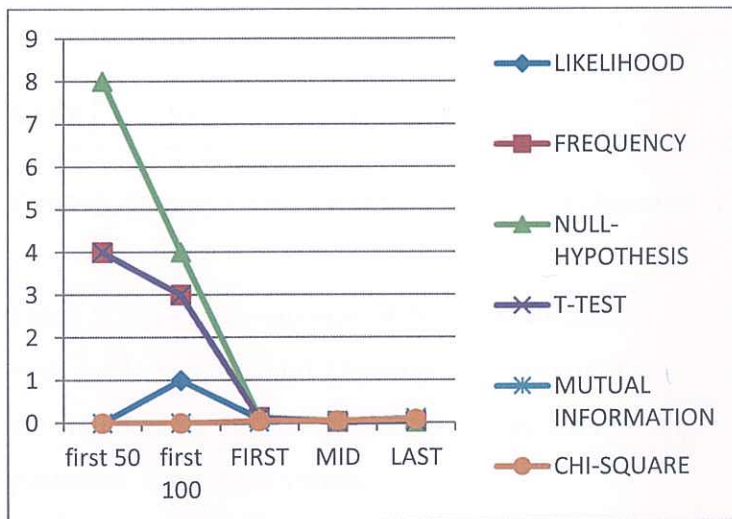


Figure 6. Precision chart on Geography Corpus

In Figure 6, precision values due to the too small ratio (denominator is high) first mid and last categories are not good results. According to precision chart on geography most effective method is null-hypothesis on first 50 and 100 part of list, second good methods are frequency and t-test.

4.1.4 CHILD DEVELOPMENT

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|--------|--------------------|------------|
| first 50 | 1 | 5 | 2 | 5 | 0 | 0 |
| first 100 | 3 | 11 | 4 | 11 | 0 | 0 |
| FIRST | 51 | 66 | 70 | 59 | 5 | 12 |
| MID | 61 | 79 | 88 | 67 | 25 | 49 |
| LAST | 91 | 91 | 91 | 91 | 91 | 91 |

Table 8. Found Collocations on Child Development Corpus

Table 8 shows the amount of found collocations on Child Development corpus by each method. On Child Development corpus it can be realized that 91 collocations are found as in the distribution given on Table 8 and recall and precision values are evaluated from these values.

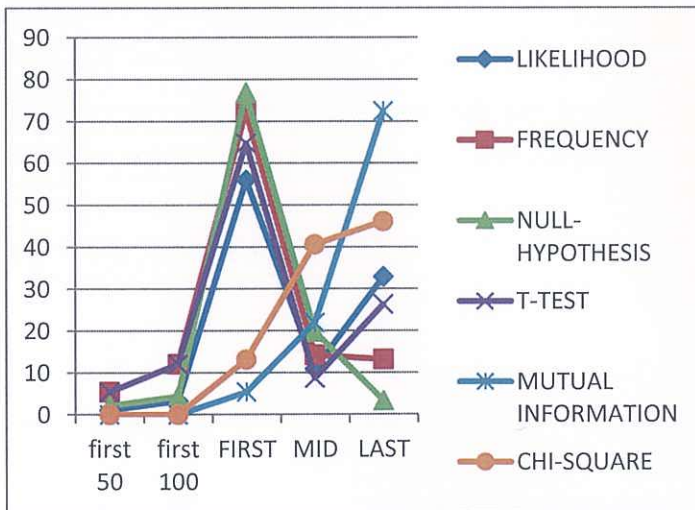


Figure 7. Recall chart on Child Development Corpus

In Figure 7, although first 50 and 100 class has less bigrams, frequency and t-test are best methods. According to recall chart on child development most effective method is null-hypothesis on first part of list, chi-square on middle part of list and mutual information on last part of list.

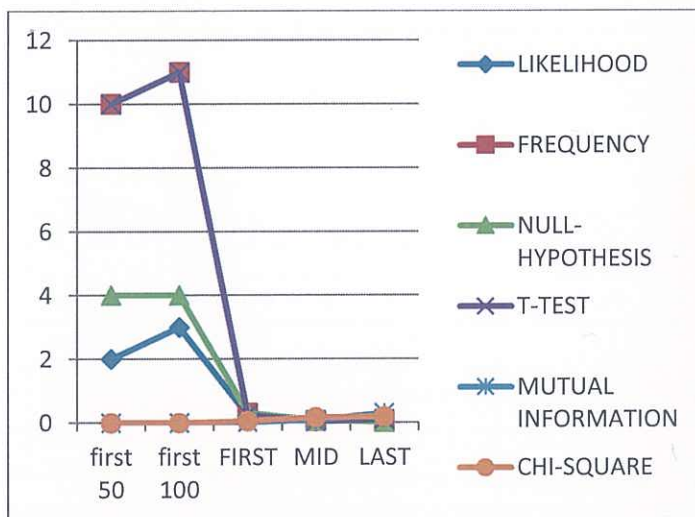


Figure 8. Precision chart on Child Development Corpus

In Figure 8, precision values due to the too small ratio (denominator is high) first, mid and last categories haven't good results. According to precision chart on child development most effective methods are frequency and t-test on first 50 and 100 part of list, second good method is null-hypothesis.

4.2 WEKA RESULTS

Tables below demonstrate the value changes between original precision, recall values of main part that I got after the application of methods and precision, recall values of list that Weka generated. After getting each list that occurs by descending order of bigram probabilities of Weka results, I found precision and recall values of each list. Then found the difference between these result and original values in percent to consider and compare the accuracy of my results.

4.2.1 ECONOMY

4.2.1.1 Decision Tree on Economy Corpus

First of all as seen in Table 9 and Table 10, as mutual and chi-square recall values are lower, Weka results improved these results. Decision tree results have better accuracy on first 100 and first part of list at all methods. All precision values are worse on Weka results.

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|---------|--------------------|------------|
| first 50 | -0,100 | -0,970 | -0,100 | -0,970 | 0,769 | 0,769 |
| first 100 | 1,572 | 0,702 | 3,311 | 0,702 | 4,615 | 4,181 |
| FIRST | 32,241 | 16,589 | 5,719 | 21,371 | 77,893 | 64,849 |
| MID | -3,278 | -11,538 | -3,712 | 3,679 | -18,930 | -21,538 |
| LAST | -28,963 | -5,050 | -2,007 | -25,050 | -58,963 | -43,311 |

Table 9. Changes in Recall values of Decision Tree on Economy Corpus (%)

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|--------|--------------------|------------|
| first 50 | -2,000 | -6,000 | -2,000 | -6,000 | 2,000 | 2,000 |
| first 100 | -1,000 | -3,000 | 3,000 | -3,000 | 6,000 | 5,000 |
| FIRST | -0,031 | -0,126 | -0,192 | -0,097 | 0,246 | 0,167 |
| MID | -0,040 | -0,089 | -0,043 | 0,001 | -0,133 | -0,149 |
| LAST | -0,196 | -0,047 | -0,028 | -0,172 | -0,383 | -0,286 |

Table 10. Changes in Precision values of Decision Tree on Economy Corpus (%)

4.2.1.2 Naive Bayes on Economy Corpus

In a difference with Decision Tree, according to Table 11 and Table 12 Naive Bayes on Economy results are less more accuracy on first part of list at all methods. In depending precision values are a bit better than Decision Tree but still all precision values are worse on weka application.

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|---------|--------------------|------------|
| first 50 | -0,100 | -0,970 | -0,100 | -0,970 | 0,769 | 0,769 |
| first 100 | 1,572 | 0,702 | 3,311 | 0,702 | 4,615 | 4,181 |
| FIRST | 29,164 | 13,512 | 2,642 | 18,294 | 74,816 | 61,773 |
| MID | -0,970 | -9,231 | -1,405 | 5,987 | -16,622 | -19,231 |
| LAST | -28,194 | -4,281 | -1,237 | -24,281 | -58,194 | -42,542 |

Table 11. Changes in Recall values of Naive Bayes on Economy Corpus (%)

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|--------|--------------------|------------|
| first 50 | -2,000 | -6,000 | -2,000 | -6,000 | 2,000 | 2,000 |
| first 100 | -1,000 | -3,000 | 3,000 | -3,000 | 6,000 | 5,000 |
| FIRST | -0,042 | -0,137 | -0,203 | -0,108 | 0,236 | 0,156 |
| MID | -0,032 | -0,081 | -0,035 | 0,009 | -0,125 | -0,141 |
| LAST | -0,193 | -0,044 | -0,025 | -0,169 | -0,381 | -0,283 |

Table 12. Changes in Precision values of Naive Bayes on Economy Corpus (%)

4.2.2 LAW

4.2.2.1 Decision Tree on Law Corpus

According to Table 13 and Table 14, Weka first 50, 100 and first part of list results are better than my recall results even on frequency method that is best method on Law corpus.

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|--------|--------------------|------------|
| first 50 | 5,114 | 0,568 | 5,114 | 0,568 | 6,250 | 6,250 |
| first 100 | 6,155 | 0,473 | 5,587 | 0,473 | 7,292 | 7,292 |
| FIRST | 19,697 | 2,083 | 13,447 | 5,492 | 61,742 | 43,561 |
| MID | -10,795 | 0,568 | -15,909 | 1,136 | -25,000 | -25,568 |
| LAST | -8,902 | -2,652 | 2,462 | -6,629 | -36,742 | -17,992 |

Table 13. Changes in Recall values of Decision Tree on Law Corpus (%)

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|--------|--------------------|------------|
| first 50 | 8,000 | -8,000 | 8,000 | -8,000 | 12,000 | 12,000 |
| first 100 | 5,000 | -5,000 | 4,000 | -5,000 | 7,000 | 7,000 |
| FIRST | -0,051 | -0,109 | -0,071 | -0,098 | 0,089 | 0,028 |
| MID | -0,055 | -0,017 | -0,072 | -0,015 | -0,102 | -0,104 |
| LAST | -0,045 | -0,024 | -0,008 | -0,038 | -0,137 | -0,075 |

Table 14. Changes in Precision values of Decision Tree on Law Corpus (%)

4.2.2.2 Naive Bayes on Law Corpus

As shown in Table 15 and Table 16, difference between Decision Tree results first part of list results are little more than Naive Bayes results so my original values are little worse than Weka order.

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|--------|--------------------|------------|
| first 50 | 5,114 | 0,568 | 5,114 | 0,568 | 6,250 | 6,250 |
| first 100 | 6,155 | 0,473 | 5,587 | 0,473 | 7,292 | 7,292 |
| FIRST | 20,739 | 3,125 | 14,489 | 6,534 | 62,784 | 44,602 |
| MID | -11,837 | -0,473 | -16,951 | 0,095 | -26,042 | -26,610 |
| LAST | -8,902 | -2,652 | 2,462 | -6,629 | -36,742 | -17,992 |

Table 15. Changes in Recall values of Naive Bayes on Law Corpus (%)

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|--------|--------------------|------------|
| first 50 | 8,000 | -8,000 | 8,000 | -8,000 | 12,000 | 12,000 |
| first 100 | 5,000 | -5,000 | 4,000 | -5,000 | 7,000 | 7,000 |
| FIRST | -0,049 | -0,107 | -0,070 | -0,096 | 0,091 | 0,030 |
| MID | -0,056 | -0,019 | -0,073 | -0,017 | -0,104 | -0,105 |
| LAST | -0,045 | -0,024 | -0,008 | -0,038 | -0,137 | -0,075 |

Table 16. Changes in Precision values of Naive Bayes on Law Corpus (%)

4.2.3 GEOGRAPHY

4.2.3.1 Decision Tree on Geography Corpus

According to recall Table 17, Decision Tree order isn't better than the original results but on first part of list it is increased minimum 8 percent. Precision values are still worse than original ones as considered in Table 18.

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|---------|--------------------|------------|
| first 50 | 0,000 | -3,226 | -6,452 | -3,226 | 0,000 | 0,000 |
| first 100 | -1,613 | -4,839 | -6,452 | -4,839 | 0,000 | 0,000 |
| FIRST | 24,899 | 8,770 | 8,770 | 13,609 | 66,835 | 52,319 |
| MID | -6,653 | -8,266 | -16,331 | -3,427 | -22,782 | -22,782 |
| LAST | -18,246 | -0,504 | 7,560 | -10,181 | -44,052 | -29,536 |

Table 17. Changes in Recall values of Decision Tree on Geography Corpus (%)

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|--------|--------------------|------------|
| first 50 | 0,000 | -4,000 | -8,000 | -4,000 | 0,000 | 0,000 |
| first 100 | -1,000 | -3,000 | -4,000 | -3,000 | 0,000 | 0,000 |
| FIRST | -0,023 | -0,052 | -0,052 | -0,043 | 0,052 | 0,026 |
| MID | -0,017 | -0,020 | -0,035 | -0,012 | -0,046 | -0,046 |
| LAST | -0,046 | -0,014 | 0,000 | -0,032 | -0,093 | -0,067 |

Table 18. Changes in Precision values of Decision Tree on Geography Corpus (%)

4.2.3.2 Naive Bayes on Geography Corpus

According to Table 19 and Table 20, On Naive Bayes it is still worse than original recall values but it is better than Decision Tree order at all first 50, 100, first and second part of list.

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|---------|--------------------|------------|
| first 50 | 3,125 | -0,101 | -3,327 | -0,101 | 3,125 | 3,125 |
| first 100 | 1,512 | -1,714 | -3,327 | -1,714 | 3,125 | 3,125 |
| FIRST | 28,024 | 11,895 | 11,895 | 16,734 | 69,960 | 55,444 |
| MID | -0,403 | -2,016 | -10,081 | 2,823 | -16,532 | -16,532 |
| LAST | -27,621 | -9,879 | -1,815 | -19,556 | -53,427 | -38,911 |

Table 19. Changes in Recall values of Naive Bayes on Geography Corpus (%)

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|--------|--------------------|------------|
| first 50 | 2,000 | -2,000 | -6,000 | -2,000 | 2,000 | 2,000 |
| first 100 | 0,000 | -2,000 | -3,000 | -2,000 | 1,000 | 1,000 |
| FIRST | -0,020 | -0,049 | -0,049 | -0,040 | 0,055 | 0,029 |
| MID | -0,012 | -0,014 | -0,029 | -0,006 | -0,041 | -0,041 |
| LAST | -0,055 | -0,023 | -0,009 | -0,041 | -0,101 | -0,075 |

Table 20. Changes in Precision values of Naive Bayes on Geography Corpus (%)

4.2.4 CHILD DEVELOPMENT

4.2.4.1 Decision Tree on Child Development Corpus

In Table 21, recall values are increased only on methods are partially bad according to original recall values except null-hypothesis at first 100 list. It is increased 1,48% unit according to previous value. Weka precision values are worse at good methods as seen in Table 22.

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|---------|--------------------|------------|
| first 50 | 0,862 | -3,534 | -0,237 | -3,534 | 1,961 | 1,961 |
| first 100 | 2,586 | -6,206 | 1,487 | -6,206 | 5,882 | 5,882 |
| FIRST | 14,544 | -1,939 | -6,335 | 5,753 | 65,094 | 57,401 |
| MID | 4,697 | 1,401 | -4,094 | 6,895 | -6,292 | -24,973 |
| LAST | -19,242 | 0,539 | 10,429 | -12,648 | -58,802 | -32,428 |

Table 21. Changes in Recall values of Decision Tree on Child Development Corpus (%)

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|--------|--------------------|------------|
| first 50 | 0,000 | -8,000 | -2,000 | -8,000 | 2,000 | 2,000 |
| first 100 | 0,000 | -8,000 | -1,000 | -8,000 | 3,000 | 3,000 |
| FIRST | -0,067 | -0,135 | -0,153 | -0,103 | 0,140 | 0,108 |
| MID | -0,009 | -0,022 | -0,045 | 0,000 | -0,054 | -0,131 |
| LAST | -0,104 | -0,022 | 0,018 | -0,077 | -0,266 | -0,158 |

Table 22. Changes in Precision values of Decision Tree on Child Development Corpus (%)

4.2.4.2 Naive Bayes on Child Development Corpus

According to Table 23 and Table 24, only difference between Decision Tree method and Naive Bayes is there is a little more accuracy on second part of list at Naive Bayes recall values but it doesn't make any sense because the purpose is to find collocations at first 50, 100 or at least first part of list.

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|---------|--------------------|------------|
| first 50 | 0,862 | -3,534 | -0,237 | -3,534 | 1,961 | 1,961 |
| first 100 | 2,586 | -6,206 | 1,487 | -6,206 | 5,882 | 5,882 |
| FIRST | 14,544 | -1,939 | -6,335 | 5,753 | 65,094 | 57,401 |
| MID | 6,658 | 3,361 | -2,133 | 8,856 | -4,331 | -23,012 |
| LAST | -21,202 | -1,422 | 8,468 | -14,609 | -60,763 | -34,389 |

Table 23. Changes in Recall values of Naive Bayes on Child Development Corpus (%)

| | LIKELIHOOD | FREQUENCY | NULL-HYPOTHESIS | T-TEST | MUTUAL INFORMATION | CHI-SQUARE |
|-----------|------------|-----------|-----------------|--------|--------------------|------------|
| first 50 | 0,000 | -8,000 | -2,000 | -8,000 | 2,000 | 2,000 |
| first 100 | 0,000 | -8,000 | -1,000 | -8,000 | 3,000 | 3,000 |
| FIRST | -0,067 | -0,135 | -0,153 | -0,103 | 0,140 | 0,108 |
| MID | -0,004 | -0,018 | -0,040 | 0,005 | -0,050 | -0,126 |
| LAST | -0,108 | -0,027 | 0,014 | -0,081 | -0,270 | -0,162 |

Table 24. Changes in Precision values of Naive Bayes on Child Development Corpus (%)

4.3 Summary of Results

At the end of these application basically result of my program is as shown below;

| | ECONOMY | | METHOD |
|-----------|-------------|--------------|-------------------|
| | RECALL(%) | PRECISION(%) | |
| first 50 | 1,739130435 | 8 | FREQUENCY, T-TEST |
| first 100 | 3,913043478 | 9 | FREQUENCY, T-TEST |
| FIRST | 80,43478261 | 0,489042798 | NULL |

Table 25. Best results of Economy Corpus

| | LAW | | METHOD |
|-----------|-------------|--------------|-------------------|
| | RECALL(%) | PRECISION(%) | |
| first 50 | 5,681818182 | 20 | FREQUENCY, T-TEST |
| first 100 | 6,818181818 | 12 | FREQUENCY, T-TEST |
| FIRST | 75 | 0,248601616 | FREQUENCY |

Table 26. Best results of Law Corpus

| | GEOGRAPHY | | METHOD |
|-----------|-------------|--------------|----------------|
| | RECALL(%) | PRECISION(%) | |
| first 50 | 6,451612903 | 8 | NULL |
| first 100 | 6,451612903 | 4 | NULL |
| FIRST | 69,35483871 | 0,124446502 | NULL,FREQUENCY |

Table 27. Best results of Geography Corpus

| | CHILD DEVELOPMENT | | METHOD |
|-----------|-------------------|--------------|------------------|
| | RECALL(%) | PRECISION(%) | |
| first 50 | 5,494505495 | 10 | FREQUENCY,T-TEST |
| first 100 | 12,08791209 | 11 | FREQUENCY,T-TEST |
| FIRST | 76,92307692 | 0,315329519 | NULL |

Table 28. Best results of Child Development Corpus

According to Table 25, Table 26, Table 27 and Table 28 it is well understood that independently frequency, t-test and null-hypothesis gave the best results. Precision and recall values are resulted unbalancedly because they are calculated differently. Precision is like “ how many true bigrams are found in first 50 bigrams” so the percentage is calculated as dividing true bigram amount by 50. Due to this situation precision is more meaningful at first 50 and 100 section. Recall is like “ how many of true bigrams are found ” so the recall percentage is calculated as dividing true bigram amount by total true bigram amount. This situation makes first 50 and 100 give less meaningful results because results are little amount at first 50 and 100 part. Considering these properties, according to precision results Law domain has best results at first 50 and first 100 part and according to recall results Economy has best results at first part of list.

CHAPTER 5

CONCLUSION

In a brief explanation my aim was to find bigrams in real existing Turkish academic articles by using frequency, null hypothesis, t-test, chi-square, likelihood and mutual information methods in this work. Although there is lots of similar works on other languages, in Turkish it is less studied and I tried to demonstrate accuracy of Turkish corpora by using these statistical methods which are well known in literature.

First of all I want to show a summary that explains briefly the results of the program that I generated. For each subject there is order as *best to worst* according to recall and precision values below:

For Economy

Recall: null hypothesis, frequency, t-test, likelihood, chi-square, mutual information

Precision: frequency, t-test, likelihood, null-hypothesis, chi-square, mutual information

For Law

Recall: frequency, t-test, null hypothesis, likelihood, chi-square, mutual information

Precision: frequency, t-test, null-hypothesis, likelihood, chi-square, mutual information

For Geography

Recall: null hypothesis, frequency, t-test, likelihood, chi-square, mutual information

Precision: null hypothesis, frequency, t-test, likelihood, chi-square, mutual information

For Child Development

Recall: frequency, t-test, null hypothesis, likelihood, chi-square, mutual information

Precision: frequency, t-test, null hypothesis, likelihood, chi-square, mutual information

One problem that I saw is frequency, t-test, null hypothesis gave best results and likelihood method has a moderate power but chi-square and mutual information have bad results. The reason must be formulation because while calculating the probabilities of bigrams due to the structure of these methods it is given the best values to less occurring biwords, therefore list has true bigrams at the end recall values became good at middle and last parts on these methods.

Other problem is even best methods like frequency has very low accuracy at first 50 and 100 section. I think the reason is source data contain few words and the verification source contains few bigrams. At the same time because of the same reason, I got better results in some specific domain. For example, Law domain has most number of tokens and bigrams and it induce to have best precision values on first 50 and first 100 part of list. On the other hand, having less words in dictionary database of Law caused less recall value on first part of list with respect to Economy. To have same amount of tokens doesn't mean it should be gotten similar results like between Economy and Geography domain. Two of them have same amount of tokens but amount of unique bigrams of Economy is more. This shows that frequencies of each bigram of Geography is higher. If we look at the results, this issue provides the more accuracy at first 50 and first 100 recall values of Geography. It means not only the amount of bigrams are important, efficient sources make the accuracy higher. So If I had more efficient articles and comparing dictionary with more expressions were available, results could have been better.

Weka results was worse than it is expected. According to differences between Weka results and my best resulted methods. In Economy corpus Decision Tree analysis increased recall values about %5,72 null hypothesis at first part, %0,7 frequency and t-test at first 100 results. In Law corpus Naive Bayes analysis increased recall values about %0,56 frequency and t-test at first 50 results, %0,47

frequency and t-test at first 100 results, %3 frequency at first part results. In Geography corpus Naive Bayes analysis increased recall values about % 11 frequency and null hypothesis at first part results.

Although there is some increase in recall values on my best methods on Weka results, recall and precision values couldn't be more higher than the original values even in some methods it had worse than original ones. I think there is a reason that effect slightly this situation. I separate the whole bigram list for having a train file. So rest of the list became test file and it was lost some non-existing and existing bigrams. Due to the less bigram amount that is already existing on original list, this decreasing may have had an impact. On the other hand, probabilities of bigrams are calculated highly close even the same so ordering couldn't be susceptible. This was also effect the results.

In some related works, I came across similar difficulties. Although integrating morphological analysis to the extraction process, making 100 percent extraction is not possible because of reasons like there are quite number of foregin multi-words that don't exist in dictionary database [16] which I struggled too. The work that contains both linguistic and statistical methods, it is noticed that frequency is the best method and after linguistic step precision values reach 47 percent without any statistical stage[2].

In addition, whatever the language is, more empirical data needs to be collected in order to improve our understanding of cooccurrence data, statistical association and its relation to collocativity[1] and in order to get useful results it is considered that there must better be applied grammatical parsing or any lexical preprocesses before applying statistical methods.

APPENDIX

By run the program that I generated , I got the results as following tables . They include best 20 bigrams and the calculated values according to relative method and corpus subject. At the end of this application I tried to apply all these methods on the data which formed with the combination of the others.

A1.CHILD DEVELOPMENT

A1.1 FREQUENCY

| BIGRAM | VALUE |
|---------------------|-------|
| özel eğitim | 275 |
| okul öncesi | 262 |
| özel gereksinimli | 123 |
| eğitim fakültesi | 121 |
| öncesi eğitim | 119 |
| üniversitesi eğitim | 117 |
| normal gelişim | 112 |
| gelişim gösteren | 109 |
| fakültesi dergisi | 104 |
| child development | 99 |
| anne baba | 95 |
| eğitim dergisi | 95 |
| erken çocukluk | 90 |
| bilişsel gelişim | 84 |
| eğitim bilimleri | 82 |
| ankara üniversitesi | 79 |
| yüksek lisans | 77 |
| down sendromlu | 72 |
| ilk yardım | 68 |
| lisans tezi | 60 |

A1.2 NULL - HYPOTHESIS

| BIGRAM | VALUE |
|-------------------|--------------|
| çocuk eğitim | 7,524812E-05 |
| eğitim çocuk | 7,524812E-05 |
| eğitim çocukların | 6,447435E-05 |
| çocukların eğitim | 6,447434E-05 |
| çocuk çocuk | 5,399023E-05 |
| çocuğun eğitim | 4,923955E-05 |
| eğitim çocuğun | 4,923955E-05 |
| özel eğitim | 4,873452E-05 |
| çocuk çocukların | 4,626009E-05 |
| çocuk oyun | 4,203266E-05 |
| oyun çocuk | 4,203266E-05 |
| anne eğitim | 4,191674E-05 |
| oyun çocukların | 3,601456E-05 |
| çocukların oyun | 3,601456E-05 |
| çocuğun çocuk | 3,532918E-05 |
| yaş eğitim | 3,316304E-05 |
| eğitim okul | 3,181632E-05 |
| çocuklar eğitim | 3,114296E-05 |
| anne çocuk | 3,007509E-05 |
| çocuk anne | 3,007509E-05 |

A1.3 T - TEST

| BIGRAM | VALUE |
|---------------------|-----------|
| özel eğitim | 16,225563 |
| okul öncesi | 16,125187 |
| özel gereksinimli | 11,037330 |
| eğitim fakültesi | 10,816595 |
| öncesi eğitim | 10,609242 |
| üniversitesi eğitim | 10,524101 |
| normal gelişim | 10,514584 |
| gelişim gösteren | 10,394865 |
| fakültesi dergisi | 10,150249 |
| child development | 9,904733 |
| anne baba | 9,691362 |
| erken çocukluk | 9,439801 |
| eğitim dergisi | 9,430535 |
| bilişsel gelişim | 9,090426 |

| | |
|---------------------|----------|
| eđitim bilimleri | 8,891401 |
| ankara üniversitesi | 8,815331 |
| yüksek lisans | 8,753843 |
| down sendromlu | 8,477750 |
| ilk yardım | 8,216112 |
| lisans tezi | 7,736046 |

A1.4 CHI – SQUARE

| BIGRAM | VALUE |
|-----------------------------------|---------------|
| cellular subscriptions | 121669,003951 |
| elektrik çarpması | 121669,003951 |
| telephone lines | 121669,003951 |
| eleştirisinin eleştirisi | 121669,002632 |
| elif sazak | 121669,002632 |
| optimumdergi.usak.edu.tr balcılar | 121669,001410 |
| cengiz gökşen | 121669,001258 |
| hayriye bilginer | 121669,001258 |
| kaynaştırma uygulamasına | 121669,001258 |
| atakurt işıl | 121669,001258 |
| bekir onur | 121669,001258 |
| vücuda kesici | 121669,001258 |
| çağlayan dinçer | 121669,001258 |
| babaanne anneanne | 121669,001258 |
| behav pediater | 121669,001258 |
| birleşmiş milletler | 121669,001258 |
| bütünleme süreçlerindeki | 121669,001258 |
| del bambino | 121669,001258 |
| dondurma yenilmez | 121669,001258 |
| eastern mediterranean | 121669,001258 |

A1.5 LIKELIHOOD

| BIGRAM | VALUE |
|------------------|-------------|
| children with | 1552,437741 |
| primary school | 1408,418367 |
| yürüttüğü down | 1382,829234 |
| merkezi'nde down | 1379,010149 |
| mit down | 1377,284058 |
| özetle down | 1377,284058 |
| kasların motor | 1376,099119 |
| goz motor | 1372,821191 |
| dönemdeki down | 1370,514086 |

| | |
|---------------------|-------------|
| duyusal motor | 1370,271290 |
| algı motor | 1369,825928 |
| developmental motor | 1365,874814 |
| ülkemizde son | 1365,572751 |
| karşı son | 1363,926619 |
| özellekle down | 1362,451992 |
| gelişimi down | 1359,973318 |
| sosyal motor | 1358,854152 |
| çocukların motor | 1355,042904 |
| normal gelişim | 1206,530322 |
| fakültesi dergisi | 1183,494385 |

A1.6 MUTUAL INFORMATION

| BIGRAM | VALUE |
|------------------------------|-----------|
| a.n gutierrez | 16,892602 |
| aba eya | 16,892602 |
| abd'deki avustralya'daki | 16,892602 |
| ablamgile gittim | 16,892602 |
| abusive versus | 16,892602 |
| acaba çocuğumuz | 16,892602 |
| acad sci | 16,892602 |
| accid emerg | 16,892602 |
| acizliği katmerleştirmesidir | 16,892602 |
| acous tic | 16,892602 |
| acqui sition | 16,892602 |
| adaylarından başlamak | 16,892602 |
| adaylarını kapsamıştır | 16,892602 |
| adler'den aktardığına | 16,892602 |
| ado lescence | 16,892602 |
| adurakoglu mynet.com | 16,892602 |
| advancing democracy | 16,892602 |
| affection envy | 16,892602 |
| afyon kocatepe | 16,892602 |
| ai'ittemı tidjs | 16,892602 |

A2.LAW

A2.1 FREQUENCY

| | |
|--------------------|-----|
| söz konusu | 589 |
| toplu iş | 396 |
| türk borçlar | 360 |
| iş sözleşmesi | 343 |
| ihtiyati tedbir | 320 |
| iş kanunu | 285 |
| iş hukuku | 274 |
| insan hakları | 263 |
| belirsiz alacak | 246 |
| alt işveren | 205 |
| sayılı iş | 204 |
| sosyal güvenlik | 201 |
| evde hizmet | 190 |
| iş sözleşmesinin | 187 |
| yer alan | 177 |
| borçlar kanunu'nun | 172 |
| belirli süreli | 164 |
| asıl işveren | 159 |
| kısmi dava | 158 |
| hukuk devleti | 157 |

A2.2 NULL – HYPOTHESIS

| BIGRAM | VALUE |
|--------------|-----------------|
| iş iş | 9,970052600E-05 |
| iş hukuk | 5,444355300E-05 |
| sayılı iş | 3,505648400E-05 |
| ilişkin iş | 3,345882000E-05 |
| çalışma iş | 3,183043000E-05 |
| işçinin iş | 2,878871800E-05 |
| iş işçinin | 2,878871800E-05 |
| işveren iş | 2,820495500E-05 |
| iş söz | 2,786698900E-05 |
| işverenin iş | 2,626932100E-05 |
| iş hukuku | 2,537831500E-05 |
| genel iş | 2,408789000E-05 |

| | |
|---------------|-----------------|
| konusu iş | 2,368847500E-05 |
| işçi iş | 2,245950000E-05 |
| tarafından iş | 2,067748800E-05 |
| iş sözleşme | 2,055459100E-05 |
| sayılı hukuk | 1,914332600E-05 |
| borçlar iş | 1,907982200E-05 |
| iş hizmet | 1,895692500E-05 |
| sözleşmesi iş | 1,846533600E-05 |

A2.3 T – TEST

| BIGRAM | VALUE |
|--------------------|-----------|
| söz konusu | 24,180660 |
| toplu iş | 19,656395 |
| türk borçlar | 18,914246 |
| iş sözleşmesi | 18,196236 |
| ihtiyati tedbir | 17,853453 |
| iş kanunu | 16,527065 |
| insan hakları | 16,176014 |
| iş hukuku | 16,054690 |
| belirsiz alacak | 15,661911 |
| alt işveren | 14,208326 |
| sosyal güvenlik | 14,127565 |
| evde hizmet | 13,735153 |
| sayılı iş | 13,485193 |
| iş sözleşmesinin | 13,388565 |
| yer alan | 13,254636 |
| borçlar kanunu'nun | 13,066797 |
| belirli süreli | 12,781374 |
| kısmi dava | 12,527276 |
| asıl işveren | 12,504008 |
| hukuk devleti | 12,444238 |

A2.4 CHI-SQUARE

| | |
|----------------------------------|-----------------|
| uğur yağcı | 324987,00716644 |
| ggüüvveennlliikk kkoonnssseeyyii | 324987,00716644 |
| hye jeong | 324987,00716644 |
| justes motifs | 324987,00716644 |
| rahime erbaş | 324987,00716644 |
| tchd cehamer | 324987,00716644 |

| | |
|-------------------------|-----------------|
| wessels beulke | 324987,00716644 |
| böy lece | 324987,00716644 |
| computer arbeitsplätze | 324987,00716644 |
| dör düncü | 324987,00716644 |
| job sharing | 324987,00716644 |
| mühf had | 324987,00716644 |
| sharing computer | 324987,00716644 |
| tomris mengüşođlu | 324987,00716644 |
| www.belgenet.com arshiv | 324987,00716644 |
| www.cvce.eu viewer | 324987,00716644 |
| affari sociali | 324987,00716644 |
| aggressive incumbents | 324987,00716644 |
| aksoyođlu necati | 324987,00716644 |
| aktiflerin satılmasını | 324987,00716644 |

A2.5 LIKELIHOOD

| BIGRAM | VALUE |
|----------------------|-------------|
| netice sebebiyle | 2233,422417 |
| sözleşmesi yapma | 1794,138414 |
| berner kommentar | 1635,004777 |
| terör yapt | 1617,799444 |
| bununla birlikte | 1606,029364 |
| yargılama sırasında | 1557,359884 |
| alt nda | 1501,648376 |
| alınan işte | 1463,071043 |
| belirsiz alacak | 1420,293400 |
| işverenin eşit | 1411,448019 |
| isteđe bađlı | 1379,199133 |
| prof dr | 1256,790003 |
| chkd cilt | 1251,089330 |
| alacak davası | 1250,001793 |
| yenilik doğuran | 1230,316276 |
| öte yandan | 1189,909026 |
| sebebiyle ađırlaşmış | 1138,117819 |
| ola rak | 1124,559344 |
| kıdem tazminatı | 1104,105030 |
| arka plandaki | 1093,978578 |

A2.6 MUTUAL INFIRMATION

| BIGRAM | VALUE |
|------------------------|-----------|
| aanndllaa mmaallaarraa | 18,310022 |

| | |
|--------------------------|-----------|
| aavvruupaa bbiirllii | 18,310022 |
| abdel nour | 18,310022 |
| abouts facts.asp | 18,310022 |
| ab`ye katılımdaki | 18,310022 |
| aczi iflası | 18,310022 |
| acısı çıkarılmamış | 18,310022 |
| acıyı azaltmakla | 18,310022 |
| adamdan oluşurlardı | 18,310022 |
| adaylarında girecekleri | 18,310022 |
| adetlere emsallere | 18,310022 |
| adlandı rılabilecek | 18,310022 |
| administratives théorie | 18,310022 |
| advice index.pdf | 18,310022 |
| aerial incident | 18,310022 |
| af duyguların | 18,310022 |
| affect reconviction | 18,310022 |
| affirmation surabondante | 18,310022 |
| afga nistan | 18,310022 |
| afrika'da fanon | 18,310022 |

A3. GEOGRAPHY

A3.1 FREQUENCY

| BIGRAM | VALUE |
|----------------------|-------|
| coğrafya dergisi | 301 |
| yer alan | 186 |
| sosyal bilgiler | 182 |
| ege coğrafya | 181 |
| sosyo ekonomik | 169 |
| aegean geographical | 163 |
| geographical journal | 163 |
| öğretmen adaylarının | 158 |
| dergisi aegean | 153 |
| journal vol | 153 |
| coğrafya öğretmen | 149 |
| yıllık ortalama | 131 |
| ekonomik seviye | 125 |
| kıyı çizgisi | 104 |
| yer almaktadır | 96 |
| orman yangınlarının | 95 |
| dergisi sayı | 87 |
| söz konusu | 86 |
| coğrafi bilgi | 80 |

A3.2 NULL – HYPOTHESIS

| BIGRAM | VALUE |
|-------------------|--------------|
| coğrafya coğrafya | 4,183310E-05 |
| arasında coğrafya | 2,046255E-05 |
| doğal coğrafya | 1,819253E-05 |
| yılında coğrafya | 1,481994E-05 |
| coğrafya dergisi | 1,413894E-05 |
| üzerinde coğrafya | 1,329579E-05 |
| alan coğrafya | 1,326336E-05 |
| ekonomik coğrafya | 1,326336E-05 |
| coğrafya sosyal | 1,280936E-05 |
| istanbul coğrafya | 1,258236E-05 |
| ege coğrafya | 1,242022E-05 |
| coğrafya ege | 1,242022E-05 |
| arasında çevre | 1,130992E-05 |
| çevre arasında | 1,130992E-05 |
| coğrafya öğretmen | 1,096092E-05 |
| öğretmen coğrafya | 1,096092E-05 |
| önemli çevre | 1,050335E-05 |
| doğal çevre | 1,005525E-05 |
| çevre doğal | 1,005525E-05 |
| arasında yer | 9,993337E-06 |

A3.3 T – TEST

| BIGRAM | VALUE |
|----------------------|-----------|
| coğrafya dergisi | 17,186810 |
| yer alan | 13,543453 |
| sosyal bilgiler | 13,461378 |
| ege coğrafya | 13,269497 |
| sosyo ekonomik | 12,971132 |
| aegean geographical | 12,751492 |
| geographical journal | 12,747278 |
| öğretmen adaylarının | 12,546750 |
| journal vol | 12,351433 |
| dergisi aegean | 12,337153 |
| coğrafya öğretmen | 12,027461 |
| yıllık ortalama | 11,364102 |
| ekonomik seviye | 11,150077 |
| kıyı çizgisi | 10,166154 |

| | |
|---------------------|----------|
| yer almaktadır | 9,760240 |
| orman yangınlarının | 9,728819 |
| dergisi sayı | 9,293161 |
| söz konusu | 9,266908 |
| coğrafi bilgi | 8,901550 |
| yılları arasında | 8,838717 |

A3.4 CHI-SQUARE

| BIGRAM | VALUE |
|-----------------------------------|--------------|
| ayşe çağliyan | 19944,800985 |
| çamlık mağaraları | 19944,800985 |
| doğrultu atımlı | 19944,800517 |
| giris cikis | 19944,800517 |
| marly bare | 19944,800517 |
| antep fıstığı | 19944,800517 |
| diñ durmaz | 19944,800517 |
| eşiği depresyonlarını | 19944,800517 |
| fi gu | 19944,800517 |
| gaussian filter | 19944,800517 |
| geka.org.tr yukleme | 19944,800517 |
| geolsci micropal | 19944,800517 |
| kontey ner | 19944,800517 |
| lokanta kahvehane | 19944,800517 |
| micropal foram.html | 19944,800517 |
| necmettin erbakan | 19944,800517 |
| sorunla karşılaşmadığı | 19944,800517 |
| talveg kotundaki | 19944,800517 |
| www.gumrukticaret.gov.tr altsayfa | 19944,800517 |
| www.jains.com.tr uploaded | 19944,800517 |

A3.5 LIKELYHOOD

| BIGRAM | VALUE |
|----------------------|-------------|
| mezunu öğretmenlerin | 2266,892463 |
| yayınları no | 1772,654693 |
| erinç kuraklık | 1668,385116 |
| ortaçağ sıcak | 1588,722641 |
| faktör analizi | 1584,762470 |
| rearranged from | 1425,582963 |
| analizler sonucunda | 1406,904730 |

| | |
|--------------------------|-------------|
| associated with | 1400,864880 |
| maden çayı | 1395,313867 |
| etkinlik indisi | 1394,902071 |
| özellekle kış | 1394,061972 |
| analizi sonucunda | 1389,335717 |
| etkisi altında | 1387,282842 |
| formasyonlarının alansal | 1385,115607 |
| deniz altında | 1383,170823 |
| su altında | 1382,051339 |
| şehircilik müdürlüğü | 1378,385491 |
| şubesi müdürlüğü | 1377,477438 |
| yanan alanlarda | 1377,146539 |
| kıyasla alansal | 1376,118246 |

A3.6 MUTUAL INFORMATION

| BIGRAM | VALUE |
|------------------------------|-----------|
| a.coğrafya öğretmeninden | 17,605653 |
| abrasion platform | 17,605653 |
| acaglayan firat.edu.tr | 17,605653 |
| accompanied assessments | 17,605653 |
| acquire guidance | 17,605653 |
| acreage shown | 17,605653 |
| adalarından endonezya'ya | 17,605653 |
| adetten başladığından | 17,605653 |
| adnksdagitapp adnks.zul | 17,605653 |
| afel günöte | 17,605653 |
| affairs reviw | 17,605653 |
| agregat stabilitesi | 17,605653 |
| ailem akrabalarımın | 17,605653 |
| akamete uğrattır | 17,605653 |
| akarsulan dağlan | 17,605653 |
| akarsularının önemlilerinden | 17,605653 |
| akarya kıt | 17,605653 |
| akaryakıtla dolduruyordu | 17,605653 |
| akdağ'ın kuzeyindekiler | 17,605653 |
| akdilek aybastı | 17,605653 |

A4.ECONOMICS

A4.1 FREQUENCY

| BIGRAM | VALUE |
|----------------------|-------|
| sosyal bilimler | 219 |
| kayıt dışı | 217 |
| yeni ekonomi | 194 |
| kayıtdışı ekonomi | 174 |
| bilimler enstitüsü | 170 |
| enstitüsü dergisi | 158 |
| söz konusu | 130 |
| dergisi cilt | 116 |
| üniversitesi sosyal | 111 |
| üçüncü yol | 105 |
| kayıtdışı ekonominin | 101 |
| sosyal bilgiler | 94 |
| ortaya çıkan | 82 |
| bilimler dergisi | 74 |
| dışı ekonomi | 74 |
| ç.ü sosyal | 72 |
| dışı ekonominin | 69 |
| yer alan | 68 |
| bütçe açıkları | 67 |
| ankara üniversitesi | 65 |

A4.2 NULL - HYPOTHESIS

| BIGRAM | VALUE |
|------------------|--------------|
| ekonomi ekonomi | 4,589258E-05 |
| ekonomi ekonomik | 4,452876E-05 |
| ekonomi sosyal | 3,300447E-05 |
| ekonomik sosyal | 3,202365E-05 |
| sosyal ekonomik | 3,202365E-05 |
| ekonomi yeni | 2,795833E-05 |
| yeni ekonomi | 2,795833E-05 |
| yeni ekonomik | 2,712748E-05 |
| sosyal yeni | 2,010674E-05 |
| yeni sosyal | 2,010673E-05 |
| ekonomi önemli | 1,899121E-05 |
| önemli ekonomik | 1,842683E-05 |
| yeni yeni | 1,703256E-05 |
| ekonomi dergisi | 1,534299E-05 |
| ekonomi türkiye | 1,503613E-05 |
| türkiye ekonomi | 1,503613E-05 |
| ekonomi ortaya | 1,496794E-05 |

| | |
|-------------------|--------------|
| türkiye ekonomik | 1,458929E-05 |
| kayıtdışı ekonomi | 1,404736E-05 |
| ekonomi kayıtdışı | 1,404736E-05 |

A4.3 T - TEST

| BIGRAM | VALUE |
|----------------------|-----------|
| sosyal bilimler | 14,709760 |
| kayıt dışı | 14,701759 |
| yeni ekonomi | 13,529561 |
| bilimler enstitüsü | 13,019436 |
| kayıtdışı ekonomi | 12,979317 |
| enstitüsü dergisi | 12,537012 |
| söz konusu | 11,389083 |
| dergisi cilt | 10,732900 |
| üniversitesi sosyal | 10,383516 |
| üçüncü yol | 10,219859 |
| kayıtdışı ekonominin | 9,968168 |
| sosyal bilgiler | 9,627020 |
| ortaya çıkan | 9,030254 |
| bilimler dergisi | 8,531238 |
| ç.ü sosyal | 8,443942 |
| dışı ekonomi | 8,383397 |
| dışı ekonominin | 8,239922 |
| yer alan | 8,216054 |
| bütçe açıkları | 8,168818 |
| ankara üniversitesi | 8,008652 |

A4.4 CHI - SQUARE

| BIGRAM | VALUE |
|----------------------|---------------|
| vrđ tmt | 198689,014368 |
| asuman oktayer | 198689,014368 |
| yaln zca | 198689,010482 |
| asia minor | 198689,007771 |
| dwlk á | 198689,007771 |
| abant izzet | 198689,007771 |
| akgul.bilkent.edu.tr | 198689,007771 |
| izzet baysal | 198689,007771 |
| içerip içermediğini | 198689,007771 |
| kapi talist | 198689,007771 |
| ker porter | 198689,007771 |
| marie claire | 198689,007771 |

| | |
|------------------|---------------|
| mektep hocası | 198689,007771 |
| tansu çiller | 198689,007771 |
| tepeden inmecei | 198689,007771 |
| thousand oaks | 198689,007771 |
| ulgen tam.doc | 198689,007771 |
| uçan şatolar | 198689,007771 |
| www.adana to.org | 198689,007771 |
| dos santos | 198689,005481 |

A4.5 LIKELIHOOD

| BIGRAM | VALUE |
|----------------------|-------------|
| test sonuçları | 1585,679007 |
| televizyon yay | 1497,630039 |
| döneme ait | 1464,124933 |
| quarterly journal | 1389,446052 |
| kalkınmış ülkelerde | 1388,573809 |
| üzerine etkileri | 1385,943418 |
| european journal | 1381,194135 |
| american journal | 1380,526090 |
| iktisadi etkileri | 1380,203765 |
| growth journal | 1376,007505 |
| uygun politikaları | 1372,179486 |
| istikrarlı risk | 1371,696365 |
| economic journal | 1367,293458 |
| kaynaklanan yapısal | 1366,711242 |
| insanların risk | 1366,656817 |
| itibaren yapısal | 1365,657527 |
| dolayısıyla yapısal | 1363,777481 |
| arasındaki yapısal | 1362,522851 |
| üçüncü yol | 1328,532381 |
| kayıtdışı ekonominin | 998,784176 |

A4.6 MUTUAL INFORMATION

| BIGRAM | VALUE |
|--------------|-----------|
| a.lıdıa'ı sö | 17,600152 |
| a.o balkanlı | 17,600152 |
| a.vası tasız | 17,600152 |
| a:reak selm | 17,600152 |

| | |
|----------------------------|-----------|
| aat mühendisleri | 17,600152 |
| abbasi s.m | 17,600152 |
| abdelma lek | 17,600152 |
| abi deye | 17,600152 |
| abul eenea | 17,600152 |
| ab'nin benimsediđi | 17,600152 |
| acti vity | 17,600152 |
| addansa sasa | 17,600152 |
| adl yazllmaltdu | 17,600152 |
| adli kitabinin | 17,600152 |
| adnıbı tmtii | 17,600152 |
| adolf sootbeer | 17,600152 |
| adun etmiştir | 17,600152 |
| ady worswick | 17,600152 |
| aeyl f.dtm | 17,600152 |
| ag.daki karakteristiklerde | 17,600152 |

REFERENCES

1. **Stefan Evert** ,(2004), "*The Statistics of Word Co-occurrences Word Pairs and Collocations*". PhD Thesis University of Stuttgart.
2. **Maria Teresa Pazienza, Marco Pennacchiotti, and Fabio Massimo Zanzotto**,(2005), "*Terminology extraction: an analysis of linguistic and statistical approaches*". University of Roma Tor Vergata, Italy.
3. **Violeta Seretan, Eric Wehrli**,(2006), "*Accurate Collocation Extraction Using a Multilingual Parser*". Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual meeting of the ACL, pages 953-960, Sydney.
4. **Ted Dunning** ,(1993), "*Accurate Methods for the Statistics of Surprise and Coincidence*". New Mexico State University.
5. **Anca Dinu, Liviu P. Dinu, Ionut T. Sorodoc**,(2014), "*Aggregation methods for efficient collocation detection*". In Proc. LREC 2014 (9th International Conference on Language Resources and Evaluation), Reykjavik, Iceland, 26-31 may 2014, pages 4041-4045.
6. **Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina and Josef van Genabith**, (2010), "*Automatic Extraction of Arabic Multiword Expressions*". Proceedings of the Workshop on Multiword Expressions: from Theory to Applications (MWE 2010), pages 18-26, Beijing
7. **Christopher D. Manning, Hinrich Schütze**, (1999), "*Foundations of Statistical Natural Language Processing*". Massachusetts Institute of Technology.
8. **Senem Kumova Metin, Bahar Karaoğlan**,(2010), "*Collocation Extraction in Turkish Texts Using Statistical Methods*". IceTAL 2010 Proceedings of the 7th International Conference on Advances in Natural Language Processing, 6233, pages 238-249.

9. **FIRTH, J. R., (1957)** "*Modes of Meaning*", Papers in Linguistics 1934-1951, London:Oxford University Press, s. 190-215.
10. **HALLIDAY, M. A. K., (1966).** "*Lexis as a Linguistic Level*" , (Ed. C.E. Bazell vd.) In Memory of F. R. Firth, London:Longman, s. 148-162.
11. **MEL'CUK, Igor, (1998),** "*Collocations and Lexical Functions (Ed. A.P. Cowie) Phraseology: theory, analysis, and applications, Oxford University Pres, s. 23-54.*
12. **NESSSELHAUF, Nadja, (2005)**" *Collocations in a Learner Corpus*". John Benjamins Publishing, 2005.
13. **CRUSE, D. A., (1986)** "*Lexical Semantics*". Cambridge University Press.
14. **KJELLMER, Goran, (1994)** , "*A Dictionary of English Collocations*". Clarendon Press Oxford.
15. **MITCHELL, T.F., (1971)** "*Linguistic goings on. Collocations and other lexical matters arising on the syntagmatic record*". Archivum Linguisticum, 2:35-69.
16. **Kemal Oflazer, Özlem Çetinoğlu ,Bilge Say, (2004),** "*Integrating Morphology with Multi-word Expression Processing in Turkish*". In Proceedings of 2nd ACL Workshop on Multiword Expressions: Integrating Processing, pages 64-71, Spain.
17. **Turkish Academy of Science www.tubaterim.gov.tr ,(2015)**
18. **J.Han, M.Kamber, J.Pei, (2011)** "*Data Mining Concepts and Techniques*". Third edition, Morgan Kaufmann.