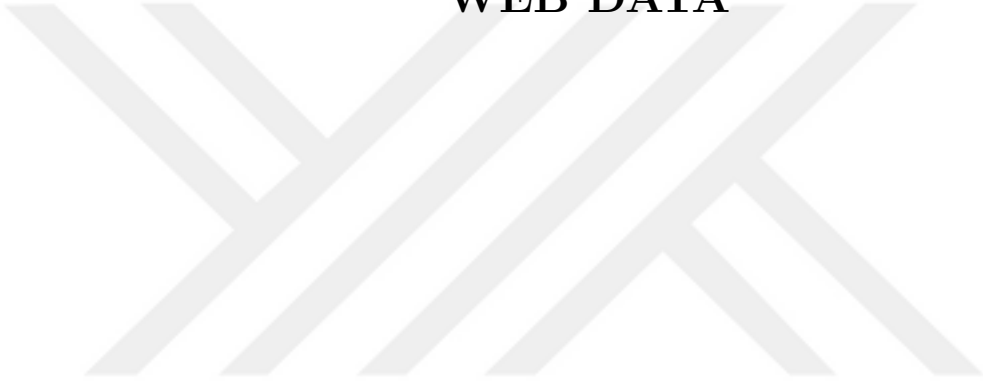


**IDENTIFICATION OF MULTIWORD
EXPRESSIONS IN TURKISH BASED ON
WEB DATA**



HANDE AKA UYMAZ

JUNE 2016

IDENTIFICATION OF MULTIWORD EXPRESSIONS IN TURKISH BASED ON WEB DATA

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF
NATURAL AND APPLIED SCIENCES OF
IZMIR UNIVERSITY OF ECONOMICS

BY
HANDE AKA UYMAZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE
IN THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

JUNE 2016

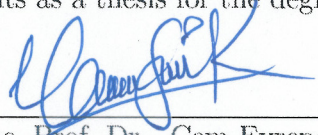
M.S. THESIS EXAMINATION RESULT FORM

Approval of the Graduate School of Natural and Applied Sciences




Prof. Dr. İsmihan Bayramoğlu
Director

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



Assoc. Prof. Dr. Cem Evrendilek
Head of Department

We have read the thesis entitled **“Identification of Multiword Expressions in Turkish based on Web Data”** completed by **Hande Aka Uymaz** under supervision of **Asst. Prof. Dr. Senem KUMOVA METİN** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



Asst. Prof. Dr. Senem KUMOVA METİN
Supervisor

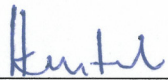
Examining Committee Members

Date: 27.06.2016


Asst. Prof. Dr. Senem KUMOVA METİN
Dept. of Software Engineering, IUE



Asst. Prof. Dr. Kaan KURTEL
Dept. of Software Engineering, IUE



Asst. Prof. Dr. Tarık KIŞLA
Dept. of Computer Education and Instructional Technologies, Ege U.



ABSTRACT

IDENTIFICATION OF MULTIWORD EXPRESSIONS IN TURKISH BASED ON WEB DATA

Hande Aka Uymaz

M.S. in Computer Engineering

Graduate School of Natural and Applied Sciences

Supervisor: Asst. Prof. Dr. Senem KUMOVA METİN

June 2016

Multiword expressions (MWEs) are recurrent combinations of words in natural languages. The extraction of MWEs in a text is significant for a number of natural language processing applications (e.g. natural language generation, computational lexicography, machine translation etc.). There are various occurrence frequency based methods (e.g. joint probability, pointwise mutual information and mutual dependency) that are used frequently for MWE extraction ([12],[13]). The major disadvantage of these methods is that extraction performance depends mainly on the size of the data set in which the occurrence frequency is measured. The main goal of this thesis is obtaining the frequency from a massive data source, the World Wide Web, in order to by-pass the negative effect of small data set.

In this thesis, we applied frequency based MWE extraction methods on two Turkish MWE data sets. The occurrence frequencies of MWE candidates in data sets are obtained from popular search engine Google. The retrieved page counts when the candidates are sent as queries to Google are employed as the occurrence frequencies. The evaluation of the 20 frequency based methods is performed by precision, recall and F-measures. The performance of web-based frequencies in identification of MWEs is compared to the traditional corpus based frequencies and it is showed that the use of web data in identification of MWEs reveals promising results.

Keywords: Multiword expression, frequency based methods, web data.

ÖZ

WEB VERİSİ KULLANILARAK TÜRKÇE ÇOK SÖZCÜKLÜ İFADELERİN BELİRLENMESİ

Hande Aka Uymaz

Bilgisayar Mühendisliği, Yüksek Lisans

Fen Bilimleri Enstitüsü

Tez Danışmanı: Yrd. Doç. Dr. Senem KUMOVA METİN

Haziran 2016

Çok sözcüklü ifade, doğal dillerde, sözcüklerin anlam bütünlüğü oluşturmak üzere tekrarlayan kombinasyonlarıdır. Metinlerden çok sözcüklü ifadelerin belirlenmesi bir çok doğal dil işleme uygulamaları (Doğal dil üretme, hesaplamalı sözlükbilim, makine çevirileri vb.) için çok önemli bir konudur. Çok sözcüklü ifadelerin belirlenmesi için gözlenme sıklığı bağımlı yöntemler (Bileşik olasılık (joint probability), noktasal karşılıklı bilgi katsayısı (pointwise mutual information), karşılıklı bağıllık (mutual dependency) v.b) sıklıkla kullanılır. Bu yöntemlerin en büyük dezavantajı, çok sözcüklü ifadelerin belirlenmesinin performansının frekansın ölçüldüğü veri kaynağının büyüklüğüne bağlı olmasıdır. Bu tezin amacı, küçük veri setlerinin yarattığı problemlerin önüne geçmek için bilinen en büyük veri kaynağı olan web'i kullanarak gözlenme sıklığını elde etmektir.

Bu tezde, 2 farklı aday veri seti kullanılarak, Türkçe dili için frekans tabanlı çok sözcüklü ifade belirleme metotlarının performansı araştırılmıştır. Veri setlerindeki adayların gözlenme sıklığı bilgisi popüler bir arama motoru olan Google kullanılarak elde edilmiştir. Aday çok sözcüklü ifadelerin arama motoruna sorgu olarak gönderildiğinde alınan sayfa sayısı (ing. page count) adayın gözlenme sıklığı olarak kabul edilmiştir. Kullanılan 20 yöntemin başarısı anma(recall), duyarlılık(precision) ve F-ölçütü (F-measure) ile değerlendirilmiştir. Web tabanlı frekans bilgisinin çok sözcüklü ifadelerin belirlenmesindeki performansı geleneksel derlem tabanlı frekans ile karşılaştırılmıştır ve çok sözcüklü ifadelerin belirlenmesinde web verilerinin kullanılması umut verici sonuçlar göstermiştir.

Anahtar Kelimeler: Çok sözcüklü ifade, sıklık tabanlı yöntemler, web verisi.

ACKNOWLEDGEMENT

Firstly, I would like to thank to my advisor, Asst. Prof. Dr. Senem Kumova Metin who gave me the opportunity to work with her, for her excellent guidance, encouragement and patience.

This thesis has been supported by the Scientific and Technological Research Council of Turkey (Tübitak Project no: 115E469). I would like to thank to Tübitak for the support.

I am very thankful to my parents Murat Tamer Aka and Filiz Aka and my sister Gözde Aka for their trust and unconditional support all the time.

Lastly, I would like to thank to my husband Mehmet Erdem Uymaz for always being with me and motivating me with his endless support.

TABLE OF CONTENTS

Front Matter	i
Abstract	iii
Öz	iv
Acknowledgement	v
Table of Contents	vi
List of Tables	vii
List of Figures	viii
1 Introduction	1
2 Literature Review	5
3 Proposed Method	9
4 Experimental Set-up	13
4.1 Base sets	13
4.1.1 Base Set 1	14
4.1.2 Base Set 2	16
4.2 Annotation of Base sets	17
4.3 Evaluation Measures	24
5 Experimental Results	26
6 Conclusion	36
A	42

LIST OF TABLES

3.1	Lexical association measures used for ranking MWE candidates . . .	10
3.2	Sample query results	12
4.1	Word & character counts of 6 corpora	14
4.2	Observed cases in annotation of BS1	21
4.3	Exceptional cases in annotation of BS1	22
4.4	Observed cases in annotation of BS2	22
4.5	Annotated base sets	23
4.6	Agreement statistics for BS 1	24
4.7	Agreement statistics for BS 2	24
4.8	Sample results after utilizing jaccard method for BS1	25
4.9	Sample results from the calculation of best case	25
5.1	Test results of the association measures for Base set 1 sorted according to F_AVG	27
5.2	Test results of the association measures for Base Set 2 sorted according to F_AVG	28
5.3	Sorted list of the association measures according to their success for BS 1	29
5.4	Sorted list of the association measures according to their success for BS 2	30
5.5	Sorted list of the association measures according to F_AVG	35

LIST OF FIGURES

1.1	Flow chart diagram	3
3.1	A sample query for bigrams	12
4.1	Flow chart diagram which shows the construction process of BS1 .	15
4.2	Flow chart diagram which shows the construction process of BS2 .	17
5.1	F-measure graph for BS1	33
5.2	F-measure for BS2	34
A.1	Precision graph for BS1	43
A.2	Recall graph for BS1	44
A.3	Precision graph for BS2	45
A.4	Recall graph for BS2	46

Chapter 1

Introduction

The multi word expression (MWE) is a combination of two or more words that correspond to some conventional way of saying things [5]. In many previous studies the term “collocation” is used instead of “MWE”. The notion of MWE has been first defined by J.R.Firth, in 1967 [1]. He states that a word can be understood by the company it keeps. In his further study, he states “collocations of a given word are statements of the habitual or customary places of that word”. Later, Sinclair, defined the same term as the occurrence of two or more words within a short space of each other in a text [2]. On the other hand, Hoey gives a definition with a different approach, stating that a collocation is the appearance of two or more lexical items together with a probability that cannot be interpreted as random [3].

It is hard to define what is multi word expression and what is not, because there are no known rules to construct an MWE. However, in previous studies, some common features that are hold by all MWEs are defined. Those features will be defined in chapter 2. For example, meaning integrity is one of the most important properties which enables collocations to create unit blocks in language. In other words, meaning of the whole is not the meaning of the constituents. For instance, the term “White House” in English is “Beyaz Saray” rather than “beyaz ev” in Turkish.

Multi word expressions are significant for a number of applications such as natural language generation, computational lexicography, parsing, machine translation, word sense disambiguation, part of speech tagging (POS) , information retrieval, corpus linguistic research, and some social studies through language ([4], [5]).

In order to be used in this wide range of applications there are a variety of methods such as statistical, rule based and linguistic methods. Generally, statistical methods utilize frequency property. For this reason, sources that is suitable for measuring frequency feature is required. Furthermore, performance of the multiword expression extraction methods depends on the corpus and the data source used to construct the corpus. If the corpus involves texts of different content it is accepted to be a better representative of the language. For example, it is difficult to see the term “Beyaz Saray” in a corpus which consists of articles about medicine. The World Wide Web, which contains heterogeneous live data, is a natural resource for human language technologies [6].

In this thesis, we aim to analyze extraction performance of frequency based MWE extraction methods when the frequency is obtained from web sources by the use of search engines. The term MWE in this thesis is limited to bigram which is the consecutive two word combination in text. We accept a bigram as a MWE if the bigram is in the one of the following groups;

- Phrasal verbs and idioms(e.g. “açığa vurmak”, “öne sürmek”)
- Stock phrases (e.g. “sert kahve”, “acı gerçek”)
- Technical terms (e.g. “moleküler genetik”, “antipsikotik ilaç”)
- Named entities including proper names and job titles (e.g. “Türkiye Cumhuriyeti” , “genel müdür”)

We built our base sets (e.g. the candidate MWE list) by utilizing different corpora. Then, the candidate bigrams are annotated by 3 to 4 human judges.

Following, we applied frequency based statistical methods in our base sets such as, pointwise mutual information, normalized expectation and first Kulczynsky. In order to utilize these methods; obtain occurrence frequency; the World Wide Web is used as data source. We have preferred the mostly used search engine, Google, to retrieve the page counts of the candidate MWEs from the internet. The page counts are accepted as occurrence frequencies of candidate data. Finally, the results are reported and the performances of the methods are measured using three metrics: precision, recall and F-measure. In Figure 1.1 flow chart which represents this process can be seen. Details about this steps can be seen in Chapter 3 and 4.

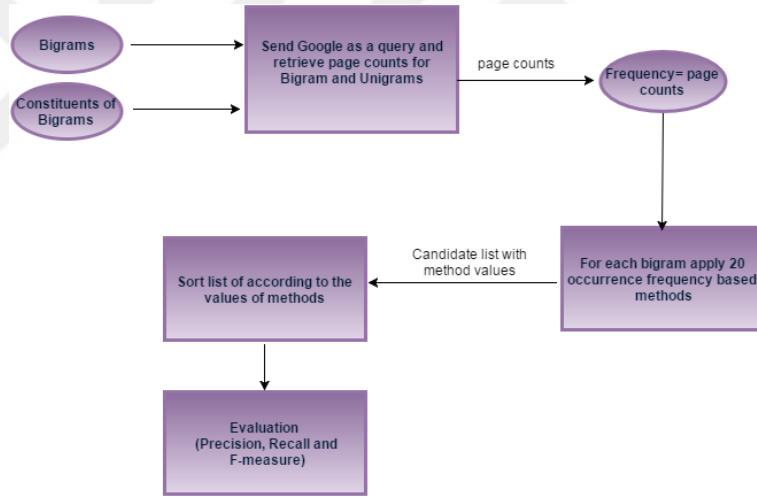


Figure 1.1: Flow chart diagram

The contribution of the thesis is summarized as follows;

1. A wide range of occurrence frequency based methods are applied on two different base sets.
2. Web data is firstly used to identify MWEs in Turkish language.

This thesis is organized as follows. In chapter 2, definition of multiword expressions and previous works about MWE extraction are presented. In section 3, the proposed method is introduced. Chapter 4 details the experimental setup

procedures including descriptions of base sets, annotation scheme and evaluation metrics. Chapter 5 includes the results, followed by a conclusion in Chapter 6.



Chapter 2

Literature Review

In recent studies, there are a variety of definitions of MWEs. In this thesis, we accept MWE and collocation as similar terms though in some studies, collocation is defined to be a type of MWEs in which the high recurrence is observed. Although each MWE definition in literature includes particular features of MWEs, there are no known rules for the formation of all types of MWEs. However, there are some common properties that are accepted in different studies.

The most widely measured and the easiest feature of the MWEs is the recurrence property. Almost all extraction techniques suggest that a MWE must differ from other word combinations in some kind of frequency measure [4], [14], [15].

The second property is being language specific. Collocations may change in different languages depending on the social or cultural behaviors of native speakers [9]. For example, in Turkish, the English MWE “wisdom teeth” is called “yirmi yaş dişleri” but the exact translation of the words to Turkish gives “akıl dişleri”. The word “wisdom” may be translated as “akıl” in Turkish. However, it gains a different meaning when it is combined with a second word. This property is very important for machine translation. For example in the English sentence from a Wikipedia article with the title wisdom tooth; “Wisdom teeth generally erupt between the ages of 17 and 25”; incorrect translation of the term “wisdom

teeth” causes meaning loss in the sentence while translating to Turkish. Furthermore, there are no known rules which define how a word chooses a particular word or word combinations from millions of different words in language while creating a MWE [9]. For instance, “sweet” is a common collocation with “dreams” in English, but there is no clear explanation for the preference of this word instead of “candy” which is a synonym of “sweet”.

Meaning integrity is another commonly accepted property of collocations. This feature enables collocations to create unit blocks which is a single word or word combination that has an individual meaning in natural language. As a result, the meaning of the whole is not the meaning of the parts [9]. The meaning integrity in collocations is also related with the property of limited compositionality. A natural language expression is called compositional if the meaning of the expression can be predicted from the meaning of the parts [5]. If constituents in the expression lose their own generally accepted meanings, then the expression is stated as non-compositional. In collocations, the meaning of a particular collocation may be predicted from a constituent because the meanings of words don’t change completely.

The last property is the domain and language dependency of the collocations. There are lots of different domain specific collocations in particular areas such as medicine, art and sports. Smadja [15] gave a descriptor example about the domain of sailing. Word combinations “dry suit” and “wet suit” are not state that a suit which is dry or wet. They are a special type of suit used by sailors to stay dry in difficult weather conditions and a special kind of suit uses for several marine activities, respectively. However, it is hard to understand these meanings easily for even native speakers.

Collocation extraction techniques can be categorized in three main groups according to recent studies which are statistical, rule-based and linguistic methods. Rule based methods have higher time complexities relative to statistical methods because they use a set of rules and require pre-processing steps to extract MWEs. For example, Ofazzer et al. [16] used rule based methods for extracting multi-word expressions in a Turkish corpus. As it is stated in their study,

they used corpora of news texts in order to evaluate their MWE extraction processor and they incrementally test and improve their semi-lexicalized rule base method. The study of Tsvetkov and Wintner [26] is an example for studies which utilize linguistic methods. They present an architecture for expressing different linguistically-motivated features which help to identify MWEs in natural language texts [26]. Furthermore, they introduce ways to compute many of these features, and define linguistically-motivated interrelationships among them that the Bayesian network models [26]. The study of Sarıkaş [27] is an example for the linguistic studies in Turkish language. They presented the difficulties and the reasons of loss of meanings such as, social, cultural differences and lexical and grammatical changes of two different languages while translating a collocation.

As it is stated in earlier studies, statistical MWE extraction typically proceeds by scoring collocation candidates with an association measure [18]. The first step is the construction of the candidate list in a corpus and the second step is the candidate ranking according to association measures. Identification of candidates is a process using specific criterion, such as frequency of candidates. Following this, a variety of mostly used statistical techniques generate a ranked list of MWE candidates and the higher scores (lower ranks) means the closer the candidate is to being a collocation. Several association measures have been utilized in the literature such as point wise mutual information, mutual dependency and t-test ([5], [17]).

In one of the studies of Pecina [12], they used 3 data sets that include corpus frequency data. For each base set they employed 55 association measures combined by standard statistical classification methods which are modified in order to provide scores for ranking [12]. They observed that methods which are the combinations of multiple association measures result in significant performance improvement [12].

In the the study of Ramisch et. al [24] a toolkit (Multiword Expression Toolkit) is presented which provides to identify type and language independent MWEs from corpora. The toolkit includes a targeted list of MWE candidates

which are extracted according to a set of standard statistical association measures and a number of user-defined criteria [24].

Wu et. al [25] used a different approach while creating a corpus for their study. They utilize a web-derived corpus and digital library software in order to produce a vast concordance and their aim was helping the students to use collocations more effectively in their writing [25].

In earlier studies on collocation extraction, a variety of methods are utilized English corpora because of the lack of tagged corpora in different languages [19]. However, recently, in a significant amount of studies, non-English corpora; such as Turkish [9], Korean [20] and Chinese [21] have been utilized in order to observe the performance of methods in different languages. For instance, in the study of Kim et.al [20], they use four statistics for dealing with the flexible word order of Korean collocations, then they separated meaningful bigrams using an evaluation function and extended the bigrams to n-gram collocations. Furthermore, Li et.al [21] presented a corpus-driven framework which generate collocations for nouns and verbs phrase, then they integrate them using statistical association measures to extract noun/verb phrase collocations.

The evaluation of MWE extraction methods is commonly performed by precision, recall and F-measure curves which is the combination of precision and recall in MWE extraction studies ([28],[29]).

Chapter 3

Proposed Method

In MWE extraction, the occurrence frequency based methods consider the occurrence frequencies of words to identify MWEs automatically and to measure how related the words are in a given MWE candidate [30].

In this thesis, occurrence frequency based methods that are listed in Table 3.1, are utilized. These methods are well-defined and commonly used in a variety of previous studies [12], [13], [18]. The methods simply try to measure the association between the constituents of the candidate based on different approaches. This is why they are named as lexical association measures [12]. The methods enable the ranking of candidates according to their tendency to be a MWE rather than classifying them as MWE or non-MWE.

# Name	Formula
1. Joint probability (JP)	$P(w_1 w_2)$
2. Conditional probability (CP)	$P(w_2 w_1)$
3. Reverse conditional probability (RCP)	$P(w_1 w_2)$
4. Pointwise mutual information (PMI)	$\log \frac{P(w_1 w_2)}{P(w_1)P(w_2)}$
5. Mutual dependency (MD)	$\log \frac{P(w_1 w_2)^2}{P(w_1)P(w_2)}$
6. Log frequency biased MD (LFMD)	$\log \frac{P(w_1 w_2)^2}{P(w_1)P(w_2)} + \log P(w_1 w_2)$
7. Normalized expectation (NE)	$\frac{2f(w_1 w_2)}{f(w_1) + f(w_2)}$
8. S cost (S)	$\log \left(1 + \frac{\min(f(w_1 \bar{w}_2), f(\bar{w}_1 w_2))}{f(w_1 w_2) + 1} \right)^{-\frac{1}{2}}$
9. U cost (U)	$\log \left(1 + \frac{\min(f(w_1 \bar{w}_2), f(\bar{w}_1 w_2)) + f(w_1 w_2)}{\max(f(w_1 \bar{w}_2), f(\bar{w}_1 w_2)) + f(w_1 w_2)} \right)$
10. R cost (R)	$\log \left(1 + \frac{f(w_1 w_2)}{f((w_1 w_2) + f(w_1 \bar{w}_2))} \right) \cdot \log \left(1 + \frac{f(w_1 w_2)}{f((w_1 w_2) + f(\bar{w}_1 w_2))} \right)$
11. First Kulczynsky (FK)	$\frac{f(w_1 w_2)}{f(w_1 \bar{w}_2) + f(\bar{w}_1 w_2)}$
12. Second Kulczynsky (SK)	$\frac{1}{2} \left(\frac{f(w_1 w_2)}{f(w_1 w_2) + f(w_1 \bar{w}_2)} + \frac{f(w_1 w_2)}{f(w_1 w_2) + f(\bar{w}_1 w_2)} \right)$
13. Braun-Blanquet (BB)	$\frac{f(w_1 w_2)}{\max(f(w_1 w_2) + f(w_1 \bar{w}_2), f(w_1 w_2) + f(\bar{w}_1 w_2))}$
14. Simpson (Simp)	$\frac{f(w_1 w_2)}{\min(f(w_1 w_2) + f(w_1 \bar{w}_2), f(w_1 w_2) + f(\bar{w}_1 w_2))}$
15. Driver-Kroeber (DK)	$\frac{f(w_1 w_2)}{\sqrt{(f(w_1 w_2) + f(w_1 \bar{w}_2)) \cdot (f(w_1 w_2) + f(\bar{w}_1 w_2))}}$
16. Piatersky-Shapiro (PS)	$P(w_1 w_2) - P(w_1)P(w_2)$
17. Jaccard (J)	$\frac{f(w_1 w_2)}{f(w_1 w_2) + f(w_1 \bar{w}_2) + f(\bar{w}_1 w_2)}$
18. Second Sokal-Sneath (SSS)	$\frac{f(w_1 w_2)}{f(w_1 w_2) + 2(f(w_1 \bar{w}_2) + f(\bar{w}_1 w_2))}$
19. Mountford (M)	$\frac{2f(w_1 w_2)}{2f(w_1 \bar{w}_2)f(\bar{w}_1 w_2) + f(w_1 w_2)f(w_1 \bar{w}_2) + f(w_1 w_2)f(\bar{w}_1 w_2)}$
20. Fager (F)	$\frac{f(w_1 w_2)}{\sqrt{(f(w_1 w_2) + (f(w_1 \bar{w}_2))(f(w_1 w_2) + f(\bar{w}_1 w_2))}} - \frac{1}{2} \max(f(w_1 \bar{w}_2), f(\bar{w}_1 w_2))$

Table 3.1: Lexical association measures used for ranking MWE candidates

In Table 3.1, $f(w_1w_2)$ is the occurrence frequency (e.g. the number of retrieved documents from Google) of a bigram “ w_1w_2 ” and $f(w_1)$ and $f(w_2)$ are the frequencies of constituents of the bigram “ w_1 ” and “ w_2 ”, respectively. $f(w_1\bar{w}_2)$ stands for a bigram that starts with word “ w_1 ” and the following word can be anything except “ w_2 ”. $f(w_2|w_1)$ is the conditional probability of w_2 given w_1 and it is calculated as follows;

$$f(w_2|w_1) = \frac{f(w_1w_2)}{f(w_1)}$$

In MWE extraction, statistical association measures are used to rank the candidates considering the association between the constituents. We utilized 20 methods for our bigram candidates. Some of them are defined below.

Joint probability is accepted to be the easiest way to score the associations between words in a text. It is the probability of the words w_1 and w_2 to occur together in the corpus. In this thesis, since it is not possible to obtain an exact number of total documents indexed in web, any probability formula given in Table 3.1 is applied without the total sample size. Point-wise mutual information is the association measure which generates a score depending on the mutual dependence of the two or more words [13]. PMI gets the highest value when $f(w_1w_2) = f(w_1) = f(w_2)$. Mutual dependency is very similar to pointwise mutual information. In MD, the term $f(w_1w_2)$ has more effect with taking the term’s square. First Kulczynski coefficient is a measure of lexical association between two consecutive words in the corpus that considers the bigrams that do not include one of the constituents [13].

In this thesis, each MWE candidate and its constituent’s occurrence frequency is accepted to be the retrieved page count from Google. Each candidate (bigram) “ $w_1 w_2$ ” and the constituents “ w_1 ” and “ w_2 ” are sent to Google search engine as queries. The number of retrieved documents is recorded in order to be used in statistical methods. Both the candidate MWEs and the the constituents of the MWEs (unigrams) are sent to Google between quotation marks. For example, Figure 3.1 presents the query for bigram “*basra körfezi*” that is sent to Google.

The retrieved number of documents, 256.000, is used as the occurrence frequency of the regarding bigram. In Table 3.2, the retrieved number of documents for the bigram “basra körfezi” and the constituting unigrams “basra” and “körfezi” are given.

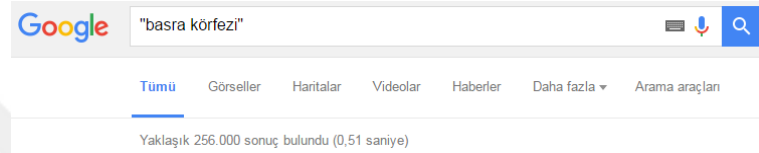


Figure 3.1: A sample query for bigrams

Search term	# retrieved documents	Notation
“basra körfezi”	256.000	$f(w_1w_2)$
“basra”	9.900.000	$f(w_1)$
“körfezi”	463.000	$f(w_2)$

Table 3.2: Sample query results

The experiments involve two candidate sets that will be mentioned as Base set 1 (BS 1), Base set 2 (BS2) from now on. The construction details of sets will be given in chapter 4. For base set 1, 6687 number of queries are sent to Google in a period of 20 days(12.01.2016-31.01.2016) and for base set 2, 4233 number of queries are sent to Google in a period of 15 days(9.03.2016-23.03.2016).

Chapter 4

Experimental Set-up

MWE extraction methods are evaluated by human annotated base sets. Base set is a set of MWE candidates that includes both positive and negative samples. Section 4.1 details the process of base set constructions in this thesis.

4.1 Base sets

In MWE extraction studies, base sets may be created by a variety of methods. For example, Pecina has studied on a base set which includes 1252 German MWE candidates randomly sampled from the 8546 distinct adjective noun pairs occurring at least 20 times in Frankfurter Rundschau corpus [22][12]. Moreover, all bigrams in the corpus (Bilkent corpus [8]), except those across sentence boundaries are retrieved to generate a base set in the study of Kumova-Metin and Karaođlan [9].

In this thesis, we constructed two base sets in order to evaluate the performance of the proposed methods in different types of MWEs. The first base set, BS 1, is formed by frequency based methods. The set includes both positive and negative samples of MWEs that are extracted from a group of corpora. The candidates in BS 1 are accepted to be representatives of MWEs that are frequently

occurring. The second base set, BS 2, is a set of idioms and bigrams that mimic the features of idioms. BS 2 is prepared to assess the proposed method on MWE candidates that are not occurring frequently in language.

The following sub-sections detail the procedures followed and the resources used in formation of BS 1 and BS 2.

4.1.1 Base Set 1

Bilkent [8], Leipzig [7], Egecorpus, BilCol [32], Muder [31] and Metu [33] corpora are utilized to construct the base set1. The number of words and characters in these corpora can be seen below, in Table 4.1.

Corpus	# Words	# Characters
Bilkent	767.132	5.111.377
Leipzig	14.279.547	110.628.416
Egecorpus	2.449.664	17.365.833
BilCol	44.150.213	347.734.602
Muder	679.750	5.391.177
Metu	1.984.634	15.222.700

Table 4.1: Word & character counts of 6 corpora

The first one is the Bilkent corpus compiled in Bilkent University in order to be used in computational linguistic studies [8]. The corpus consists of articles from popular newspapers that is collected in several years [9]. It has been morphologically analyzed by a finite state machine and sentence boundaries and stemmed forms of words have been tagged automatically ([8],[9]).

Leipzig corpora collection is compiled by Leipzig University, Department of Natural Language Processing [7]. The corpus is collected from the web and contains newspaper texts and randomly collected web pages [7].

EgeCorpus is a collection of documents in a variety of topics. The corpus is built in Ege University in International Computing Institute. In order to be used in different natural language processing studies, the documents in corpus are collected from different sources.

Muder corpus is built in Muğla Sıtkı Koçman University. Table 4.1 gives the details about the word and character counts of the corpus [31].

BilCol, is a corpus constructed in Bilkent University. It includes news on 13 different topics that are collected from 5 different Turkish news web sources throughout the year 2005 [32].

METU Turkish Corpus consists of words of post-1990 written Turkish samples [33]. The words of the corpus were taken from 10 different genres [33].

We applied normalized frequency, pointwise mutual information(PMI), chi-square test and t-score methods to bigrams that have occurrence frequency more than or equal to 5 in these corpora, similar to the study of Kumova-Metin and Karaoğlan [9]. Bigrams are sorted according to obtained values. Then, for every measure, the first 200 bigrams in the corpora are selected. The bigram lists that are obtained for a specific method are merged. Figure 4.1 represents the construction process of base set 1. In integration procedure, if any bigram is selected from more than one corpus the average of the related measure is considered.

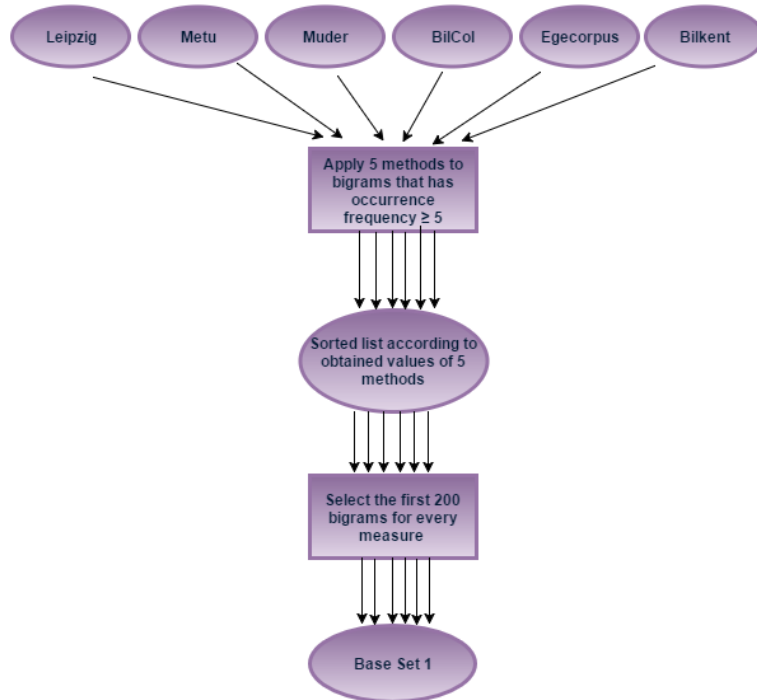


Figure 4.1: Flow chart diagram which shows the construction process of BS1

4.1.2 Base Set 2

Firstly, to form the base set 2, bigrams are retrieved from three different corpora; Leipzig [7], Bilkent [8] and Egecorpus. We merged these three corpora and listed bigrams. Secondly, a dictionary of idioms is built from four online sources; Atasözü arşivi[10], Wikisözlük[11], www.turkedebiyati.org and www.netdata.com; that accessed between 10-21.10.2015. Idiom dictionary consists of 15008 records in which 10218 are idioms of two words (bigrams). For each idiom, a list of bigrams is created by selecting the bigrams from the merged corpus that starts with the first word regarding idiom. The list of bigrams are merged. To reduce the size of merged set of 59170 items, all the bigrams that include the same first word are removed except the most frequently occurring bigram with same first word. This reduced set includes 2313 candidates. Finally, the candidates whose second word is a number (“açığı 2001”) or a single character (“ikiz i”) or is a predetermined stop word are removed from the set(“ilkel bir”). Predetermined stop words are the ones that may not be second words of MWEs. For instance, the bigram “buz gibi” is tagged as MWE and “korku ve” is tagged as non MWE. The predetermined stop words are “ama, bile, bir, bu, da, daha, de, en, için, ile, ki, ve, veya”. The final BS 2 includes 1411 MWE candidates. Figure 4.2 represents the construction process of base set 2.

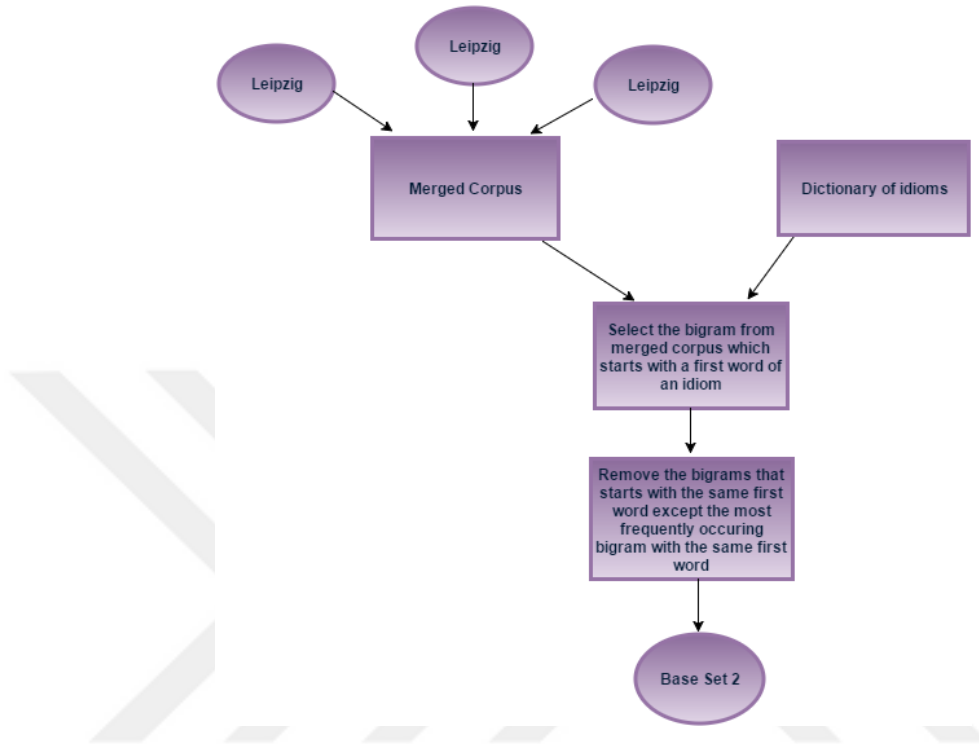


Figure 4.2: Flow chart diagram which shows the construction process of BS2

4.2 Annotation of Base sets

The reliability of base set annotation has a high impact on the evaluation of MWE extraction methods. The reliability of the annotation may be strengthened by increasing the number of judges and providing an annotation guideline to the judges. Annotation guideline is a document that includes a set of rules to be followed and exceptional cases that must be considered while annotating the base set.

In our study, 3-4 native speaker judges who are postgraduate students in computer engineering are employed in annotation of base sets. While deciding if a candidate bigram is a multi word expression or not, a common guideline is used by the judges. Well-known dictionaries and an open content online encyclopaedia, Wikipedia, that is created through the collaborative effort of user community are used for annotation of the candidates. Annotation procedure in our study has

two phases. Firstly, all the candidates in the sets are searched through Wikipedia and four online dictionaries: www.tdk.gov.tr, www.tureng.com, a collection of field specific terms ¹ and dictionary of idioms. A bigram candidate is tagged as MWE if it has at least one exact match in one of these sources. Secondly, the remaining candidates (still untagged) are reassessed based on the following 6 exceptional cases:

- Case 1: Location

If the judge thinks that the bigrams which indicates a location such as a street, an avenue or a bay but, there is no exact match for that bigram in Wikipedia, (s)he must search first word of the bigram again in Wikipedia. If it can be understood from the text that it refers to a location then, it is accepted as a multi word expression.

For instance, for the bigram “Adrasan koyu”, there is no exact match in Wikipedia. However, there exists a Wikipedia article for the word “Adrasan” that includes the sentence:

“ Sörf, su kayağı gibi aktivitelere kucak açan koyun 25 metre sualtı görüş mesafesinin olması, balıkadamları yöreye çekiyor.”

It means that “ Since the bay which has the possibilities such as surfing and water ski, has 25 meters underwater visibility range so, it is an attractive place for divers”.

It can be comprehended from the text that, “Adrasan” is a bay, so “Adrasan Koyu” is tagged as MWEs.

- Case 2: Proper names

The candidates that hold the distinguishing constituents of the company names are tagged as MWEs. For example, the candidate “Coca Cola” is considered as MWE, since Wikipedia has an article with title “Coca

¹The collection is prepared within the context of Tübitak project (115E469)

Cola Company” that includes the first two distinguishing constituents of company name.

Parts of distinctive constituents of the company names are not tagged as MWE. For instance, “American Tobacco” is not a MWE because, it is stated as “British American Tobacco” in both Wikipedia and the company’s web site.

- Case 3: Personal names

Some personal names have more than 2 constituents. Therefore, bigram candidates which compose of personal names can be in the form of, “first name+second name”, “first name+ family name” or “second name+family name”. Only the candidates with last two constituents of the personal names are accepted as MWEs. For instance, the candidate “Reşit rey” is tagged as MWE based on the Wikipedia article with title ”Ahmet Reşit Rey”². However, “Ahmet Reşit” is tagged as non MWE.

- Case 4: Typing errors

While tagging candidates that have typing errors the judges are free to tag such candidates as MWE if the typing error doesn’t change the bigram meaning and if the judge believes that the regarding error is a common misuse in language. For example, the bigram “alış veriş” is tagged as MWE by our judges although the correct form is “alışveriş”. On the other hand, the judges tend to tag bigrams(especially technical terms) that have typing errors in letters as non MWE. For example, the candidate bigram “aköz humor” is tagged as non MWE despite the fact that the correct form “aköz hümör” is a MWE.

- Case 5: Verb in the second word

Some candidates contain verbs in its second word. If an exact match can’t be found for this types of bigrams, the bigram is searched again modifying the verb to infinitive form. For example, “aklını çalıştırıp” modified into “aklını çalıştırmak” in the second turn. If there is a match in the dictionary for the infinity form “aklını çalıştırıp” is tagged as MWE.

²Ahmet Reşit Rey is a statesman in the ottoman empire era.

- Case 6: Bigrams in Trigrams

Some bigram candidates are the parts of 3 word MWEs. Even if there is an exact match in the dictionaries or Wikipedia article for that MWE of 3 consecutive words(trigram), constituents of it can't accepted as a MWE. For instance, there is a Wikipedia article with the title "çoklu çekirdekli işlemci", but bigram candidate "çoklu çekirdekli" is not tagged as MWE.

Following the guideline if a candidate is not tagged as MWE or non MWE by a judge, it is tagged as "NONE" mentioning that the judge was unable to annotate the regarding candidate though (s)he followed the whole procedure.

In Table 4.2, the observed annotation combinations are presented for base set 1. In table, 1 means the judge tagged the candidate as MWE, 0 means the judge tagged the candidate as non-MWE. For example, "1-0-0-NONE" means that the first judge tagged the candidate as MWE, following two judges tagged as non-MWE and the last judge did not tag the candidate.

In order to tag the candidates, we counted the number of ones for each candidate. If the number of ones for a candidate is greater than or equals to three, then it is accepted as MWE. If this number equals to two, then the candidate is re-assessed by another judge, such cases are represented by "X" in tag column in Table 4.2. The final decision statistics for such situations are given in Table 4.3.

In Table 4.4 the observed annotation combinations are presented for base set 2 where 3 judges decided the tag for candidates. For each candidate in base set 2, if the number of MWE decisions is greater than or equals two, the candidate is accepted as MWE. For the other cases, it is tagged as non-MWE.

Cases	Total count	Tag	Cases	Total count	Tag
0-0-0-0	783	non-MWE	1-1-0-0	13	X
0-0-0-1	84	non-MWE	1-1-0-1	49	MWE
0-0-0-NONE	3	non-MWE	1-1-1-0	40	MWE
0-0-1-0	13	non-MWE	1-1-1-1	884	MWE
0-0-1-1	18	X	1-1-1-NONE	18	MWE
0-0-NONE-0	5	non-MWE	1-1-NONE-1	2	MWE
0-0-NONE-1	2	non-MWE	1-1-NONE-NONE	10	X
0-1-0-0	40	non-MWE	1-NONE-0-0	2	non-MWE
0-1-0-1	37	X	1-NONE-0-1	3	X
0-1-1-0	2	X	1-NONE-1-1	8	MWE
0-1-1-1	27	MWE	NONE-0-0-0	5	non-MWE
0-NONE-0-0	14	non-MWE	NONE-0-0-1	3	non-MWE
0-NONE-0-1	14	non-MWE	NONE-0-0-NONE	1	non-MWE
0-NONE-0-NONE	4	non-MWE	NONE-0-1-1	3	X
0-NONE-1-0	2	non-MWE	NONE-1-0-0	3	non-MWE
0-NONE-NONE-1	1	non-MWE	NONE-1-1-1	3	MWE
1-0-0-0	32	non-MWE	NONE-1-NONE-1	1	X
1-0-0-1	14	X	NONE-NONE-0-0	1	non-MWE
1-0-0-NONE	1	non-MWE	NONE-NONE-0-1	1	non-MWE
1-0-1-0	9	X	NONE-NONE-1-1	1	X
1-0-1-1	72	MWE	1-0-NONE-0	1	non-MWE

Table 4.2: Observed cases in annotation of BS1

EXCEPTIONAL CASES	TOTAL COUNT	MWE	non-MWE
0-0-1-1	18	15	3
0-1-0-1	37	32	5
0-1-1-0	2	2	0
1-0-0-1	14	11	3
1-0-1-0	9	4	5
1-1-0-0	13	10	3
1-1-NONE-NONE	10	10	0
1-NONE-0-1	3	3	0
NONE-0-1-1	3	3	0
NONE-1-NONE-1	1	1	0
NONE-NONE-1-1	1	0	1

Table 4.3: Exceptional cases in annotation of BS1

CASES	TOTAL COUNT	TAG
0-0-0	437	non-MWE
0-0-1	23	non-MWE
0-1-0	28	non-MWE
0-1-1	25	MWE
0-NONE-0	0	non-MWE
1-0-0	31	non-MWE
1-0-1	62	MWE
1-1-0	63	MWE
1-1-1	741	MWE

Table 4.4: Observed cases in annotation of BS2

For example, as it can be seen in the right part of the Table 4.2, there are 49 cases in BS1 that 3 judges tagged the bigram candidates as MWE and 1 judge tagged as non-MWE. Because of the majority, it is tagged as MWE in the final set. Table 4.3 includes situations such as 2 judges give the same decision for a candidate, and the others tagged different than these 2 judges. For instance, 18 MWE candidates are tagged as MWE by 2 judges and the same candidates are

tagged as non-MWE by the other 2 judge. In this case, the fifth judge decided for that bigram and among 18 candidates, 15 bigrams are tagged as MWE and 3 of them are tagged as non-MWE.

	Number of Bigrams annotated as MWE	Number of Bigrams annotated as non-MWE	Total
Base Set 1	1194 ($\sim 53.56\%$)	1035 ($\sim 46.43\%$)	2229 (100%)
Base Set 2	891 ($\sim 63.14\%$)	520 ($\sim 36.85\%$)	1411 (100%)

Table 4.5: Annotated base sets

Table 4.5 gives the proportions and numbers (#) of candidates that are annotated as MWE and non-MWE in the resulting base sets. For example, in base set 2 there exists 891 candidates ($\sim 63.14\%$ of the entire bigram set) annotated as MWE and 520 candidates ($\sim 36.85\%$ of total) annotated as non-MWE.

Inter-rater agreement among the annotators is measured by Fleiss Kappa [23] which is a statistical measure for assessing the reliability of agreement between a fixed number of raters when they assign categorical ratings to a number of items. If the raters are in complete agreement then the kappa value k is 1 and if k is 0 there is no agreement among the raters. The results are, ~ 0.728 and ~ 0.767 , respectively for base set 1 and base set 2.

In Tables 4.6 and 4.7, the statistics on agreement among the judges are given. For instance, there are 1667 and 415 bigrams in base set 1 that 4 judges and 3 judges have the same decision, respectively. Furthermore, for 111 bigrams there is a different case such as two judges tagged the bigram as MWE and the other ones tagged the same bigram as non MWE. As it can be seen from the Tables 4.6 and 4.7 a great amount of bigrams are annotated with the same tag (MWE or non-MWE) by the majority of judges.

Number of Judges	# Bigrams
4	1667
3	415
2	36
2-2	111
Total	2229

Table 4.6: Agreement statistics for BS 1

Number of Judges	# Bigrams
3	1178
2	233
Total	1411

Table 4.7: Agreement statistics for BS 2

4.3 Evaluation Measures

In this thesis, we evaluated the performance of MWE identification methods by precision, recall and F-measures. In information retrieval, precision can be considered as the fraction of retrieved documents that are relevant to the given query, and recall is the fraction of relevant documents that are retrieved for the given query. F-measure is the combination of the precision and the recall values. Simply, it is the harmonic mean of precision and recall that is calculated as follows;

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

In MWE extraction, precision may be defined as the fraction of true MWEs retrieved from the base set and recall is the fraction of extracted true MWEs retrieved from the entire set of true MWEs in the base set [19].

A	B	"A B"	TAG	Jaccard	SET SIZE	P	R	F
şunları	kaydetti	şunları kaydetti	0	31.04	1	0	0	0
semsettin	gunaltay	semsettin gunaltay	1	28.5454	2	0.5	0.0008	0.0016
dayanıklı	tüketim	dayanıklı tüketim	0	3.3197	3	0.3333	0.0008	0,0016
zülfü	livaneli	zülfü livaneli	1	3.3195	4	0.5	0.0016	0,0033
gümbür	gümbür	gümbür gümbür	1	2.9689	5	0.6	0.0025	0.0049

Table 4.8: Sample results after utilizing jaccard method for BS1

In Table 4.8, P, R and F refers precision, recall and F-measure, respectively. It represents the top first 5 candidates sorted according to their values after jaccard measure is applied. For example, when set size is equals to 3 precision, recall and f-measure are measured 0.333, 0.008 and 0.0016 , respectively. In this case all measures are calculated considering only the first 3 candidates. In the first 3 candidate there is only 1 true MWE. This is why precision is equals to $\frac{0+1+0}{\text{set size}}$ and recall is equals to $\frac{0+1+0}{\text{total number of MWEs in BS1}}$.

A	B	"A B"	TAG	SET SIZE	Precision	Recall	F-measure
new	york	new*york	1	1	1	0.0008	0.0016
hong	kong	hong*kong	1	2	1	0.0016	0.0033
los	angeles	los*angeles	1	3	1	0.0025	0.005
know	how	know*how	1	4	1	0.0033	0.0066
las	vegas	las*vegas	1	5	1	0.0041	0.0083

Table 4.9: Sample results from the calculation of best case

Table 4.9 represents the first 5 candidates of base set 1 in the sorted list according to their tags. Best case represents the sorted list in which MWEs hold the top most ranks and non-MWEs are at the bottom of the list.

Chapter 5

Experimental Results

In this thesis, after utilizing each frequency based association measure listed in Table 3.1, the base sets (BS 1 and BS 2) are sorted according to the association values of the candidates. We calculated the precision, recall and F-measure for first N candidates of the sorted base sets where N is varied from 1 to total number of candidates in set to obtain the curves. Then, we calculated the average values of precision (PRECISION_AVG), recall(RECALL_AVG) and F-measure(F_AVG) values and the area under F curve(F_AREA) of every measure. In Tables 5.1 and 5.2, sorted lists according to F_AVG of 20 association measures can be seen for BS 1 and BS 2, respectively.

#	Measures	F_AREA	F_AVG*	PRECISION_AVG	RECALL_AVG
1	BEST	1593.0310	0.7146	0.8707	0.7317
2	LFMD	1271.9823	0.5706	0.6696	0.5986
3	R cost	1266.6687	0.56823	0.6682	0.5964
4	CP	1266.5252	0.5682	0.6662	0.5955
5	MD	1266.1319	0.5680	0.6679	0.5963
6	DK	1266.1319	0.5680	0.6679	0.5963
7	SSS	1265.3916	0.5676	0.6841	0.5941
8	Jaccard	1264.8019	0.5674	0.6785	0.5943
9	SK	1261.9638	0.5661	0.6599	0.5948
10	NE	1256.1829	0.5635	0.6642	0.5911
11	Simpson	1256.0658	0.5635	0.6543	0.5926
12	S cost	1255.7905	0.5633	0.6543	0.5924
13	FK	1253.9225	0.5625	0.6775	0.5889
14	BB	1253.3080	0.5622	0.6634	0.5897
15	RCP	1214.4310	0.5448	0.6334	0.5742
16	JP	1207.9420	0.5419	0.6281	0.5750
17	PMI	1182.4557	0.5304	0.6207	0.5604
18	Mountford	1159.1292	0.5200	0.6232	0.5470
19	U cost	1093.4338	0.4905	0.5721	0.5212
20	Fager	1041.6869	0.4673	0.5305	0.5028
21	PS	1010.4959	0.4533	0.4959	0.4913

Table 5.1: Test results of the association measures for Base set 1 sorted according to F_AVG

#	Measures	F_AREA	F_AVG*	PRECISION_AVG	RECALL_AVG
1	BEST CASE	1003.4654	0.7111	0.9216	0.6846
2	LFMD	825.4159	0.5849	0.7488	0.5707
3	JP	820.6075	0.5815	0.7414	0.5684
4	CP	820.119	0.5812	0.7434	0.5674
5	Jaccard	814.0536	0.5769	0.7370	0.5641
6	MD	813.3099	0.5764	0.7355	0.5631
7	DK	813.3099	0.5764	0.7355	0.5631
8	R cost	812.8324	0.5760	0.7347	0.5629
9	NE	812.7905	0.5760	0.7325	0.5633
10	BB	811.3062	0.5749	0.7308	0.5625
11	SSS	810.8749	0.5746	0.7302	0.5620
12	SK	805.6121	0.5709	0.7337	0.5569
13	FK	803.7424	0.5696	0.7283	0.5566
14	Simpson	801.5928	0.5681	0.7305	0.5543
15	S cost	801.5909	0.5681	0.7305	0.5543
16	PMI	784.3281	0.5558	0.7121	0.5438
17	RCP	772.3949	0.5474	0.7030	0.5364
18	Mountford	752.3191	0.5331	0.6915	0.5203
19	Fager	745.7406	0.5285	0.6731	0.5197
20	U cost	740.6867	0.5249	0.6555	0.5174
21	PS	728.6997	0.5164	0.6469	0.5094

Table 5.2: Test results of the association measures for Base Set 2 sorted according to F_AVG

In Tables 5.1 and 5.2, it is observed that LFMD measure produces the maximum values for both base sets. In addition, the scores of the second maximum values for all measures are significantly lower than the scores of LFMD. The minimum evaluation values are obtained for PS measure in both base sets.

#	Measure	Formula
1	BEST	
2	LFMD	$\log \frac{P(w_1 w_2)^2}{P(w_1)P(w_2)} + \log P(w_1 w_2)$
3	R cost	$\log \left(1 + \frac{f(w_1 w_2)}{f((w_1 w_2) + f(w_1 \bar{w}_2))} \right) \cdot \log \left(1 + \frac{f(w_1 w_2)}{f((w_1 w_2) + f(\bar{w}_1 w_2))} \right)$
4	CP	$P(w_2 w_1)$
5	MD	$\log \frac{P(w_1 w_2)^2}{P(w_1)P(w_2)}$
6	DK	$\frac{f(w_1 w_2)}{\sqrt{(f(w_1 w_2) + f(w_1 \bar{w}_2)) \cdot (f(w_1 w_2) + f(\bar{w}_1 w_2))}}$
7	SSS	$\frac{f(w_1 w_2)}{f(w_1 w_2) + 2(f(w_1 \bar{w}_2) + f(\bar{w}_1 w_2))}$
8	Jaccard	$\frac{f(w_1 w_2)}{f(w_1 w_2) + f(w_1 \bar{w}_2) + f(\bar{w}_1 w_2)}$
9	SK	$\frac{f(w_1 w_2)}{f(w_1 w_2) + 2(f(w_1 \bar{w}_2) + f(\bar{w}_1 w_2))}$
10	NE	$\frac{f(w_1) + f(w_2)}{2f(w_1 w_2)}$
11	Simpson	$\frac{f(w_1 w_2)}{\min(f(w_1 w_2) + f(w_1 \bar{w}_2), f(w_1 w_2) + f(\bar{w}_1 w_2))}$
12	S cost	$\log \left(1 + \frac{\min(f(w_1 \bar{w}_2), f(\bar{w}_1 w_2))}{f(w_1 w_2) + 1} \right)^{-\frac{1}{2}}$
13	FK	$\frac{f(w_1 \bar{w}_2) + f(\bar{w}_1 w_2)}{f(w_1 w_2)}$
14	BB	$\frac{1}{\max(f(w_1 w_2) + f(w_1 \bar{w}_2), f(w_1 w_2) + f(\bar{w}_1 w_2))}$
15	RCP	$P(w_1 w_2)$
16	JP	$P(w_1 w_2)$
17	PMI	$\log \frac{P(w_1 w_2)}{P(w_1)P(w_2)}$
18	Mountford	$\frac{2f(w_1 \bar{w}_2)f(\bar{w}_1 w_2) + f(w_1 w_2)f(w_1 \bar{w}_2) + f(w_1 w_2)f(\bar{w}_1 w_2)}{2f(w_1 w_2)}$
19	U cost	$\log \left(1 + \frac{\min(f(w_1 \bar{w}_2), f(\bar{w}_1 w_2)) + f(w_1 w_2)}{\max(f(w_1 \bar{w}_2), f(\bar{w}_1 w_2)) + f(w_1 w_2)} \right)$
20	Fager	$\frac{1}{\sqrt{(f(w_1 w_2) + (f(w_1 \bar{w}_2))(f(w_1 w_2) + f(\bar{w}_1 w_2)))} - \frac{1}{2} \max(f(w_1 \bar{w}_2), f(\bar{w}_1 w_2))}$
21	PS	$P(w_1 w_2) - P(w_1)P(w_2)$

Table 5.3: Sorted list of the association measures according to their success for BS 1

#	Measure	Formula
1	BEST	
2	LFMD	$\log \frac{P(w_1 w_2)^2}{P(w_1)P(w_2)} + \log P(w_1 w_2)$
3	JP	$P(w_1 w_2)$
4	CP	$P(w_2 w_1)$
5	JACCARD	$\frac{f(w_1 w_2)}{f(w_1 w_2) + f(w_1 \bar{w}_2) + f(\bar{w}_1 w_2)}$
6	MD	$\log \frac{P(w_1 w_2)^2}{P(w_1)P(w_2)}$
7	DK	$\frac{f(w_1 w_2)}{\sqrt{(f(w_1 w_2) + f(w_1 \bar{w}_2)) \cdot (f(w_1 w_2) + f(\bar{w}_1 w_2))}}$
8	R cost	$\log \left(1 + \frac{f(w_1 w_2)}{f((w_1 w_2) + f(w_1 \bar{w}_2))} \right) \cdot \log \left(1 + \frac{f(w_1 w_2)}{f((w_1 w_2) + f(\bar{w}_1 w_2))} \right)$
9	NE	$\frac{2f(w_1 w_2)}{f(w_1) + f(w_2)}$
10	BB	$\frac{f(w_1 w_2)}{\max(f(w_1 w_2) + f(w_1 \bar{w}_2), f(w_1 w_2) + f(\bar{w}_1 w_2))}$
11	SSS	$\frac{f(w_1 w_2)}{f(w_1 w_2) + 2(f(w_1 \bar{w}_2) + f(\bar{w}_1 w_2))}$
12	SK	$\frac{f(w_1 w_2)}{f(w_1 w_2) + 2(f(w_1 \bar{w}_2) + f(\bar{w}_1 w_2))}$
13	FK	$\frac{f(w_1 w_2)}{f(w_1 \bar{w}_2) + f(\bar{w}_1 w_2)}$
14	Simpson	$\frac{f(w_1 w_2)}{\min(f(w_1 w_2) + f(w_1 \bar{w}_2), f(w_1 w_2) + f(\bar{w}_1 w_2))}$
15	S cost	$\log \left(1 + \frac{\min(f(w_1 \bar{w}_2), f(\bar{w}_1 w_2))}{f(w_1 w_2) + 1} \right)^{-\frac{1}{2}}$
16	PMI	$\log \frac{P(w_1 w_2)}{P(w_1)P(w_2)}$
17	RCP	$P(w_1 w_2)$
18	Mountford	$\frac{2f(w_1 \bar{w}_2)f(\bar{w}_1 w_2) + f(w_1 w_2)f(w_1 \bar{w}_2) + f(w_1 w_2)f(\bar{w}_1 w_2)}{2f(w_1 w_2)}$
19	Fager	$\frac{1}{\sqrt{(f(w_1 w_2) + (f(w_1 \bar{w}_2))(f(w_1 w_2) + f(\bar{w}_1 w_2)))} - \frac{1}{2} \max(f(w_1 \bar{w}_2), f(\bar{w}_1 w_2))}$
20	U cost	$\log \left(1 + \frac{\min(f(w_1 \bar{w}_2), f(\bar{w}_1 w_2)) + f(w_1 w_2)}{\max(f(w_1 \bar{w}_2), f(\bar{w}_1 w_2)) + f(w_1 w_2)} \right)$
21	PS	$P(w_1 w_2) - P(w_1)P(w_2)$

Table 5.4: Sorted list of the association measures according to their success for BS 2

In Tables 5.3 and 5.4, the measures and regarding formulas for base sets are sorted in decreasing order according to F_AVERAGE is given. Two important remarks from Tables 5.3 and 5.4 are ;

1. LFMD, MD and CP are in 5 best performing measures for both base sets.
2. The 5 best performing measures involve commonly operand $f(w_1w_2)$, $f(w_1)$ and $f(w_2)$. It is observed that in most of the measures in this group, $f(w_1w_2)$ is divided by the multiplication of $f(w_1)$ and $f(w_2)$.

Following, we plotted the graphs of precision, recall and F-measure values of all competing measures. In Figures 5.1 and 5.2, F-measure curves and in Appendix A, precision and recall graphs are presented. In this graphs, horizontal axis represents the number of MWE candidates in base sets and vertical axis represents the values of the evaluation measures. Observing the F measure curves for BS 1 and BS 2, it can be stated that for both data sets Fager, PS and U cost measures are not successful in ranking the bigram candidates. LFMD and DK measures generate higher F-values for BS1 and LFMD and CP measures perform better for BS2.

The performances of association measures may vary when the frequencies are obtained from different sources. In this thesis, in order to examine whether the web may be used as a source in MWE identification task or not, the methods are employed with frequency values that are extracted from Leipzig corpus [7] and the web, individually. We examined that Leipzig corpus include 1245 ($\sim 55.85\%$) number of candidates of BS1. This reduced set of BS1 is composed of 733($\sim 58.87\%$) MWEs and 512($\sim 41.124\%$) non-MWEs. The first part of the Table 5.5 shows the performance results that are obtained from the corpus and the second part presents the results for web frequencies. Performance results in Table 5.5 show that:

1. LFMD is the best performing measure when the frequencies are obtained from Leipzig corpus and CP is the best performing measure when the frequencies are obtained from web.
2. MD and R cost methods are in the first 5 best performing measures for both sources.
3. The best performing association measures, LFMD and CP respectively for

the Leipzig corpus and the web, generate almost same average F-values, precision and recall.



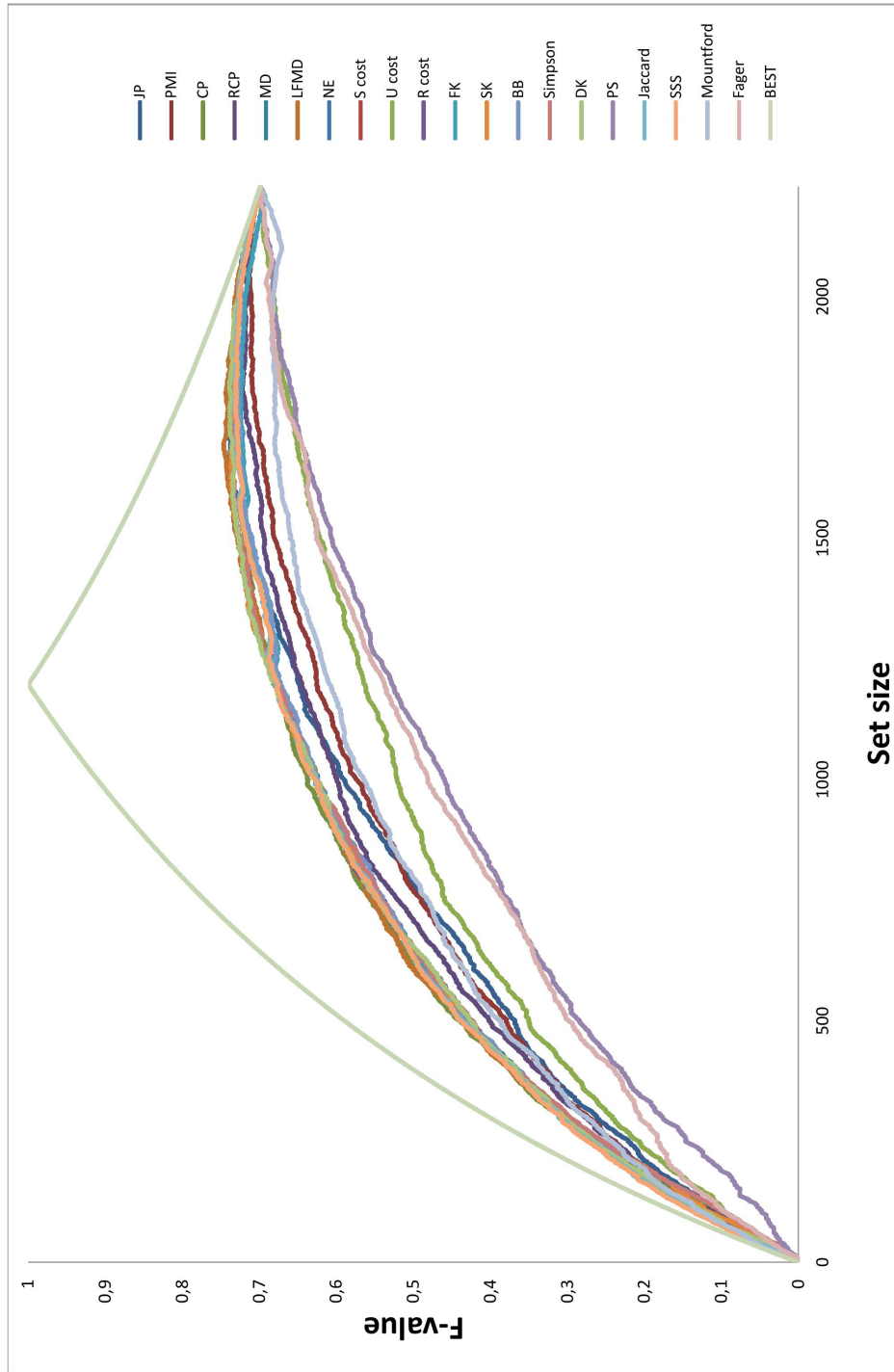


Figure 5.1: F-measure graph for BS1

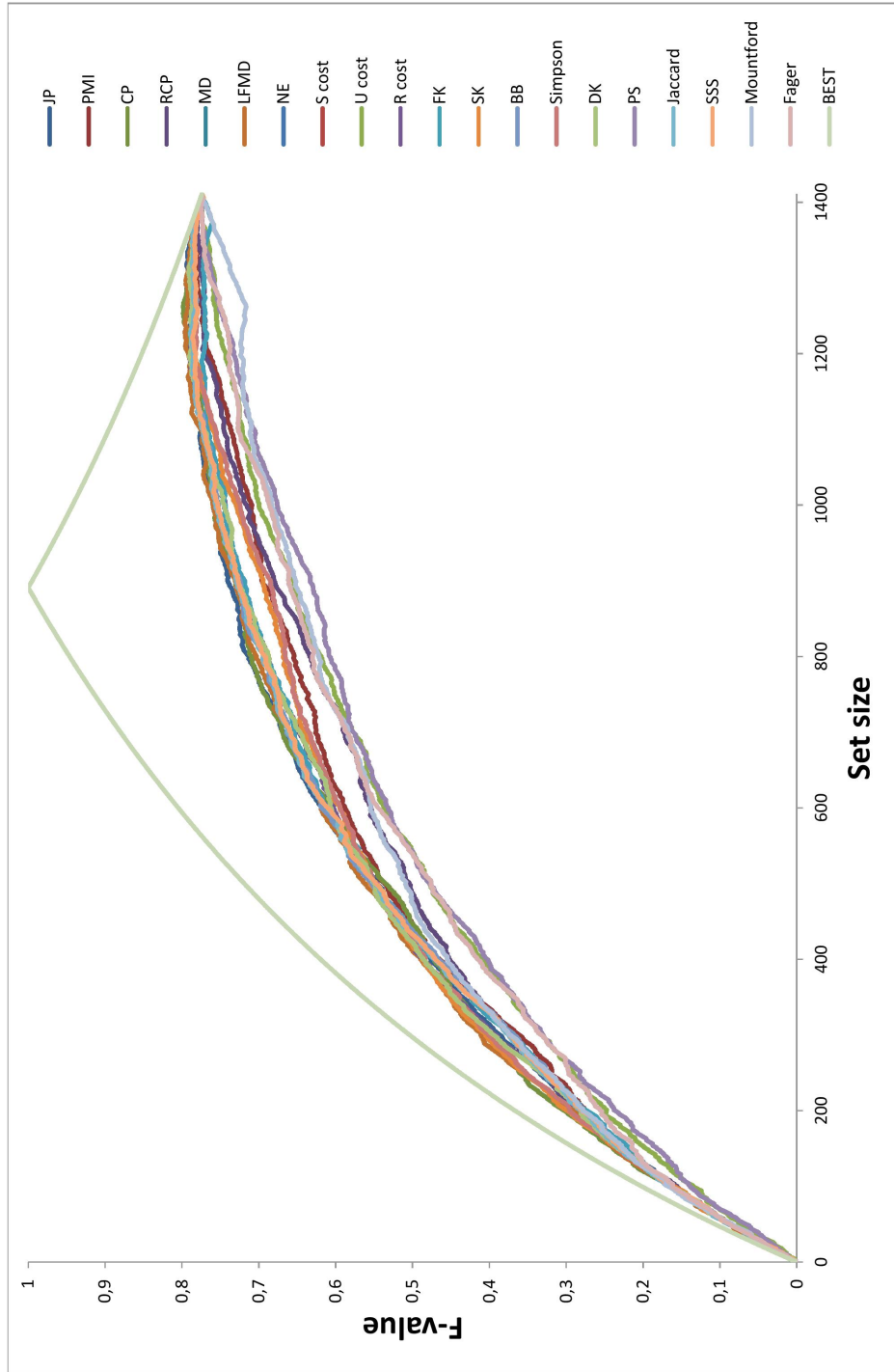


Figure 5.2: F-measure for BS2

Leipzig Corpus		Web(Google)							
	F_AREA	F_AVG	Precision_AVG	Recall_AVG		F_AREA	F_AVG	Precision_AVG	Recall_AVG
BEST	889,354	0,714	0,900	0,706		889,354	0,714	0,900	0,706
LFMD	705,036	0,566	0,713	0,566	CP	705,452	0,566	0,693	0,5700
MD	676,585	0,543	0,652	0,553	Jaccard	705,289	0,566	0,707	0,569
DK	676,578	0,543	0,652	0,533	SSS	704,08	0,565	0,707	0,567
R cost	676,497	0,543	0,652	0,553	R cost	703,878	0,565	0,693	0,569
BB	675,643	0,542	0,652	0,552	MD	703,444	0,565	0,694	0,569
NE	675,617	0,542	0,651	0,552	DK	703,444	0,565	0,694	0,569
FK	675,617	0,542	0,651	0,552	LFMD	701,539	0,563	0,69	0,567
Jaccard	675,617	0,542	0,651	0,552	NE	700,315	0,562	0,691	0,565
SSS	675,617	0,542	0,651	0,552	BB	699,929	0,562	0,691	0,565
CP	668,785	0,537	0,650	0,546	FK	699,667	0,561	0,706	0,564
SK	668,06	0,536	0,645	0,546	SK	698,614	0,561	0,682	0,566
Simpson	663,42	0,532	0,642	0,542	Simpson	693,408	0,556	0,675	0,562
S cost	663,03	0,532	0,642	0,542	S cost	693,322	0,556	0,675	0,562
Mountford	660,954	0,530	0,634	0,542	PMI	680,927	0,546	0,674	0,553
PMI	656,433	0,527	0,628	0,54	RCP	670,328	0,538	0,653	0,544
RCP	654,468	0,525	0,631	0,534	Mountford	667,386	0,536	0,672	0,539
U cost	653,684	0,525	0,643	0,533	U cost	650,449	0,522	0,645	0,528
Fager	649,953	0,522	0,622	0,534	Fager	650,192	0,522	0,627	0,531
PS	643,305	0,516	0,616	0,529	JP	642,040	0,515	0,623	0,526
JP	610,061	0,490	0,600	0,500	PS	637,423	0,511	0,602	0,523

Table 5.5: Sorted list of the association measures according to F_AVG

Chapter 6

Conclusion

Recurrent combinations of words in natural languages compose MWEs. In identification of MWEs, statistical methods that are known as association measures are widely used. Such statistical methods mainly consider the occurrence frequency property in MWE extraction. Therefore, a data source that is suitable for measuring frequency of words or word combinations is required. The corpus and the data source that is used to construct the corpus affect the performance of the MWE extraction methods.

In this thesis, we analyze the extraction performance of the frequency based MWE extraction methods when the frequency is obtained from web sources by the use of search engines. The term MWE in this thesis is limited to bigrams and the MWEs are annotated by 3 to 4 human judges according to some predefined rules. We construct 2 base sets by utilizing 3-6 different corpora. The first base set is built by statistical methods and the second base set is built by selecting candidates based on idioms dictionary.

A set of 20 renowned frequency based statistical methods is applied in the thesis. In order to obtain the occurrence frequency, the World Wide Web is used as the first data source. The mostly used search engine, Google, is preferred to retrieve the occurrence frequencies; the page counts; of the candidate MWEs from the internet. It is observed that in base sets the measure LFMD produced

the highest scores in all evaluation measures when the frequencies are obtained from the web. In order to compare the performance of web when used as a data source in MWE identification task, we repeated the same tests with an alternating well-known data source, Leipzig corpus. Occurrence frequency is accepted to be the page counts retrieved from Google and frequency in Leipzig corpus in our web based and corpus based study, respectively. It is observed that the average best F-value for web and the corpus is 0.566 with the method CP and LFMD, respectively. As a result, it can be stated that when different sources of frequency are used, though the best performing methods differ, the similar performance scores may be obtained.

As a future work, some of the occurrence frequency based statistical methods can be merged and the success can be measured or a different search engine can be utilized in order to obtain frequencies of candidate MWEs instead of Google.

BIBLIOGRAPHY

- [1] Firth, J.R.: Modes of Meaning. Papers in Linguistic 1934-51. Oxford University Press (1967)
- [2] Sinclair, J.M.: Corpus, Concordance, Collocation. Oxford University Press, Oxford (1991)
- [3] Hoey, M.: Patterns of Lexis in Text. Oxford University Press (1991)
- [4] Bisht, R. K., Dhama H.S. and Neeraj Tiwari, N.: An evaluation of different statistical techniques of collocation extraction using a probability measure to word combinations. Journal of Quantitative Linguistics, Vol.13,161-175 (2006)
- [5] Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press. England (1999)
- [6] Jianfang, L., Sheng, L., Yuhua, C. : Collocation Extraction Using Web Feedback Data. Chinese Journal of Electronics, Vol 18, No.2, (Apr.2009)
- [7] Quasthoff U, Richter M, Biemann C. Corpus portal for search in monolingual corpora. Proceedings of the fifth international conference on language resources and evaluation. (2006)
- [8] Tür, G., Hakkani-Tür, D and Oflazer K. : A statistical Information Extraction System for Turkish. Natural Language Engineering. Vol 9 No.2, 181-210 (2003)

- [9] Kumova Metin, S., Karaođlan, B., Collocation Extraction in Turkish Texts Using Statistical Methods, 7th International Conference on Natural Language Processing (LNCS-ISI) IceTAL 2010, Reykjavik, Iceland (2010)
- [10] Web site link: <http://www.atasozuarsivi.com/deyimler-sozlugu.html>
- [11] Web site link: <https://tr.wiktionary.org/wiki/Kategori:Deyim>
- [12] Pecina, P., A Machine Learning Approach to Multiword Expression Extraction, Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions (2008)
- [13] Kumova Metin, S., Kışla, T., Karaođlan B., Named Entity Recognition in Turkish Using Association Measures, *Advanced Computing: An International Journal (ACIJ)*, Vol.3, No.4 (July 2012)
- [14] Church, K. W., Hanks, P.: Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, Vol. 16 No.1, 22-29 (1990)
- [15] Smadja, F.A.: Retrieving Collocations from Text: Xtract. *Computational Linguistics*, Vol. 19 No. 1, 143-177 (1993)
- [16] Oflazer, K., Çetinođlu, Ö., Say, B.: Integrating morphology with multi-word expression processing in Turkish. Proceedings of the Workshop on Multiword Expressions: Integrating Processing. p. 6471. (2004)
- [17] Pecina P.: Lexical association measures and collocation extraction. *Lang Resour Eval.* 2010;44(1-2):13758.
- [18] Bouma G.: Collocation Extraction beyond the Independence Assumption. *Proc ACL 2010 Conf Short Pap.*;10914 (2010)
- [19] Kumova Metin, S.: Neighbour Unpredictability Measure in Multiword Expression Extraction, *International Journal of Computer Systems Science and Engineering (SCI-Expanded)* (2016)
- [20] Kim S, Yoon J, Song M.: Automatic Extraction of Collocations From Korean Text. *Comput Hum*;35:27397. (2001)

- [21] Li W, Lu Q, Liu J.: Chinese typed collocation extraction using corpus-based syntactic collocation patterns. IEEE NLP-KE 2007 - Proceedings of International Conference on Natural Language Processing and Knowledge Engineering.p. 24855. (2007)
- [22] The FR corpus is part of ECI Multilingual Corpus I distributed by ELSNET.(<http://www.elsnet.org/eci.html>) (1994)
- [23] Fleiss,J.L.: "Measuring nominal scale agreement among many raters." Psychological Bulletin 378-382 (1971)
- [24] Ramisch, C., Villavicencio, A., Boitet, C.: mwetoolkit: a Framework for Multiword Expression Identification, LREC , (2010)
- [25] Wu, S., Witten H., I., Franken, M.: Utilizing lexical data from a Web-derived corpus to expand productive collocation knowledge. Recall,22, pp 83-102 (2010)
- [26] Tsvetkov, Y., Wintner, S.: Identification of Multi-word Expressions by Combining Multiple Linguistic Information Sources. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 836845, Edinburgh, Scotland, UK, July 2731 (2011)
- [27] Sarıkaş, F.: Problems in Translating Collocations. Elektronik Sosyal Bilimler Dergisi www.e-sosder.com C.5 S.17 (33-40) (2006)
- [28] Evert S, Krenn B.: Methods for the qualitative evaluation of lexical association measures. Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL 01. p. 18895. (2001)
- [29] Evert S.: Significance tests for the evaluation of ranking methods. Proc 20th Int Conf.945 es. (2004)
- [30] Antunes S., Mendes A.: An Evaluation of the Role of Statistical Measures and Frequency for MWE Identification, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) (2014)
- [31] Dinçer T.: Trkçe için istatistiksel bir bilgi geri-getirim sistemi, Doktora Tezi, U.B.E.,Ege niversitesi (2004)

- [32] Can, F, Kocberber, S, Baloglu, O, Kardas, S, .calan, HC & Uyar, E: New event detecDon and topic tracking in Turkish, Journal of the American Society for InformaDon Science and Technology, vol 61, no. 4, pp. 802-819, (2010)
- [33] Say, Bilge, Deniz Zeyrek, Kemal Oflazer and Umut Özge. "Development of a Corpus and a Treebank for Present-day Written Turkish", (Proceedings of the Eleventh International Conference of Turkish Linguistics, August, 2002) Imer, Kamile and Gürkan Doğan (eds), Current Research in Turkish Linguistics, pp.183-192, Eastern Mediterranean University Press (2004)

Appendix A



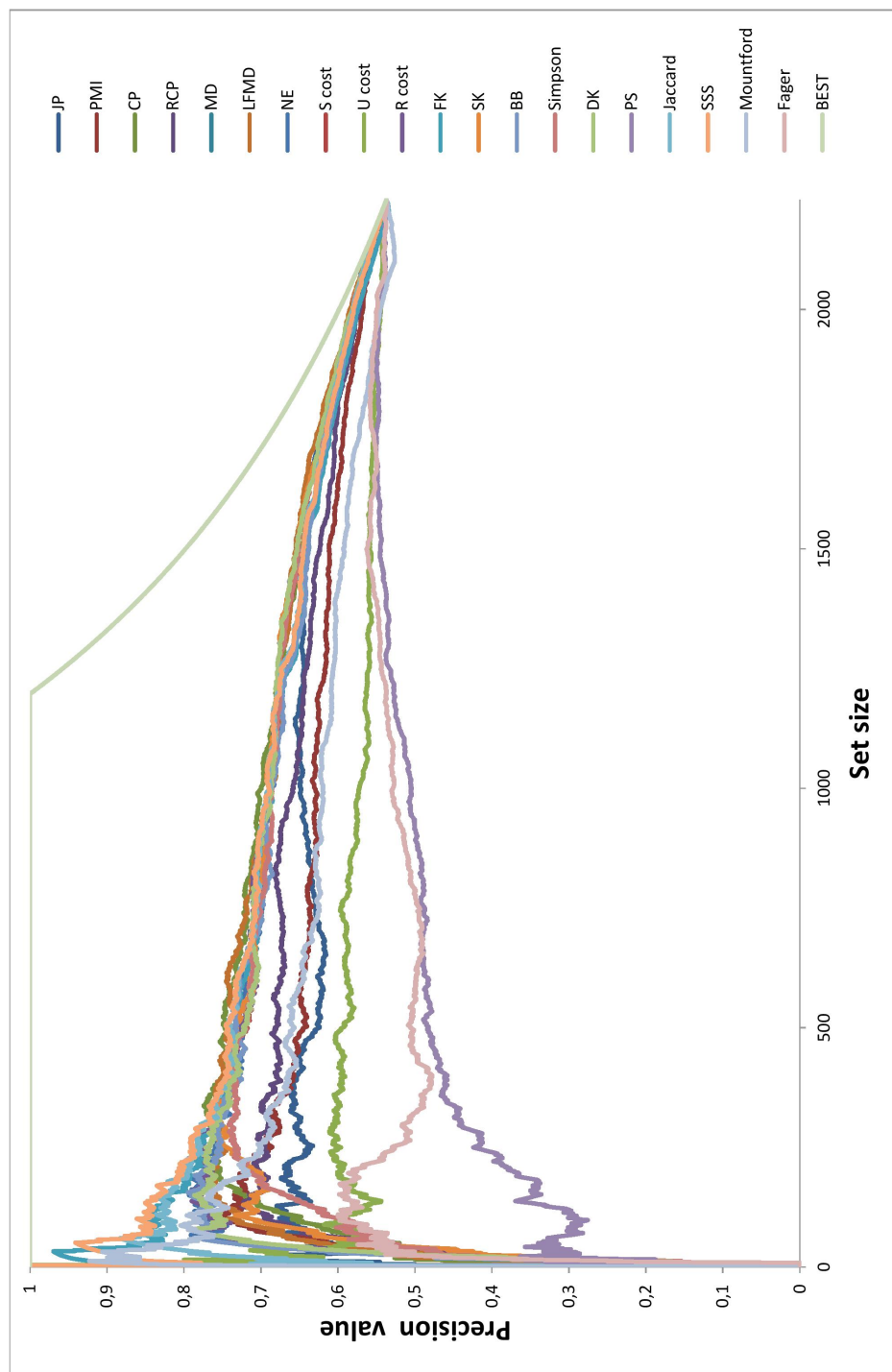


Figure A.1: Precision graph for BS1

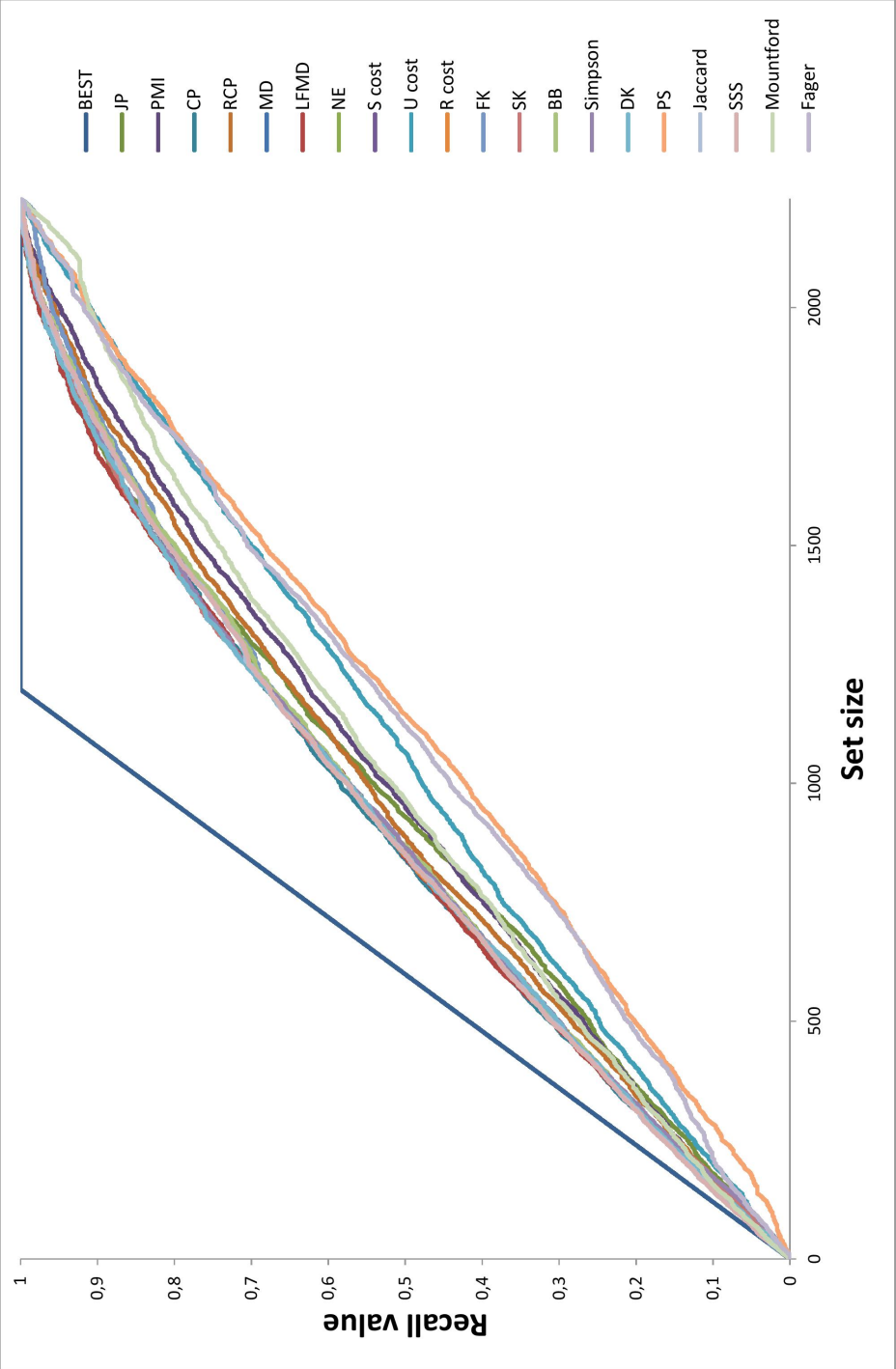


Figure A.2: Recall graph for BS1

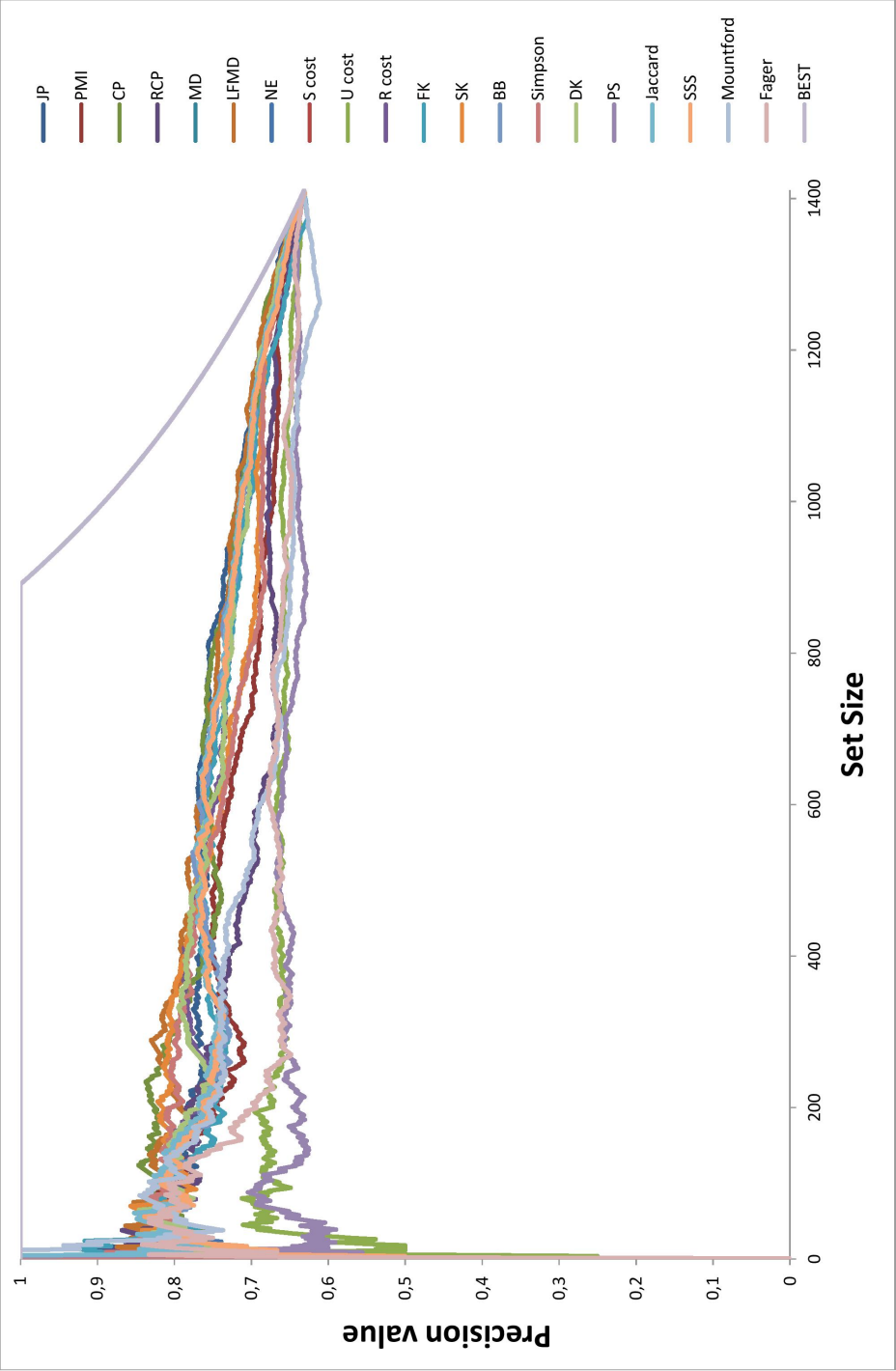


Figure A.3: Precision graph for BS2

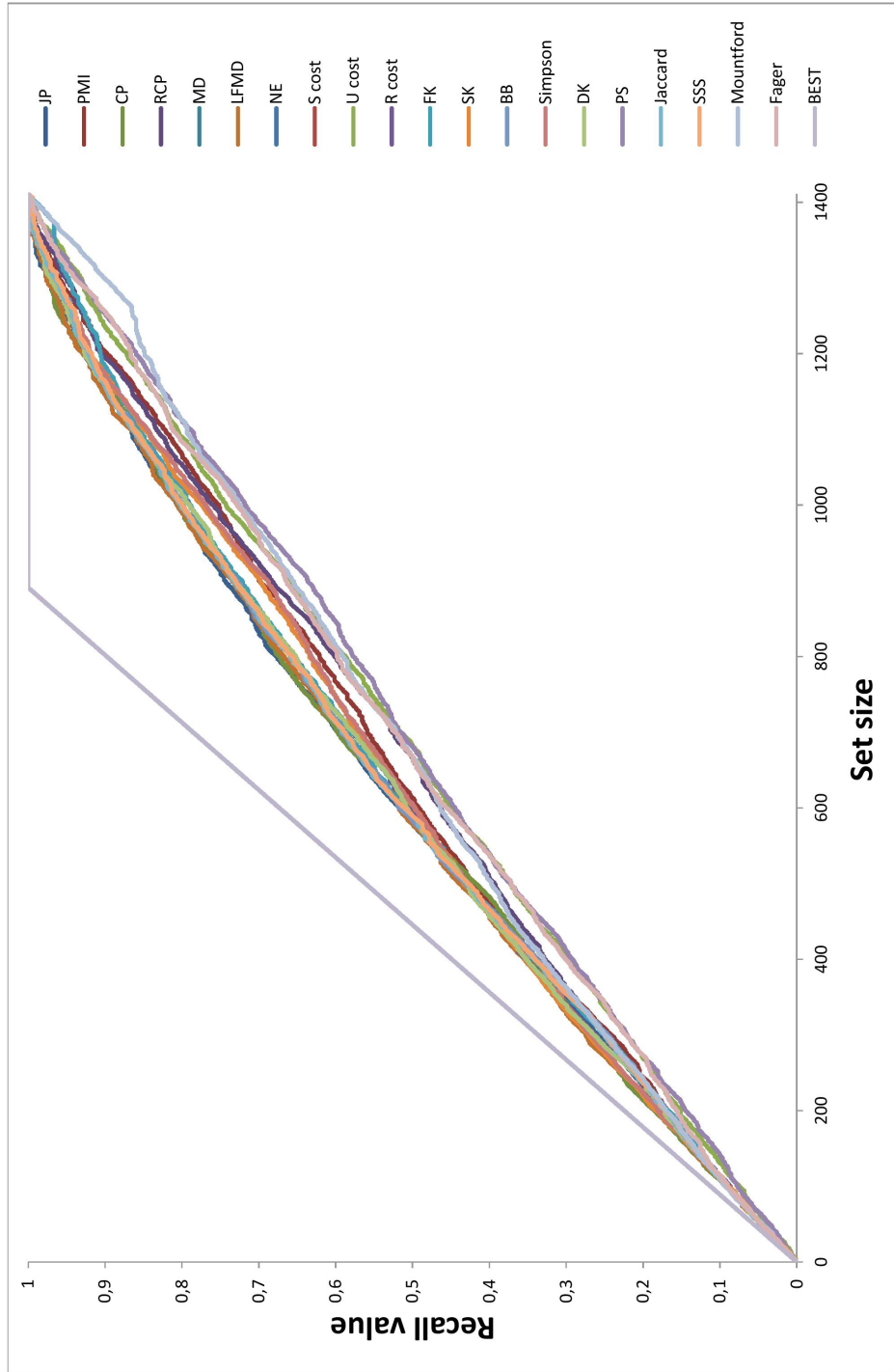


Figure A.4: Recall graph for BS2