PERSON NAME RECOGNITION IN TURKISH FINANCIAL TEXTS BY USING LOCAL GRAMMAR APPROACH

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÖZKAN BAYRAKTAR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

SEPTEMBER 2007

Approval of the Graduate School of Informatics

____

Prof. Dr. Nazife BAYKAL
Director

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

____

Assoc. Prof. Dr. Yasemin YARDIMCI
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.

________________________	______________________________
Prof. Dr. Nazife BAYKAL	Dr. Tuğba TAŞKAYA TEMİZEL
Co-Supervisor	Supervisor

Examining Committee Members

Prof. Dr. Deniz ZEYREK (METU, FLE) ____________________

Dr. Tuğba TAŞKAYA TEMİZEL (METU, IS) ____________________

Prof. Dr. Nazife BAYKAL (METU, IS) ____________________

Doç. Dr. Nihan KESİM ÇİÇEKLİ (METU, CENG) ____________________

Doç. Dr. Erkan MUMCUOĞLU (METU, IS) ____________________

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 wok.

Name, Last Name : Özkan BAYRAKTAR

Signature : ____________________

ABSTRACT

PERSON NAME RECOGNITION
IN TURKISH FINANCIAL TEXTS
BY USING LOCAL GRAMMAR APPROACH

Bayraktar, Özkan
M.Sc., Department of Information Systems
Supervisor: Dr. Tuğba Taşkaya Temizel
Co-Supervisor: Prof. Dr. Nazife Baykal

September 2007, 147 pages

Named entity recognition (NER) is the task of identifying the named entities (NEs) in the texts and classifying them into semantic categories such as person, organization, and place names and time, date, monetary, and percent expressions. NER has two principal aims: identification of NEs and classification of them into semantic categories. The local grammar (LG) approach has recently been shown to be superior to other NER techniques such as the probabilistic approach, the symbolic approach, and the hybrid approach in terms of being able to work with untagged corpora. The LG approach does not require using any dictionaries and gazetteers, which are lists of proper nouns (PNs) used in NER applications, unlike most of the other NER systems. As a consequence, it is able to recognize NEs in previously unseen texts at minimal costs. Most of the NER systems are costly due to manual rule compilation especially in large tagged corpora. They also require some semantic and syntactic analyses to be applied before pattern generation process, which can be avoided by using the LG approach.

In this thesis, we tried to acquire LGs for person names from a large untagged Turkish financial news corpus by using an approach successfully applied to a Reuter's financial English news corpus recently by H. N. Traboulsi. We explored its applicability to Turkish language by using frequency, collocation, and concordance analyses. In addition, we constructed a list of Turkish reporting verbs. It is an important part of this study because there is no major study about reporting verbs in Turkish.

Keywords: Local Grammar, Named Entity, Named Entity Recognition, Proper Noun, Turkish Reporting Verbs.

ÖZ

TÜRÇE FİNANS METİNLERİNDE YEREL DİLBİLGİSİ YAKLAŞIMI KULLANARAK KİŞİ İSMİ TANIMA

Bayraktar, Özkan
Yüksek Lisans, Bilişim Sistemleri Bölümü
Tez Danışmanı: Dr. Tuğba Taşkaya Temizel
Yardmcı Tez Danışmanı: Prof. Dr. Nazife Baykal

Eylül 2007, 147 sayfa

Varlık ismi tanıma varlıkların (örneğin, kişi ismi, organizasyon ismi, yer ismi, zaman deyimi, tarih deyimi ve yüzde deyimi) bulunup, anlamsal açıdan sınıflandırılmasıdır. Varlık ismi tanımanın iki temel amacı vardır. Birincisi varlıkların bulunup, tanınmasıdır. İkinci ise bu varlıkların sınıflandırılmasıdır. Son zamanlarda, yerel dilbilgisi yaklaşımı diğer varlık tanıma tekniklerine (örneğin, olasılıksal yaklaşım, sembolik yaklaşım ve hibrit yaklaşım) olan üstünlüğü işaretlenmemiş derlemler üzerinde çalışması açısından kanıtlanmıştır. Yerel dilbilgisi yaklaşımı varlık tanıma esnasında diğer varlık tanıma sistemlerinin aksine hiç bir genel sözlük, isim, organizasyon yada yer sözlüğüne ihtiyaç duymamaktadır. Sonuç olarak yerel dilbilgisi yaklaşımı daha önce görülmemiş metinlerde en az maliyet ile varlıkları tanımakta ve sınıflandırmaktadır. Diğer varlık tanıma sistemleri yerel dilbilgisi yaklaşımının aksine örüntü oluşturmadan önce bazı anlamsal ve yapısal analizlere ihtiyaç duymaktadır.

Biz bu tezde işaretlenmemiş büyük bir Türkçe finansal haber derleminde daha önce H.N. Traboulsi tarafından Reuters'ın bir finansal haber derlemine denenmiş ve başarılı olmuş yerel dilbilgisi yaklaşımı kullanarak kişi isimlerinin tanınmasında kullanabileceğimiz örüntüleri oluşturmaya çalıştık. Kısacası, yerel dilbilgisi yaklaşımının sıklık analizi, uygunluk analizi ve eşdizimlik analizi kullanarak Türkçe'ye uygulanabilirliğini araştırdık. Bunun yanı sıra, bu tezin önemli bir aşamasını oluşturan ve daha önce hiç çalışılmamış Türkçe rapor etme eylemlerinin bir listesinin oluşturulmasını da gerçekleştirdik.

Anahtar Kelimeler: Yerel Dilbilgisi, Varlık İsmi, Varlık İsmi Tanıma, Özel İsim, Türkçe Rapor Etme Eylemleri.

This thesis is dedicated to:

My Father and Mother
My Brother, Sisters, and Their Families
My Girlfriend (Minimik)

For their endless support,
For their Love …

ACKNOWLEDGEMENT

I would like to thank my supervisor Dr. Tuğba Taşkaya Temizel for her advice, encouragement and support she provided to me in supervising this thesis. I would also like to thank my co-supervisor Prof. Dr. Nazife Baykal and my examiner Prof. Dr. Deniz Zeyrek for their kind help and suggestions. Thanks go to Assist. Prof. Dr. Bilge Say and Dr. Hayssam N. Traboulsi for their suggestions and comments throughout this thesis.

I appreciate to Informatics Institute's administrative staff, Sibel Gülnar, Ali Kantar, Ayşe Ceylan, Nejla Işıklar, Nuray Tekin, Songül Acar, Yunus Canatan and Yaşar Sayın for their kindness. I am very grateful to Informatics Institute's research assistants who support me and review my thesis. Especially, I never forget my dear friend Selim Nar for his encouragement throughout this research.

And, very special thanks go to my girlfriend Pelin Beriat for her endless support, encouragement, and suggestions since the beginning of this research. I really appreciate to her for helping me during all the hard times. I also thank to her family for their encouragement during this study.

Finally, I express sincere appreciation to my parents Dursun and Safiye Bayraktar, my brother Osman and his family Selda, Berkant & Serhat, my sister Nermin and her family Cengiz, Gözde & Batuhan, and my sister Gülderen and her family Meltem & Erkin for their loves and endless support.

TABLE OF CONTENTS

ABSTRACT i
ÖZ vi
DEDICATION viii
ACKNOWLEDGEMENT ix
TABLE OF CONTENTS x
LIST OF TABLES xii
LIST OF FIGURES xiii
LIST OF ABBREVIATIONS xiv
CHAPTER
1 INTRODUCTION 1
1.1 DATA MINING (DM) 4
1.2 TEXT MINING (TM) 5
1.3 NAMED ENTITY RECOGNITION (NER) 6
1.4 THESIS OUTLINE 7
2 LITERATURE REVIEW 9
2.1 PROPER NOUN (PN) 9
2.2 NAMED ENTITY RECOGNITION (NER) 10
2.2.1 INTRODUCTION 10
2.2.2 APPLICATIONS 10
2.2.3 EXISTING APPROACHES 11
2.2.4 EVALUATION STRATEGY 16
2.2.5 SUMMARY 18
2.3 LOCAL GRAMMAR (LG) 19
2.3.1 INTRODUCTION 19
2.3.2 ADVANTAGES 20
2.3.3 APPLICATION AREAS 21
2.3.4 METHODS AND TECHNIQUES 22
2.3.5 THE APPROACH 25
2.3.6 THE EVALUATION STRATEGY 27
2.3.7 SUMMARY 28
2.4 PREREQUISITES 28
2.4.1 CORPUS CREATION 28
2.4.2 FREQUENCY ANALYSIS 30
2.4.3 COLLOCATION ANALYSIS 32
2.4.4 CONCORDANCE ANALYSIS 34
3 METHODOLOGY 36
3.1 INTRODUCTION 36
3.2 DATA SET 37
3.2.1 THE EC2000 DATA SET 38

3.2.2 THE METU TC DATA SET 39
3.3 ENVIRONMENT 40
3.3.1 THE NOOJ SOFTWARE 41
3.3.2 THE TEXT ANALYSIS TOOL 44
3.4 METHODOLOGY 45
3.4.1 PREPROCESSING THE EC2000 AND WORD FREQUENCY ANALYSIS 45
3.4.2 EXTRACTION OF TURKISH REPORTING VERBS 48
3.4.3 SIGNIFICANCE TEST OF TURKISH REPORTING VERBS 50
3.5 NAMED ENTITY EXTRACTION 55
3.5.1 CONCORDANCE ANALYSIS 57
3.5.2 COLLOCATION ANALYSIS 70
3.5.3 CONCORDANCE ANALYSIS FOR FOURTH FORM OF REPORTING VERBS 72
3.5.4 CREATING THE LIST OF ALL PATTERNS FOUND 77
3.5.5 EXTRACTING PERSON NAMES 84
3.6 SUMMARY 86
4 EVALUATION 88
4.1 THE EVALUATION STRATEGY AND THE EXPERIMENTAL RESULTS 88
4.1.1 EVALUATION STRATEGY 88
4.1.2 EXPERIMENTAL RESULTS 91
4.1.3 DISCUSSION 92
4.2 CONCLUSION 97
5 CONCLUSION & FUTURE WORK 98
5.1 CONTRIBUTIONS 99
5.2 FUTURE WORK 100
REFERENCES 102
APPENDICES
A:Most Frequent Top 1000 Words in the FEC2000 109
B:Frequency Distribution of Top 100 Reporting Verbs 130
C:Maximum Likelihood Values of Top 100 Reporting Verbs 133
D:Descriptions of Columns in Concordance Lines 136
E:Concordance Lines of First Type of Verb Form Including Seven Reporting Verbs 137
F:Concordance Lines of Second Type of Verb Form Including Seven Reporting Verbs 139
G:Concordance Lines of Third Type of Verb Form Including Seven Reporting Verbs 141
H:Concordance Lines of Fourth Type of Verb Form Including Six Reporting Verbs 143
I:Collocations of Reporting Verb “söyledi” 145

LIST OF TABLES

Table 3-1 – Data set characteristics of EC2000 from 2000 to 2007......... 38
Table 3-2 – Data set characteristics of the EC2000................................ 38
Table 3-3 – The one-sample Kolmogorov-Smirnov non-parametric test.. 51
Table 3-4 – The two-sample Kolmogorov-Smirnov non-parametric test .. 52
Table 3-5 – The Mann-Whitney non-parametric test............................... 53
Table 3-6 – The Mann-Whitney non-parametric test after removing top forty reporting verbs... 54
Table 3-7 – Notations which are used in the descriptions of patterns...... 58
Table 3-8 – Pattern 1 for RV_Form_1... 58
Table 3-9 – Pattern 2 for RV_Form_1... 59
Table 3-10 – Pattern 3 for RV_Form_1... 60
Table 3-11 – Pattern 4 for RV_Form_1... 60
Table 3-12 – Open class word pattern for RV_Form_1........................... 62
Table 3-13 – Pattern 1 for RV_Form_2... 63
Table 3-14 – Pattern 2 for RV_Form_2... 64
Table 3-15 – Open class word pattern for RV_Form_2........................... 65
Table 3-16 – Pattern 1 for RV_Form_3... 66
Table 3-17 – Pattern 2 for RV_Form_3... 67
Table 3-18 – Pattern 3 for RV_Form_3... 68
Table 3-19 – Open class word pattern for RV_Form_3........................... 69
Table 3-20 – Top ten collocations of reporting verb “söyledi” 72
Table 3-21 – Pattern 1 for RV_Form_4... 73
Table 3-22 – Pattern 2 for RV_Form_4... 73
Table 3-23 – Pattern 3 for RV_Form_4... 74
Table 3-24 – Pattern 4 for RV_Form_4... 75
Table 3-25 – Pattern 5 for RV_Form_4... 75
Table 3-26 – Pattern 6 for RV_Form_4... 76
Table 3-27 – Pattern 7 for RV_Form_4... 76
Table 4-1 – F-measure, recall, and precision results for local grammar approach with C.R. (Capitalization Rule)... 91
Table 4-2 – F-measure, recall, and precision results for local grammar approach and capitalization rules. C.R. stands for Capitalization Rule 92
Table 4-3 – Materials containing NEs which should have been categorized as organization names.. 93

LIST OF FIGURES

Figure 3-1 – Text distribution according to genre in the METU TC 40
Figure 3-2 – Sorted word frequency list of the EC2000 January news in the Nooj software 41
Figure 3-3 – Locate a pattern interface 42
Figure 3-4 – The Nooj regular expression of abbreviation pattern 42
Figure 3-5 – The Nooj syntactic grammar of abbreviation pattern 43
Figure 3-6 – Concordance lines of abbreviation list 43
Figure 3-7 – First step is to find word frequencies in the EC2000 46
Figure 3-8 – Graph of sentence splitter 48
Figure 3-9 – Second step is to find all significant Turkish reporting verbs 49
Figure 3-10 – Third step is to extract person names in the FEC2000 56
Figure 3-11 – Collocation analysis of the reporting verb “söyledi” 71
Figure 3-12 – The graphical representation of pattern 1 78
Figure 3-13 – The graphical representation of pattern 2 79
Figure 3-14 – The graphical representation pattern 3 79
Figure 3-15 – The graphical representation of pattern 4 80
Figure 3-16 – The graphical representation of WORD1 81
Figure 3-17 – The graphical representation of WORD2 81
Figure 3-18 – The graphical representation of CHAR1 81
Figure 3-19 – The graphical representation of CHAR2 82
Figure 3-20 – The graphical representation of DIGIT 82
Figure 3-21 – The graphical representation of UPPER 83
Figure 3-22 – The graphical representation of LOWER 83
Figure 3-23 – The graphical representation of SPECIAL CHARACTERS 84

LIST OF ABBREVIATIONS

DM	: Data Mining
IE	: Information Extraction
IR	: Information Retrieval
LG	: Local Grammar
MUC	: Message Understanding Conference
NE	: Named Entity
NEE	: Named Entity Extraction
NER	: Named Entity Recognition
NLP	: Natural Language Processing
OCW	: Open Class Word
PN	: Proper Noun
RV	: Reporting Verbs
TM	: Text Mining

CHAPTER 1

INTRODUCTION

Proper noun (PN) is a language expression that represents a unique named entity (NE) such as an organization name, a person name, or a place name. PNs are usually considered as the opposite of the common nouns. PNs are different from common nouns because they have unique referents and do not exhibit the semantic properties of these referents, whereas common nouns have common semantic properties of their referents and do not have unique referents at all **[1]**.

In **[1]**, it is stated that all kinds of names are studied in the field of Onomastics. There are many sub-categories of Onomastics such as Anthroponymy, which studies person names, and Toponomy, which studies place names. Since we try to find person names in financial news, we are more interested in Anthroponymy. Naming systems in Anthroponymy differ according to how the types of person names are used. The types of person names in Anthroponyms can be defined as;

- *First Name*: Süleyman is the first name of a person named Süleyman Demirel.
- *Surnames*: Surnames are shared by members of a nuclear family in societies. Demirel is the surname of a person named Süleyman Demirel. In most of societies, surnames pass from father to son.

- *Clan Name*: A clan is a traditional social unit which comprises of a number of families that are claimed to have a common ancestor, share identifying marks or slogans which are associated with particular natural phenomena such as red American Indian in America
- *Patronyms*: A patronym is a name which is used to identify someone by referring to him or her as the son or daughter of someone else. For example, Johnson is the son of John or Macdonald is the son of Donald.
- *Teknonyms*: A teknonym is a name which is used to address the parents (father or mother) of someone. It is common in parts of Africa and the Arab World. For example, Abu Ali means that Abu is the father of Ali in most Arab countries.
- *Nickname*: A nickname is an informal name given to people and is usually not used in formal documents. For example, Abu Ammar is the nick name of Yasser Arafat, the former Palestinian president.
- *Ethnonyms*: An ethnonym is a name which is used to refer to an ethnic group such as Englishman, Arab, Indian, Russian, and Slav.

If the naming systems for these types of person names and their nature are understood correctly, then the effort that is to be spent on the recognition of person names can be reduced. Therefore, successful NER systems can be developed by using that knowledge. In addition to understanding the naming systems, linguistic aspects of PNs should be examined in detail in order to develop a successful NER system. Early linguistic studies of PNs usually focused on their meaning. However, today, modern linguistic studies focus more on the semantic categories and the syntactic aspects of PNs **[1]**.

As cited in **[1]**, Quirk defined syntactic categories of PNs as **[2]**;

- PNs with articles: person names, temporal names, and geographical names.
- PNs without articles: PNs without modifications, PNS with pre modifications, and PNs with post modifications.

As cited in **[1]**, Allerton defined semantic categories of PNs as **[3]**;

- *Human Beings*: For example, Mustafa Kemal Atatürk.
- *Vessels, Vehicles and Machines,* such as Metro.
- *Geographical Locations,* such as Turkey.
- *Social Organizations,* such as Atatürkçü Düşünce Derneği.
- *Publications and Works of Art,* such as Bilim ve Teknik.
- *Languages and Dialects,* such as Turkish.

In **[4]**, PNs are categorized as geographical entity (such as city, port, airport, island, county, province, country, continent, region, and water), affiliation (religion and nationality), organization (company and company type), human (person and title), document, equipment (software, hardware, and machines), scientific (diseases, drugs, and chemicals), temporal (date and time), and miscellaneous.

Although PN is a subtype of NE in linguistics, PNs and NEs are used interchangeably in the information extraction (IE) and information retrieval (IR) fields. NEs include PNs (organization names, person names, and location names), temporal expressions (dates and times), and number expressions (monetary values, percentages) **[5]**. The task of recognizing NEs in free texts in computational linguistics is called NER. NER systems try to identify NEs and categorize them into semantic sub-categories that are described above.

Most of the NER systems require linguistic resources like list of markers, large dictionaries of general nouns, namely gazetteers, and general dictionaries. They use dictionaries and gazetteers to classify the NEs. Therefore, we can state that most of the time, recognition of PNs is required in NER systems. However, creating dictionaries and gazetteers is a difficult task. Consequently, the approaches which do not require dictionaries and gazetteers such as local grammar approach are preferred.

NER is related with many disciplines. Data mining (DM), text mining (TM), natural language processing (NLP), IR, and IE are the important ones of these disciplines.

1.1 DATA MINING (DM)

Today, computer usage has become inevitable because of the increasing technology in the world. Consequently, the available information has increased very rapidly. It is very difficult and costly to extract meaningful information when the amount of available information that is stored in large databases reaches gigabytes, sometimes terabytes. DM methods simplify this costly task by automatically acquiring useful information from these large databases. DM is the process of identifying patterns, extracting previously unknown but useful information and discovering novel relationships in the data to make crucial decisions by applying statistical and machine learning methods. It is applied to common business problems to increase the productivity of people working in businesses **[6]**. DM assumes that the information that is to be mined to be in the form of relational databases. Relational databases are structured databases which consist of contextually and semantically well-defined data. However, the available electronic information in natural language documents is not always in the form of structured databases for many applications. Most of the time, it is in the form of unstructured or semi-structured databases.

1.2 TEXT MINING (TM)

TM is the process of discovering novel information from a collection of texts which is also known as corpus **[7]**. TM searches for patterns in unstructured or semi-structured texts. Thus, it can be considered as an extension of DM. However, TM differs from DM in that in TM, patterns are extracted from natural language texts rather than from structured databases **[8]**. TM applies DM techniques to unstructured or semi-structured textual data, whereas DM applies these techniques to structured data.

TM uses the disciplines of IR, NLP, IE and DM to discover structure, patterns, and knowledge in large textual corpora. IR systems identify the documents in a collection and provide the relevant set of documents by applying IR methods such as pattern matching, keyword matching, and word frequency analysis to discover what a document is about. IR reduces the number of documents that are to be analyzed by finding the relevant documents; therefore, it speeds up the analysis considerably. TM is related to IR because it is an application of IR but it is limited to texts. TM is also related to DM because it goes beyond search and retrieval.

TM also uses NLP which is a range of computational techniques for analyzing and representing naturally occurring texts to enable computers to have a human-like understanding of language ability **[9]**. Types of analysis that NLP can perform are part of speech tagging that is used to classify words into categories such as noun, verb or adjective, word sense disambiguation that is used to assign appropriate meanings to words from the set of possible meanings that the word may have, and parsing texts by performing grammatical analysis of sentences. As cited in **[7]**, computational linguistics extrapolates knowledge from numerical data to corpora. Statistical methods can be used over large corpora to extract useful patterns by the help of computational linguistics. These patterns can be used to solve the problems within NLP methods together with speech tagging and word sense disambiguation.

Another approach that TM uses is IE, which is the process of obtaining structured data from unstructured natural language documents. IE facilitates the examination of text by partially analyzing it to find specific target terms that can be used for further analysis. In **[10]**, IE is defined as text understanding that locates specific pieces of data in natural language documents and transforming unstructured texts into structured databases. It is different from IR in the sense that IE tries to extract meaningful structured information inside the related documents rather than searching for documents and their metadata and finding the desired related documents.

IE systems rely heavily on data which are generated by NLP systems. NLP provides the linguistic data to IE systems in order to be used in IE phase. In NLP, corpora of textual documents can be transformed into more structured databases by using IE methods and then novel relationships can be discovered in the resultant relational databases by applying DM techniques. The tasks of an IE system include term analysis, which identifies the terms in a document, fact extraction, which identifies and extracts complex facts from documents, and NER that identifies the NEs in documents such as person names and organization names.

1.3 NAMED ENTITY RECOGNITION (NER)

NER aims to identify the NEs in the texts and classify them into semantic categories such as person names, organization names, place names, time expressions, date expressions, monetary expressions, and percent expressions. LGs have recently been shown to be superior to other NER techniques such as the probabilistic approach (supervised machine learning), the symbolic approach (rule based), and the hybrid approach in terms of being able to work with untagged corpora **[1]**. Unlike most of the other NER systems, in the LG approach, it is not required to use any dictionaries and gazetteers. As a consequence, the LG approach is able to recognize NEs in previously unseen texts at minimal costs. Most of the NER systems are costly due to manual rule compilation especially in large tagged corpora. They also require some semantic and syntactic analyses to be applied before the pattern generation process, which can be avoided by using the LG approach.

To sum up, we believe that the power of NER of today makes many difficult tasks much easier. The recognition of person names has become a necessity in many of the fields for some special purposes, such as person profiling issues. Thus, we decided to apply NER to Turkish language in the financial domain.

1.4 THESIS OUTLINE

This thesis is composed of five chapters. Chapter one is the introduction chapter that introduces key terminologies in the scope of this thesis. We first give introductory information about NEs and PNs. After introducing these and other key terminologies, we introduce the related fields, which are data mining, text mining, information retrieval, information extraction, and natural language processing, and NER. Then, we give some brief information about NER. We conclude this chapter explaining why this research has been undertaken.

Chapter two is dedicated to literature review. In this chapter, we first give detailed information about PNs, their types, and their usage in NER. Next, we describe what NER is, where it is used, the existing approaches in NER, and the evaluation strategy in detail. Then, we present the LG approach, which is applied in this thesis. We also describe what LG is, its advantages, where it is used, methods and techniques for LG, description of the LG approach, and the evaluation strategy in detail. Finally, we discuss some prerequisites of the LG approach, which are: corpus creation, frequency analysis, collocation analysis, and concordance analysis.

In Chapter three, we first introduce the overall description of the methodology used in this thesis briefly. Next, we outline the characteristics of two data sets (the economy corpus and the reference corpus) that we used in this thesis. Then, we explain the tools used in our study. Finally, we define our methodology and describe the steps that we follow in NEE in detail.

In Chapter four, we present the evaluation strategy that we followed and the experimental results that we obtained. Then we discuss them in detail.

In chapter five, we summarize what we have covered in this thesis. We give information about what this thesis contributes to the literature. Finally, we conclude with a discussion about possible future work in this field.

CHAPTER 2

LITERATURE REVIEW

2.1 PROPER NOUN (PN)

PNs are language expressions that represent unique NEs such as organization names, person names, and place names. Most of the time, they are considered to be the opposite of common nouns because, as I mentioned before, PNs are different from common nouns because they have unique referents and they do not exhibit the common semantic properties of their referents **[1]**.

Common nouns are general nouns and in writing and they are not capitalized. They are capitalized only when they are used at the beginning of a sentence or in the title of PNs. On the other hand, PNs can be people names, places names, things, and ideas, such as Mary, Confucianism, Douglas College, Vancouver, and Superman. They are capitalized in most of the languages; therefore, it is usually easier to identify PNs than common nouns **[11]**.

Although PNs can be considered as a sub type of NEs in linguistics, PNs and NEs are used interchangeably in the IR and IE fields. The computational linguistic task of recognizing NEs in free texts is called NER. It tries to identify and categorize the NEs into semantic sub categories.

In the following subsection NER systems and techniques will be discussed in detail.

2.2 NAMED ENTITY RECOGNITION (NER)

2.2.1 INTRODUCTION

NER was born in the Message Understanding Conference (MUC), which aimed for the standardization of IE tasks. MUC was sponsored by the US Defense Advanced Research Projects Agency (DARPA), in the 1990s. MUC defined three subtasks for NER which were to extract entity names, temporal expressions, and number expressions **[5]**. The tasks of finding temporal expressions and number expressions are both much easier than the task of finding NER because of their simple structure.

The task of NER is to identify NEs in the texts and classify them into semantic categories such as person names, organization names, place names, time expressions, date expressions, monetary expressions, and percent expressions. NER is frequently used in IE, which aims to understand the texts from specific pieces of data in natural language documents and to convert these texts into structured databases. This is different from IR, which only makes searches to find the desired documents.

2.2.2 APPLICATIONS

Today, NER systems have an important role in many application areas. NER is used in;

- Machine Translation Systems in order to translate unknown words and to reduce the ambiguities **[12]**.

- IR in order to have more relevant texts and to reach more accurate search results by disregarding the irrelevant documents **[12]**.
- Document organization in order to find all documents related to a particular topic **[12]**.
- Answering questions. It associates words with their semantics. For example, “Africa” is associated with “country”, “John” with “person” and “1.2 inches” with “length” **[13]**.

2.2.3 EXISTING APPROACHES

There are three widely used approaches for NER systems: the rule based (symbolic) approach, the supervised machine learning (probabilistic) approach, and the hybrid approach in order to acquire PNs in the texts. In the symbolic approach it is required that the natural language descriptions, models, and linguistic resources to be provided explicitly. However, the probabilistic approach does not require any explicit natural language descriptions. Instead, it builds its own models by learning from the corpora **[1]**.

In the probabilistic approach, *statistical language models* which are trained on large, manually annotated corpora to learn identification patterns are used. For example, *Resolve* uses decision trees and requires tagged coreference examples to learn how to classify pairs of phrases as coreferent, which means having a common referent, or not coreferent **[14]**. *Crystal* automatically extracts a conceptual dictionary in order to identify information by generalizing the patterns from an annotated training corpus **[15]**. On the other hand, *AutoSlog* automatically constructs a domain-specific dictionary of concepts in order to extract information from an annotated corpus of text by using thirteen general syntactic patterns **[16]**.

The probabilistic approach displays good performance if it is trained with a large corpus. For example, *Nymble* of BBN corporation achieved high success rates (more than 90%) for English and Spanish languages in MUC-7 after manually tagging words in the large corpora **[17]**. Other statistical systems in **[18]** that participated in MUC-7 achieved a lower success rate (at most 90%) for English.

In the symbolic approach, the *rules* that are developed intuitively are used to define NEs in terms of their syntactic and lexical structure in manually annotated corpora. Symbolic approach requires linguistic resources such as list of markers, gazetteers (large dictionaries of general nouns) and general dictionaries in order to classify NEs **[1]**.

The hybrid approach is the combination of probabilistic and statistical approaches. As cited in **[1]**, this approach has been used in NEE systems that were developed at the University of Edinburgh (Mikheev, C. & Moens 1998), the University of Manitoba (Lin 1998), and by Cucchiarelli & Velardi (Cucchiarelli & Velardi 1999).

Leading systems that are based on symbolic approach are:

NetOwl is an entity extraction product which finds and classifies key concepts in unstructured texts accurately in multiple languages by using advanced computational linguistics. In addition to NEE, NetOwl also provides links and events between extracted entities. It is also used for resolving aliases of extracted entities and associates them to the same real-world object. Because NetOwl is a rule based system, where as certain rules can work efficiently in a specific domain, most rules may not be carried into a new domain **[19][20][21]**.

Facile is a rule-based NER system which aims to perform NE task defined in MUC-7. It performs some processes such as preprocessing the text, tokenizing, tagging, recognizing special formatting, and searching single and multi-word tokens in the database on input. And then it recognizes

PNs and classifies them **[22]**. However, in **[1]** it is stated that Facile joined in MUC-7 and scored 93% F-Measure on training data where as 79.50% on the test data which was provided by the MUC organizers. This low performance can be explained by the lack of the rules which require a great amount of manual effort to be developed.

Gate is a tool that is used to construct, test, and evaluate the language engineering systems **[23]**. It reduces the difficulty of integration, documentation, and data visualization because it reuses NLP components that can be reused. However, Gate has some drawbacks too. First, Gate has complex and non standard interface. Second, installing systems of Gate is very difficult and they do not run on all platforms. Finally it does not support all character sets other than 8-bit character sets **[24]**.

In **[25]**, it is denoted that early studies had been mostly based on the symbolic (rule based) approach while recent studies are mostly based on the probabilistic approach (supervised machine learning) because the probabilistic approach automatically extracts rules from training sets. However it should be stated that when training examples are not available, the symbolic approach is preferred to the probabilistic approach. Supervised machine learning approach needs a large annotated corpus. If an annotated corpus is not available, or it is costly to create, supervised machine learning approach becomes inconvenient. For these cases, semi-supervised machine learning and unsupervised learning are two alternative learning methods. The main technique used in semi supervised machine learning is "*bootstrapping*" and includes little supervision like giving a set of seed for starting the learning process. For example, if a system tries to find names of the diseases in texts, a small number of example names can be given to the system. The system then tries to find some common clues about the given disease names and then tries to find other instances of disease names which are used in similar context. On the other hand, the typical approach in unsupervised machine learning is “*clustering*”. Unsupervised machine learning uses lexical resources, lexical patterns, and statistics which are calculated on a large unannotated corpus. It is costly to create a large annotated corpus and maintain rules because rules and dictionaries have to be changed according to the application. Therefore, unsupervised machine learning has the advantage of recognizing NEs without requiring the cost for the construction of a large manually annotated training corpus or a lot of rules **[26]**.

Although there are many NER systems which are successful in recognizing NEs, there are still many problems in the process of identifying and classification of PNs. As seen in **[1]**, first difficulty is that NER systems should recognize the places following to people names, organizations following their founders, person names that contain foreign first names, and the names of organizations which consist of single common words. Second difficulty for NER systems is that they should handle the problems of structural and semantic ambiguities that are occurred with PNs. Third important difficulty for NER systems is that they should decide on splitting a sentence into pieces or tagging whole sentence. For example, the phrase Fenerbahçe'li Cemil Turan should not be tagged as one PN. Actually, it has to be split into two different PNs: an organization name (i.e. Fenerbahçe) and person (i.e. Cemil Turan) name. Similarly, the organization name Rahmi Koç Müzesi should not be split and tagged as one organization name.

NER systems are expected to correctly recognize NEs in new domains at minimal costs; however, it is very difficult to carry out such a task because of the arising ambiguities during the NER process. Examples of these ambiguities are given above. This difficulty can be handled by using manual effort.

In **[1]**, Hayssam denoted that most of the NER systems use parts of speech, syntactic taggers, or capitalization rules to locate candidates of

PNs and try to categorize these PNs by using the data bases of first names, job titles, organizations and reporting verbs (RVs) to identify their semantic categories. However, his NER prototype (NExtract) does not use any syntactic taggers or part of speech taggers to recognize PNs.

It only uses LGs, which are found by the LG-Finder that Hayssam developed for finding LGs automatically, in order to extract proper names. The LG-Finder covers the materials that contain human-subject-reported verbs to locate person names and organization names. Proper names extracted by the NExtract are used with the lists of first names, job titles, and markers for organization names to disambiguate the sequences of initially capitalized words that are extracted by using capitalization rules but missed by his LGs. This strategy has two main advantages. First, it does not use part of speech taggers to specify the PN candidates. Second one is that it reduces the high cost associated with compilation of symbolic rules.

There is a limited work for NER systems applied to Turkish texts. In **[27]**, preprocessing Turkish texts by using a multi-word expression processor is studied. In other words, the multi-word expression extraction that recognizes the fragments of input texts that contain word materials, which do not have compositional syntactic and semantic structures, is studied. It is focused on four different forms of multi-word expressions in Turkish. These four forms can be defined as lexicalized collocations, semi-lexicalized collocations, non-lexicalized collocations, and multi-word NEs denoting person names, organization names, and location names. They used a simple approach which employs a huge database of person, organization, and location names instead of using complex NEE scheme in order to extract the multi-word NEs. Oflazer et al. tested their multi-word expression processor on a large corpus and a small corpus. While the large one comprised 730,000 tokens, small one consisted of 4,200 tokens. They used recall and precision metrics as evaluation metrics and obtained

65.2% overall recall and 98.9% overall precision. Although they achieved a high precision rating, they did not succeed in obtaining a high recall value. In the evaluation phase, they did not use an international evaluation tool like MUC scoring software to evaluate their performance.

In **[28]**, a probabilistic model for automatically tagging names in Turkish text is also presented. For name tagging, which can be considered as the subtask of finding only names in NEE, they used an approach which consists of four models. The first one is the lexical model which tries to find lexical information by using only word tokens. The second one is the contextual model which tries to find the context information for the word tokens. The third one is a morphological model which tries to find the morphological information with respect to the corresponding case and name tag information. The fourth one is the name tag model which tries to find the name tag information such as information of word tokens which can be categorized as person, organization, and location names. Finally, they combined their four models and based their testing and evaluation on the combinations of these four models by using MUC scoring software. They trained their system on 492,822 words of newspapers articles containing 37,277 names and tested their system on 28,000 words of newspaper articles containing 2,197 names. They used the F-measure metric to evaluate their system. They reached 91.56% F-measure for named tagging tasks in Turkish for combined model.

2.2.4 EVALUATION STRATEGY

Most of the NER systems which are based on rule based (symbolic) approach, supervised machine learning (probabilistic) approach, or hybrid approach usually use F-Measure which is computed by the uniformly weighted harmonic mean of precision and recall as an evaluation metric to measure their performance and compared their results with the test corpus which is annotated by human annotators. Precision is the measure of how much the response results are actually in the test set. Recall is the measure of how much of the test set are covered by the response results. Most of the time MUC scoring software is used to test the performance of NER systems.

In **[1]**, Hayssam used MUC scoring software and its primary measures, which are precision and recall, in order to test their NER tool. The performance of the system was checked by using only capitalization rules and by using capitalization rules and the LG approach at the same time in order to measure the effect of the LG approach. The NExtract performs much better when the capitalization rules are used with LGs. He also used F-measure as an evaluation metric. Although, he achieved 90% recall, 54% precision, and 79% F-measure scores when they use only capitalization rule, they get 90% recall, 88% precision, and 90% F-measure scores when he used capitalization rules with LGs. Consequently, they get remarkably better result when they used capitalization rules with LGs.

The testing data set is tested by using MUC-7 NE evaluation tool. Two test sets of news articles are used in the evaluation phase. He used 120 news articles from Reuter’s news published in 2002 as the first test set. This set is manually annotated by 6 PhD students in the Computing Department of the University of Surrey. While annotating the first test set, students used MUC international standard tags. These tags are person tags and organization tags which can be seen below;

- <**ENAMEX** TYPE="ORGANIZATION"> Organization name</**ENAMEX**>
- <**ENAMEX** TYPE="PERSON">Person name<**ENAMEX**>

As seen above organization and person names are bracketed with MUC tags for organization names and person names, respectively.

He also used 120 news articles which are chosen from the MUC-6 corpus randomly as the second test set. This second test set was initially taken from the Wall Street Journal (WSJ) and was manually annotated by the Language Data Consortium-8 (LDC-8).

In **[28]**, MUC scoring tool was also used to test their NEE tool. The overall accuracy result was computed by using the F-measure. In this study 91,56% F-measure was achieved.

In **[29]**, Mason performed automatic tests in order to see how using different sample words in patterns affect the result of his system. He annotated a number of sample lines with the correct pattern that should be identified by his system in order to evaluate the system performance. In this way, whether using different sample words in patterns increase the overall performance of the system or not can be tested. Four possible outcomes of a pattern match from the system are;

- The system can find correct pattern with highest ranking
- The system can find correct pattern but not with highest ranking
- The system cannot find correct pattern
- The system can find a pattern which is not in the data

They used precision, recall, and F-measure metrics to evaluate the outcomes of the pattern matches.

2.2.5 SUMMARY

When we consider the fact that most of the NER systems are based on rules that are manually compiled from tagged corpora, it can be stated that NER systems like the one in **[1]** have notable advantages to other NER systems because of their automatically generated LG patterns. In **[1]**, it is stated that after being implemented into finite state transducers, LGs can be reused in larger LGs. This strategy helps making the systems that uses the LG approach much more portable. When it is needed to examine texts from another domain or to extract new categories of PNs, this strategy also reduces the need for re-engineering the existing NER system.

LGs that are widely used in NER will be explained in the following sub-section.

2.3 LOCAL GRAMMAR (LG)

2.3.1 INTRODUCTION

“The term local grammar is used to suggest that the local one extends the general grammar when one is required to express information which is related to a specific class of named entities” **[1]**. In **[30]**, it is stated that global grammar covers dependencies between words which are far away from each other by using several difficult and complex grammatical rules, whereas LG covers the dependencies between neighboring words.

More specifically, LG is a way of recognizing the behavior of words that are used in a specific domain, finding how these words are used in sentences and inferring their usage patterns. In the literature, Oliver Mason defined LG “a way of describing the syntactic behavior of groups of individual elements, which are related but whose similarities cannot easily be expressed using phrase structure rules” **[29]**. As cited in **[1]**, Harris defined LG as a way of describing syntactic restrictions of certain subsets of sentences, which are closed under some or all of the operations in the language **[31]** and Gross defined the LG as a finite state grammar and used it for finding words related by prefixation, suffixation, and sentences having similar syntax **[32].**

Because LGs are finite state transducers, they do not have length and complexity limit. Therefore, any number of LGs can be used at the same time to disambiguate texts. LGs recognize the word sequences in the sentences and find the appropriate tags for these word sequences. It is highly possible that, there can be lexical ambiguities in the sequences of words in the sentences. These lexical ambiguities are needed to be reduced. LGs can be used for this purpose. Goran Nenadić states that “local grammar does not describe the whole sentence; it defines and gives lexical constraints to the word sequences in the sentences” **[33]**.

Lexicon grammar tables and graphs are used to represent LGs. The graph representation is more practical because it is easy to read and understand it. A large set of syntactic structures can be represented by lexicon grammar tables. Lexicon grammar tables, however, are needed to be transformed into graphs when they are used for computational purposes **[1]**.

2.3.2 ADVANTAGES

LGs detect errors or unnatural word sequences and reduce most of the lexical ambiguities by defining and applying lexical constraints to the word sequences in the texts. Therefore, they find patterns in the sequences of words that obey some strict rules in the texts more easily **[34]**.

Automatically extracting rules by using LGs from corpora instead of finding rules manually reduces the high cost perceptibly. Besides, translating the rules into the LG as finite state transducers has a speed and accuracy advantage.

It is highly possible for an expert to miss some of the patterns that contain PNs while constructing LGs manually because people are always prone to error. Therefore, LGs sometimes may have low performance and this can be considered as the drawback of LGs.

2.3.3 APPLICATION AREAS

LGs have a wide range of application areas and can be used for many different purposes. These are as follows;

- LGs are used for describing the linguistic description of a language that covers the texts in a specific domain and for extracting the patterns in previously unseen texts **[35]**.
- LGs are used for adaptive learning tasks because after identifying the patterns in the unseen texts, they add identified patterns to the LG base to cover unrecognized patterns **[35]**.
- LGs are used for lemmatization and lexical disambiguation. Lemmatization is the process of determining the lemma for a given word. Lexical disambiguation is a process that describes lexically ambiguous sentences within contextually appropriate sentences **[36][37][38]**. In **[39]**, Nouns and prepositions that are used with other nouns are found. After finding the nouns and their prepositions, the results may contain some ambiguities. It is stated that LGs were used for eliminating these ambiguities.
- LGs are used for dividing the input texts into the sentences. Some punctuation can be used to understand the end of the sentences such as the full stop, the exclamation, the quotation mark, the question mark, and etc. **[1]**.
- LGs are used for translation of any word structure such as time expressions from one language to another **[1]**.
- LGs are used for string matching that is finding occurrences of a string within another string or body of text **[38].**

In **[40]**, five LG types are defined for describing the person name contexts in Korean texts and nouns that indicate a family relation such as son, father, etc. that frequently occur in the neighborhood of Korean person names. An auxiliary analysis tool was used to analyze proper names rather than a dictionary which can provide all proper names in the Korean language. As cited in **[1]**, Baptista also applied LGs to find PNs in Portuguese **[41]**. Friburger & Maurel used LGs to extract person names in French texts by investigating the right and the left contexts of person names in a corpus of Le Monde newspaper (about 165000 words) **[42]**.

2.3.4 METHODS AND TECHNIQUES

There are several methods or approaches for finding patterns and constructing LGs. One of them is the *bootstrap* method. In **[39]**, steps of the *bootstrap* method are defined as;

- The dictionaries of the usage of words according to their context are defined syntactically by linguists and stored in electronic dictionaries.
- Entries available in the electronic dictionaries are extracted.
- Concordance analysis is performed for every occurrence of each word.
- All new occurrences are sorted manually according to their lexical status by linguists. Concordance lines can be sorted according to either left or right contexts of words.
- Patterns are generated according to these concordance lines manually by linguists.
- NEs are extracted from concordance lines by using these patterns.

Six steps are carried out in the LG approach in **[1]**;

- Considering the assumption that proper names occur in materials that contain reporting verbs as the subject argument, "phrase-long collocations" of frequent RVs such as said, told, and added are generated.
- Concordance analysis is performed on found phrase-long collocations.
- Lexical feature vectors whose dimensions describe the number and positions of predefined sets of lexical units are constructed for each concordance line.
- LGs are constructed by classifying similar lexical feature vectors using a similarity measure, such as Euclidean distance.
- Most frequent constructions are formed.
- Extract PNs from sentences by using these most frequent constructions.

In **[43]**, it is questioned that whether information extracted at an earlier period of time is contradicted with or supported by information extracted at a later period of time by performing automatic analysis of specialist corpus. The algorithm used to acquire the LG patterns can be defined as;

- Identify the keywords and extract collocates of these keywords from a specialist corpus.
- Construct a sub corpus by using sentences that contains key collocates.
- Analyze the sub corpus and extract the trigrams (i.e. two collocate of a root node and root node itself) above a frequency threshold in the sub corpus. Note the position of the trigrams in each of the sentences.
- Analyze the sentences for the existence of the trigrams in the correct position. If a trigram that is found to co-occur with another frequent trigram that exists at the next position, call these two trigrams a pattern.

- Continue this process until all the trigrams in the sentence are matched with the significant trigrams.

In order to validate the patterns that are found in the training corpus, extracted patterns are used to find similar patterns and information from the test corpus.

In **[29]**, there are three steps in the approach that is applied. These are;

- The input was split into sentences.
- Ambiguity classes were assigned to the input tokens
- A set of patterns for a word by using recursive transition networks parsers were described. Transition network parsers use finite-state networks that can describe the individual elements (e.g. nouns, verbs, or other clauses) of the pattern found.

Most of the leading NER systems that are mentioned in the previous chapter are not able to learn linguistic patterns. They should perform the pattern generation as a prerequisite of NER. In **[1]**, it is indicated that there are few NER systems developed to learn extraction rules and mainly based on an iterative approach, where users should give base patterns for the entities that are tried to be found. Moreover, It is stated that their LG generation tool (*LG-Finder)* relies only on corpus linguistics techniques, which include frequency, concordance, and collocation analyses. This tool does not require an annotated corpus and base patterns.

As cited in **[1]**, there are also two other systems which generate linguistic patterns. These are *Dipre* (Dual Iterative Pattern Relation Expansion) and *Snowball*. Dipre is a method which is proposed in Brin **[44]** to learn linguistic patterns that relate two entities from the texts e.g. book title and author. Snowball is a system which generates linguistic patterns showing relations between entities and extracts entities from these relations in the free texts without human intervention. Hayssam argues that the main problem of Dipre and Snowball is due to the manual effort needed during the creation of regular expressions and training named entity component on a manually tagged corpus respectively. Tagging corpora manually requires great effort. Therefore, it can be stated that the LG-Finder takes the advantages of automatically acquiring LGs (i.e. linguistic patterns) for PNs from untagged corpora. Moreover, the LG-Finder does not need the construction of regular expressions and NE tagging.

2.3.5 THE APPROACH

Manual acquisition of LGs from corpora is a slow and exhausting task which causes problems for the portable NER systems. Therefore, automatically extracting LGs like in **[1]** can be considered as a better approach than others. The steps in this approach were explained in the section **2.3.4** in more detail.

In **[1]**, the frequency distributions of RVs in both the Reuters Finance Corpus **(RFC)** and the British National Corpus **(BNC)** which can be considered as representative samples of general language are analyzed in detail. When comparing the distribution of RVs in the BNC corpus with the distribution of those in the RFC, it was realized that frequencies of some RVs are considerably higher in the RFC, which is a corpus in specific domain. Therefore, it was focused on LGs around the materials that contain frequent RVs. Actually, the most frequent reporting verb occurs in RFC 4 times more than the one in the BNC. All the most frequent verbs totally occur in RFC 34 times more frequent than the ones in the BNC. Consequently, the materials that contain these most frequent RVs are examined and collocation analysis were performed on these RVs. Hayssam used three criteria supported by Smadja **[45]** to define the number of words to consider to the left and right of the reporting verb. Smadja's three criteria are;

- Collocates with low frequencies are eliminated assuming that the frequency of a word in the neighborhood of a lexical item must be greater at least one standard deviation than the average.
- Histogram of the frequencies of a word in the neighborhood of a lexical item should have at least one spike, which is a sharp turning point in the histogram. In other words, if the histogram of collocates for a lexical item does not contain any spikes, these collocates should not be considered as significant collocates of this lexical item.
- The frequency threshold of a word which has a certain distance from a lexical item has to be at least one standard deviation above the average frequency of that lexical item.

In **[1]**, the text fragments that have high frequency collocation patterns were used to create lexical feature vectors. The lexical feature vectors were created by calculating the frequency of each string in the collocation pattern and classifying these strings according to their lexical feature types. After creating the lexical feature vectors, similar structures were clustered together using the Euclidian distance similarity measure. During this process, only structures that have a Euclidian distance less or equal to a threshold distance are removed from the clusters and added to the new cluster. The clustering process continues until there is no more structure left. By finding collocates, a large number of concordance lines were decreased to a lower level, thus manual analysis and interpretation could be performed more easily.

In **[1]**, the LG-Finder extracts linguistic patterns of person names from an untagged text collection. It does not attempt to capture every instance of such patterns. Person names occur frequently in structurally similar patterns of RVs, which are sub-categorized for high animacy nouns in the test collection.

In **[1]**, Hayssam indicated that although traditional IE systems try to extract all relevant information from each document, the LG-Finder extracts only patterns from all of the documents in the test set and does not attempt to find every instance of these extracted patterns.

In **[46]**, LGs were used to find positive and negative sentiment words. They applied their method, which is called Laissez-faire, to the financial corpus in order to find frequent key terms and to disambiguate these key terms automatically. Although Laissez-faire takes advantage of LGs, it has three main disadvantages. The first one is that there are considerably high ambiguities in free natural language texts. For example, the verbs “fall” and “rise” can indicate positive and negative meanings in the financial domain. The second one is that some verbs have more than one meaning. For example, “rise” does not have positive meaning for flowers. The third one is that the selection of sentiment words is subjective. The Laissez-faire method has five steps. These are selecting the training corpus randomly, extracting the keywords, extracting candidate collocates, extracting LG patterns, and asserting LG patterns.

2.3.6 THE EVALUATION STRATEGY

The evaluation strategy in **[1]** can be defined as follows;

- Recall **(R)**, Precision **(P)** and F-Measure **(F)** metrics were used to test the accuracy of the system.
- Different encoding algorithms, which are used to generate the lexical feature vectors, were used to see how performance of the system was affected by this change.
- Euclidian distance was computed in different ways to see how its maximum estimated value was affected by this change.

In **[46]**, LGs generated by the LG-Finder and manually obtained LGs are compared with each other. Manual acquisition of LGs includes three steps; finding concordances for collocations, generating patterns from concordances, and grouping similar patterns. The effect of using LGs on reducing ambiguities and increasing the throughput is estimated and the system was evaluated on this criteria.

2.3.7 SUMMARY

To sum up, in this section we explained what LG is, where LG is used, advantages of LG, LG methods and techniques, and evaluation strategy in LG in this sub section.

2.4 PREREQUISITES

There are four prerequisite steps which are carried out in the LG approach. These steps are corpus creation, frequency analysis, collocation analysis, and concordance analysis.

2.4.1 CORPUS CREATION

2.4.1.1 INTRODUCTION

Corpus is a collection of large amount of written and spoken material about a language. Corpora of texts are used typically to study the structure and function of a language. The distribution of various linguistic units, which comprises the texts in a corpus, is used to make and test hypotheses about different linguistic descriptions of a language. However, “the specification of a corpus - the types and proportions of material in it - is hardly for linguists at all, but more appropriate to the sociology of culture” (pp. 13 - 23**, [47]**). Linguists also have to make selection of texts, and have to make serious and difficult decisions while compiling corpora. Therefore, they should be eager to describe and analyze language.

The general purpose of corpus creation is to make a good selection of the language which is an important prerequisite for dictionary compilation, machine learning, and IE. While creating a corpus, its size must be decided. The corpus should be as large as possible and should have an ability to grow, because large number of words is required to study the behavior of words in texts, which is finding the relation between words and their frequencies (pp. 13 - 23**, [47]**).

However, it should be stated that constructing a corpus with an appropriate size is a very difficult task. The type of the corpus is very important when constructing a large corpus. In (pp. 13 - 23**, [47]**), it is stated that the material of the corpus can be in electronic form, in the printed form, or in text processing form such as word processing, electronic mail etc. It is also indicated that there are three methods to use these forms of texts as an input. The first one is converting handwritten material into electronic form by keyboarding and transcription of spoken language. The second one is scanning mass printed books and converting them into electronic form. The third one is formatting materials that are already in the electronic form. While converting texts into electronic form, permission problems for converting these texts may occur. The copyright holders may not want their texts to be copied and converted into electronic form. Although, most of the copyright holders seem not to give permission for copying of their text, if it is explained that why their texts were desired, and what precautions will be taken against the exploitation to copyright holders, then it would be possible to avoid this largely unproductive labor (pp. 13 - 23**, [47]**).

Another difficulty during the creation of a corpus is that whether it should contain only written texts, or only transcriptions of spoken language, or both. In (pp. 13 - 23**, [47]**), it is stated that collecting spoken language is very difficult at the beginning of a project and collections of film scripts, drama texts, etc. are used instead of spoken language in this phase.

Which type of material should be used in creating corpus (the formal or informal material)? It is also explained the decision is made by the designer of the corpus. Corpus designers should consider the domain of corpus, styles of writers of texts in a corpus, and the period of the corpus.

2.4.1.2 MUST A CORPUS BE BALANCED?

As mentioned before, it is an important task to decide which size of corpus is appropriate in corpus creation. Another important task in corpus creation is the decision about whether to use a balanced corpus or an unbalanced one. As cited in **[48]**, Sinclair stated that a corpus does not need to be balanced because there can be number of serious problems in creating a balanced corpus in the Lexicography Conference. Sinclair also stated that it may even not be possible to correctly balance a corpus. In addition, it is expressed that there is more probability to find less frequent words like “imaginable” in these larger corpora rather than smaller one. Besides, as cited in **[48]**, Hanks states that “only a large corpus of natural language enables us to identify recurring patterns in the language and enables us to observe collocational and lexical restrictions accurately” **[49]**. It is also stated that if such phrases exist in large corpora, then it may be worth to investigate these phrases in more detail. Therefore, sometimes we do not have to use a balanced corpus for some special purposes. However, it is also denoted that the quirks in unbalanced corpora should be uncorrelated with different corpora **[48]**. As a consequence, if it is tried to find some less frequently used words in a corpus for special purposes, an unbalanced but a large corpus can be used instead of a smaller balanced corpora.

2.4.2 FREQUENCY ANALYSIS

Word frequency analysis is used for determining the distribution of each word in a given text. It is used as an alternative method of spell checking that is used to find and correct problematic words such as wrongly used homonyms. Word frequency analysis is used mainly for four purposes.

These are automatic classifications of texts, checking *keyword stuffing* in texts, plagiarism detection, and corpus comparison.

Frequency analysis is used by search engines in order to make automatic classification. In **[50]**, for example, the web documents are classified into pre-defined categories in order to increase the precision of web search. Classification process in this study is based on frequency analysis. They first defined a set of categories and a pre-defined training set of web pages. Second, they built a normalized vector of word frequencies for each of the categories from these web pages. Third, when a new document arrived, they computed the normalized word frequency vector of this new document. Then they compared it with the vectors of pre-defined categories and classified according to these pre-defined vectors by using a similarity matrix. As a consequence, it can be concluded that texts can be classified according to their subjects by using linguistic analyses without human support. Frequency analysis can be used for this purpose because the relation between the words and their frequencies is regular (p. 43, **[47])**.

Frequency analysis is used for checking whether there is any keyword *stuffing* in the texts or not. Words normally have expected frequency in terms of their usage in the texts. If a word is used with the frequency that extremely exceeds the expected frequency in the text, then this can be named as *keyword stuffing.* Frequency analysis tries to find the words with frequencies that are used much more than the expected frequency and assumes that these words cause *keyword stuffing*. *Keyword stuffing* can be handled by using the synonyms instead of the words that cause *keyword stuffing* **[51]**.

Frequency analysis is used for plagiarism detection which compares the given text in large texts. Because it is a very difficult task, the plagiarism detection system does not know where to start and how to find the plagiarism. Since authors have their own writing styles for usage of words (especially technical terms) in texts, the frequency of words that authors use have their own distributions. Therefore, frequency analysis can help to understand whether the contents of a given text and large texts are similar or not **[51]**.

In **[52]**, a copy detection mechanism is presented for digital documents. In this mechanism, the word frequency in the new document was compared with the frequency in the registered documents to detect whether this new document was copied from any document from registered documents or not. They first registered original documents and stored them in a document database. And then, if a new document arrives, a vector that contains the frequency of each possible word occurrence in the new document was created. Finally, this vector was compared with similar vectors in the document database.

Frequency analysis is also used to compare two or more corpora in terms of their similarities. In **[53]**, a method is presented to measure corpus similarity and corpus homogeneity by using word frequency lists as a sub step. To measure corpus homogeneity, the corpus was divided into two sub corpora by randomly placing texts in one sub corpus until all texts were placed in one sub corpus. Then, they produced frequency lists for each sub corpus and calculated the χ^2 statistics for testing the difference between two sub corpora. Finally, the process was iterated for randomly chosen different sub corpora. To measure corpus similarity, the same procedure was followed for two different corpora by taking first sub corpus from the first corpus and the second sub corpus from the second corpus.

2.4.3 COLLOCATION ANALYSIS

Collocation is the co-occurrence of two or more words which are related semantically. In **[54],** collocation analysis is defined as a statistical test that tells how likely two or more words can co-occur in texts. In (p. 170, **[47]**), collocation is also defined as “the occurrence of two or more words within a short space of each other in a text”. Collocation analysis aims to find the collocate–nodes of a given word. In **[55]**, it is stated that the word whose collocate-nodes are searched is called root node and its left and right words are called collocate–nodes. The number of words in neighboring of root node is called word span (or window span). If the word span is two, one word on each side of the root node is considered as a collocate node. It is also denoted that, although collocations are made up of n-grams, most of the time analysis is performed by using bi-gram or tri-gram. Bi-gram collocations comprise a root node and a collocate–node.

Root node and each candidate node in its neighboring are compared in collocation analysis. Each candidate node is given a score according to the overlap with its neighboring. At the end, candidate nodes with high scores are considered as correctly found collocate-nodes.

Two words can collocate with different frequencies. In (pp. 115 - 116, **[47]**), it is stated that if a root node is more frequent than its collocate–node, it is called downward collocation. On the other hand, if a root node is less frequent than its collocate–node, it is called upward collocation. Moreover, it is expressed that upward and downward collocation are systematically different and being stronger pattern, downward collocation gives much more semantic information for a word**.**

There are number of existing collocation applications which are;

- **TIGERSearch:** It is a software tool that tries to search annotated corpora syntactically, gets the sample sentences, and extracts the lexical properties of a given word such as extracting the collocate-nodes of a given node **[56]**.

- **Xaira:** It is a corpus indexing and searching tool that performs word search, concordance generation and manipulation, collocation analysis, lexical analysis. In Xaira, indexing of small or huge corpora can be performed efficiently **[57]**.
- **BNCweb:** It gives users access the BNC (British National Corpus) and its meta-textual annotation. BNCweb is based on SARA which is a predecessor of Xaira. BNCweb makes it possible to analyze the retrieved data by displaying the sorted search results, collocations, frequency distributions. It uses three different statistical association measures; log-likelihood, mutual information (point-wise) and chi-square to score the significance of word-pairs **[58]**.

2.4.4 CONCORDANCE ANALYSIS

Humans cannot discover all the significant patterns in large corpora. Therefore, concordance analysis can be used for this purpose. It lets linguists to access many important language patterns in texts. Concordance analysis extracts concordance lines which contain materials that include the keywords, which are tried to be searched by users. Computer generated concordance is very flexible because it lets users select the number of character or words on each side of the keyword. However, it should be stated that “concordance analysis is still highly labor-intensive and prone to errors of omissions made by humans” **[60]**.

In (pp. 170 - 171, **[47])**, it is stated that computers have made concordances to be compiled easily with the help of KWIC (key word in context) index convention. KWIC index, a way of displaying concordance lines in which the keywords are located in the middle of each line, and each line has some pre-defined number of characters, has been used widely for many years. Concordance lines exhibit left and right context of keywords. Although KWIC index is widely used, it is not the only way of displaying concordances. There are other possible techniques used for concordance analysis. For example, the whole sentence or the paragraph that contains keywords can be used in concordance analysis rather than KWIC, which have a fixed length for left and right context of a keyword. However, KWIC is still the most popular convention for computer generated concordance, because it is easy to scan concordance lines with KWIC convention.

Concordances are affected by various factors. In (p. 43, **[47])** these factors are defined as; first factor is to decide whether the concordance should be selective or exhaustive. Exhaustive concordance exhibits the available information when no strict rule about selection is required. Nevertheless, there can be cases in which selection must be made. If texts contain many of the commonest words, then some words may occur too much and selection becomes a necessity. Second factor is to decide the length of citation. In concordance analysis, KWIC convention is usually used in concordance analysis. As defined above, KWIC locates keywords in the middle and counts the lengths of the citations by characters or by words. It uses punctuation marks to identify sentences. Third factor is to decide how to order the citations. It is very important to decide how citations of word forms can be represented. The first and simplest way to present the citations of word forms is text order. Citations of word forms can be exhibited according to the preceding words, succeeding words, or both at the same time.

CHAPTER 3

METHODOLOGY

3.1 INTRODUCTION

Recently, automatic analysis of text in natural languages has been widely used in IE. As described in the Introduction chapter, IE is the process of obtaining structured data from unstructured natural language documents and facilitating the examination of text by partially analyzing to find specific target terms that can be used for further analysis.

In **[1]**, it is stated that there are three sub tasks of IE. These are template element construction which associates descriptive information with the entities, template relation construction identifying relations between entities in texts, and scenario template construction which ties entities with events and descriptions of relations between entities. All three sub tasks of IE are built on NER since, prior to an information system can determine the relationships between the entities, it should correctly recognize these entities, their categories, and their relationships with each other. In other words, they are all dependent on NER. Consequently, we can say that NER is a prerequisite for performing these three tasks of IE.

NER is the task of identifying the NEs, phrases each of which uniquely refers to entity object (some object in the world – for instance, a place or a person) by its proper name, acronym, nickname or abbreviation **[61]**, in the texts and classifying them into semantic categories such as person names, organization names, place names, time expressions, date expressions, monetary expressions, and percent expressions. In NER, finding only names is called name tagging.

In this thesis, we tried to extract NEs**,** specifically person names, from the Turkish financial news corpus by using the LG approach that has also been successfully applied to Reuter's financial English news corpus recently in **[1]**.

In this chapter, we introduce the methodology that we used to extract person names in a large untagged Turkish financial news corpus by using the LG approach. This approach aims to find similar sentences syntactically and capture their shared features by parsing the sentences in the texts to be found. Before searching LGs in the sentences in texts, frequency analysis, concordance analysis, and collocation analysis were performed and also a list of RVs was constructed. In chapter 3.2, we describe the characteristics of the EC2000, which is our financial corpus, and the METU TC, which is our reference corpus. In chapter 3.3, we introduce the Nooj software and the text analysis tool that we used to perform linguistic analyses. In chapter 3.4, we explain our methodology in detail. In chapter 3.5 we give information about NER. Finally, we summarize and conclude our work in chapter 3.6.

To sum up, in this thesis, we explore the applicability of the LG approach to Turkish language in order to extract person names from Turkish financial news.

3.2 DATA SET

Two corpora are used for NER in this thesis. The first one is Economy Corpus of the year 2000 (hereafter called "EC2000"), which is taken from

Anadolu Agency (A.A.) and contains 11.518.306 words. The second one is METU Turkish Corpus (hereafter called "METU TC") taken from Informatics Institute official web site **[62]** and includes 1.889.080 words. We used this corpus as a reference corpus because it is one of the best studies prepared for Turkish.

3.2.1 THE EC2000 DATA SET

Although we obtained news from the year 2000 to year 2007 from A.A, we used the financial news of the year 2000 as training data set because of performance issues. The data characteristics of Economy Corpus from the year 2000 to the year 2007 without cleaning processing are given in the Table 3-1.

Table 3-1 – Data set characteristics of EC2000 from 2000 to 2007

Year	Number of news items	Size	Size on Disk
2000	39.650	99,6 MB	190 MB
2001	44.027	136 MB	233 MB
2002	46.573	145 MB	245 MB
2003	49.032	134 MB	243 MB
2004	51.749	113 MB	234 MB
2005	49.738	113 MB	229 MB
2006	46.262	108 MB	217 MB
2007	10.925	27,2 MB	52,8 MB

Data characteristics of the training data set are given in Table 3-2.

Table 3-2 – Data set characteristics of the EC2000

Size	99,6 MB
Size on Disk	190 MB
Number of Lines	368.270
Number of Words	9.154.458
Number of Characters	54.579.390

The news in the EC2000 is provided in XML format, which comprises various tags as can be seen below:

```
<haber2000>
        <haber_id>News id of the news</haber_id>
        <kategori_id>Category id of the news</kategori_id>
        <il_isim>City of the news</il_isim>
        <tarih>Date of the news</tarih>
        <saat>Hour of the news</saat>
        <baslik>Title of the news </baslik>
        <icerik>Content of the news</icerik>
        <oncelik>Priority id of the news</oncelik>
</haber2000>
```

3.2.2 THE METU TC DATA SET

The METU TC is a balanced electronic corpus comprising written texts in Turkish, which was prepared by a team of academicians from METU and Sabancı University **[63]**. In building the METU TC, these academicians aimed at building a corpus of post-1990 written Turkish samples from various genres. They did not include any spoken component because of the lack of resources and experience. They decided that around 2.000.000 words are reasonable for their aim. After some interactions with publishers, they set their sample size to 2000 words (or whenever the last sentence finishes); including up to three samples taken from each source if its publisher allows it. The METU TC now comprises exactly 1.889.080 words, which were taken from 520 samples belonging to 291 different sources.

In **[63]**, authors of the METU TC stated that they tried to be as balanced as they could be, but they did not base their sampling on statistical sampling of all the works produced for their chosen period (post-1990) and did not base their sampling on whatever a typical reader in Turkey called written language. They used 14 main text types in the METU TC which are articles, essays, interviews, novels, research monographies, short stories, travels and other types such as biography, memoirs, personal development literature, and short columns. The distribution of the text types in the METU TC, which was taken from **[63]**, is given in Figure 3-1 in detail.

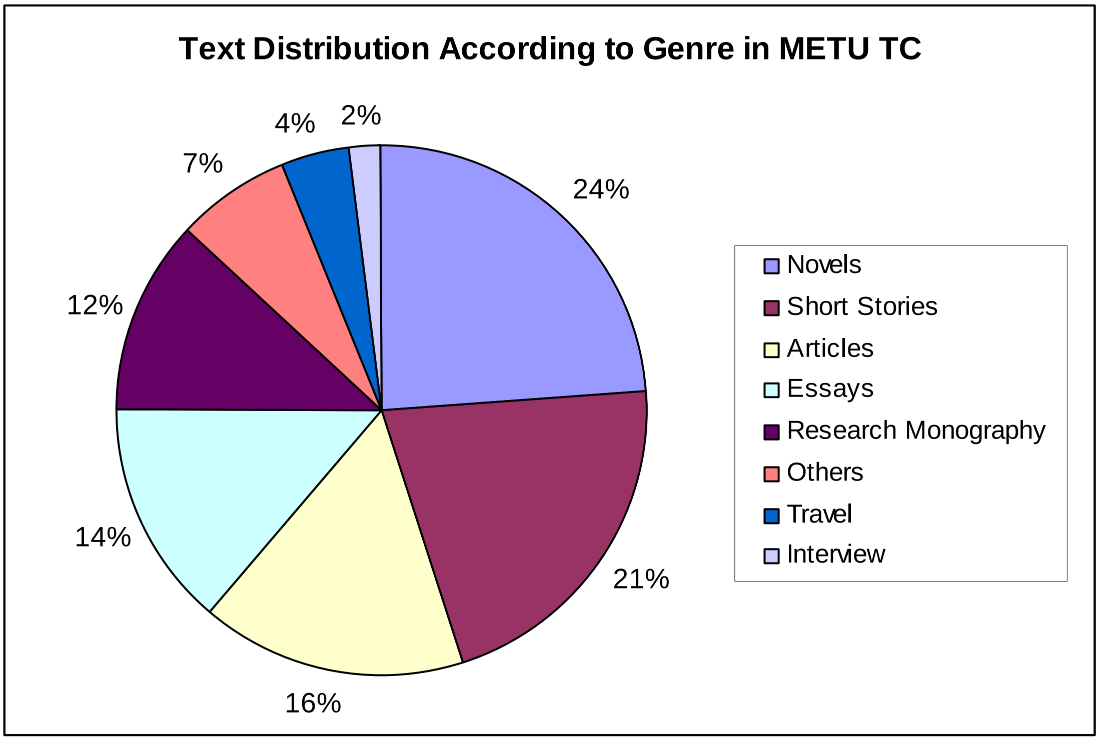


Figure 3-1 – Text distribution according to genre in the METU TC

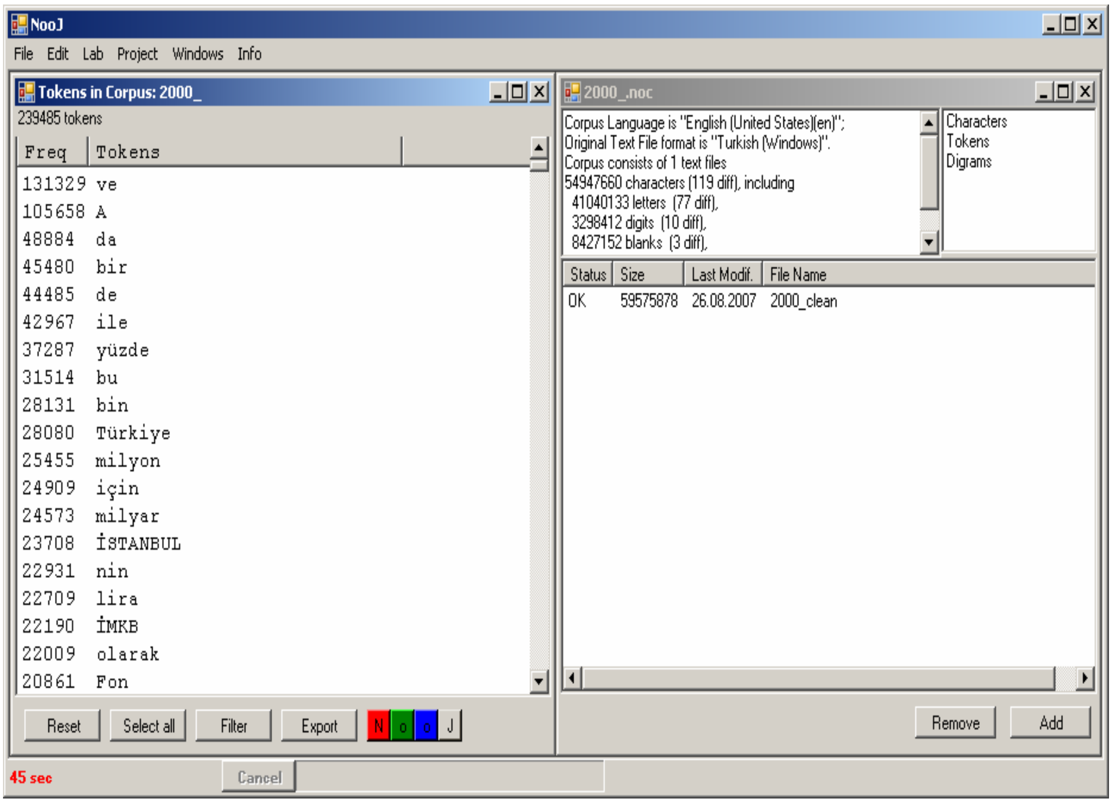
3.3 ENVIRONMENT

The Nooj software and the text analysis tool were mainly used in this thesis. We used the Nooj software to perform frequency analysis, concordance analysis, and draw local grammar graphs. The text analysis tool was used to perform collocation analysis.

3.3.1 THE NOOJ SOFTWARE

The Nooj software, a linguistic development environment used to construct descriptions of natural languages by parsing large sets of texts (large corpora) in real time, was employed in this thesis. It was developed on the .Net platform and runs on Microsoft Windows. It can build complex concordances, with respect to given patterns. Semantic units in large texts, such as names, easily identified and extracted by using the Nooj software. The Nooj software can process texts and corpora made of hundreds of text files. Lexical, syntactic and semantic annotations can be inserted in the text in cascade, without destroying the text **[64]**.

All tokens used in large texts with their frequencies can be found by using the Nooj software. They can be sorted both alphabetically and according to their frequencies. For example, Figure 3-2 shows the sorted word frequency list of the EC2000 January news.



239485 tokens

Freq	Tokens
131329	ve
105658	A
48884	da
45480	bir
44485	de
42967	ile
37287	yüzde
31514	bu
28131	bin
28080	Türkiye
25455	milyon
24909	için
24573	milyar
23708	İSTANBUL
22931	nin
22709	lira
22190	İMKB
22009	olarak
20861	Fon

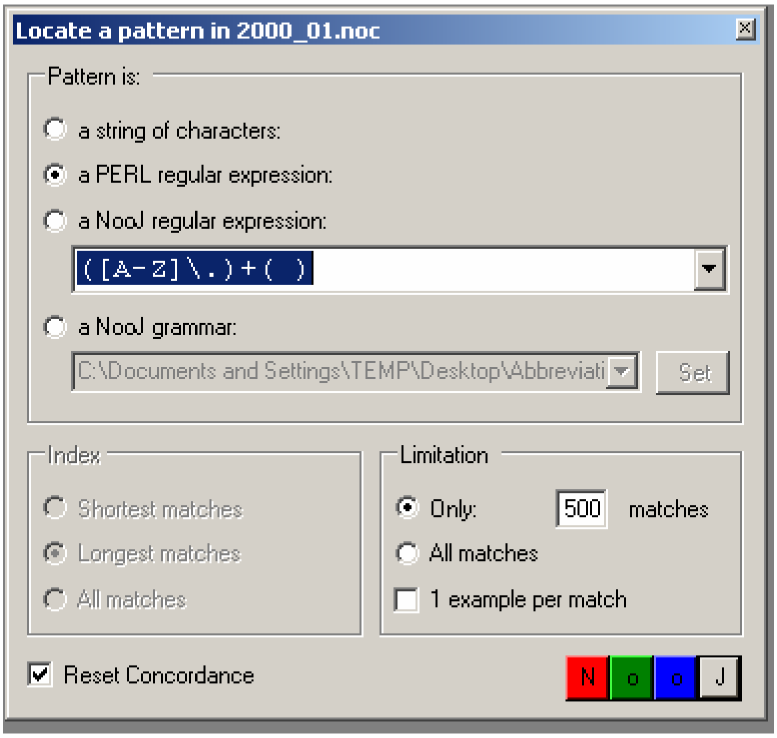
Corpus Language is "English (United States)(en)";
Original Text File format is "Turkish (Windows)".
Corpus consists of 1 text files
54947660 characters (119 diff), including
41040133 letters (77 diff),
3298412 digits (10 diff),
8427152 blanks (3 diff),

Characters
Tokens
Digrams

Status	Size	Last Modif.	File Name
OK	59575878	26.08.2007	2000_clean

Figure 3-2 – Sorted word frequency list of the EC2000 January news in the Nooj software

All concordance lines according to the given patterns can be obtained with the help of the Nooj software by using its “Locate a pattern” menu. The patterns for concordance analysis can be created in four ways (see Figure 3-3): by using a string of characters, a PERL regular expression, a Nooj regular expression, or a Nooj grammar.



Locate a pattern in 2000_01.noc

Pattern is:

○ a string of characters:

● a PERL regular expression:

○ a NooJ regular expression:

([A-Z]\.)+( )

○ a NooJ grammar:

C:\Documents and Settings\TEMP\Desktop\Abbreviati

Set

Index

○ Shortest matches

● Longest matches

○ All matches

Limitation

● Only: 500 matches

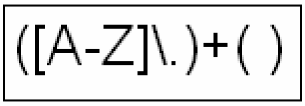
○ All matches

□ 1 example per match

☑ Reset Concordance

Figure 3-3 – Locate a pattern interface

For example, the abbreviation list of January in the EC2000 can be found by using the regular expression in Figure 3-4 through the “a PERL regular expression” option in “Locate a pattern” interface.



([A-Z]\.)+( )

Figure 3-4 – The Nooj regular expression of abbreviation pattern

The abbreviation list in January in the EC2000 can also be found by using the following the Nooj syntactic grammar in Figure 3-5 through the “a Nooj grammar” option in “Locate a pattern” interface.

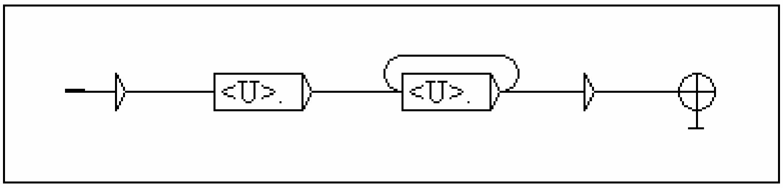
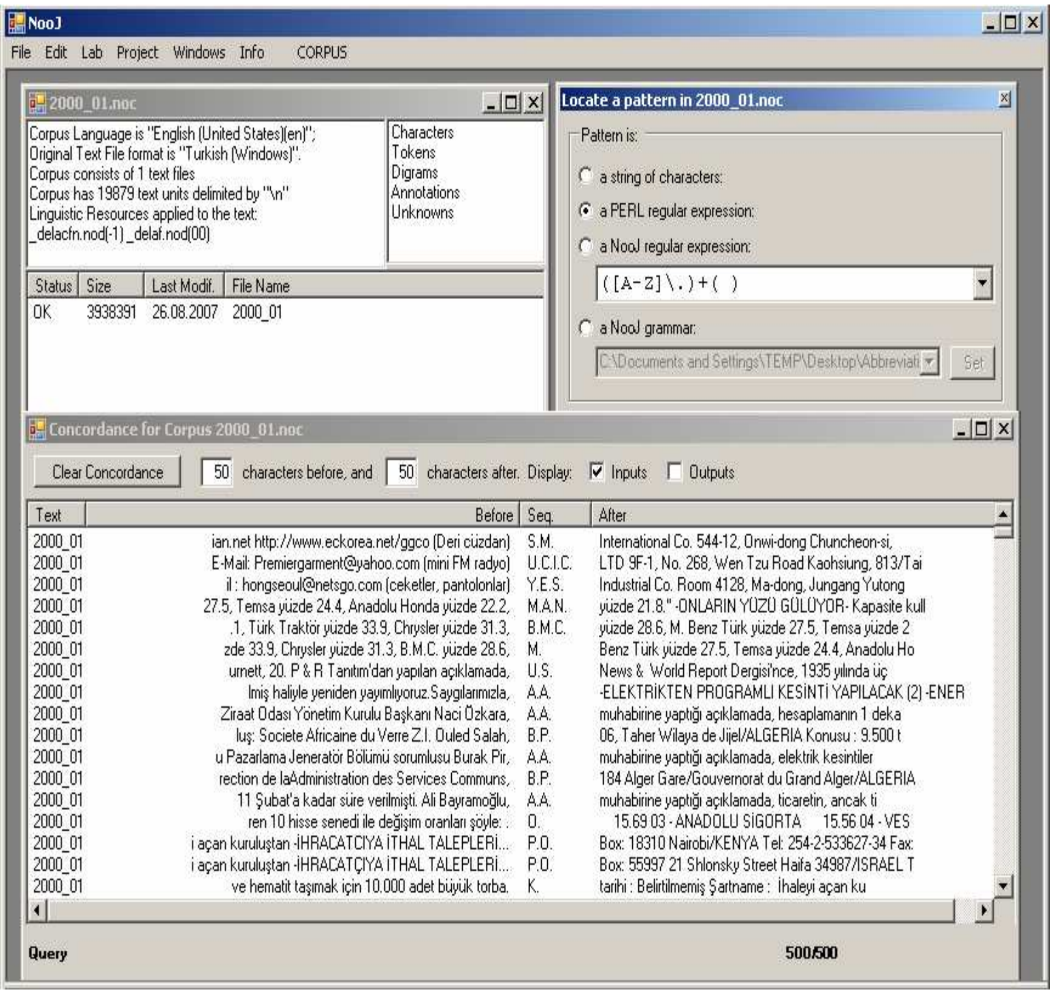


Figure 3-5 – The Nooj syntactic grammar of abbreviation pattern

By using the criteria given in Figure 3-5, all of the concordance lines in the abbreviation list can be found by using regular expression **“([A-Z].\)+( )”** from January news of the EC2000 in the Nooj software. The results can be seen in Figure 3-6.


NooJ
File Edit Lab Project Windows Info CORPUS
2000_01.noc
Characters
Tokens
Digrams
Annotations
Unknowns
Locate a pattern in 2000_01.noc
Set
Concordance for Corpus 2000_01.noc
Clear Concordance
Query

Corpus Language is "English (United States)(en)";
Original Text File format is "Turkish (Windows)".
Corpus consists of 1 text files
Corpus has 19879 text units delimited by "\n"
Linguistic Resources applied to the text:
_delacfn.nod(-1) _delaf.nod(00)

Status	Size	Last Modif.	File Name
OK	3938391	26.08.2007	2000_01

Pattern is:

○ a string of characters:
● a PERL regular expression:
○ a NooJ regular expression:
([A-Z]\.)+( )
○ a NooJ grammar:
C:\Documents and Settings\TEMP\Desktop\Abbreviati

50 characters before, and 50 characters after. Display: ☑ Inputs □ Outputs

Text	Before	Seq.	After
2000_01	ian.net http://www.eckorea.net/ggco (Deri cüzdan)	S.M.	International Co. 544-12, Onwi-dong Chuncheon-si,
2000_01	E-Mail: Premiergarment@yahoo.com (mini FM radyo)	U.C.I.C.	LTD 9F-1, No. 268, Wen Tzu Road Kaohsiung, 813/Tai
2000_01	il : hongseoul@netsgo.com (ceketler, pantolonlar)	Y.E.S.	Industrial Co. Room 4128, Ma-dong, Jungang Yutong
2000_01	27.5, Temsa yüzde 24.4, Anadolu Honda yüzde 22.2,	M.A.N.	yüzde 21.8." -ONLARIN YÜZÜ GÜLÜYOR- Kapasite kull
2000_01	.1, Türk Traktör yüzde 33.9, Chrysler yüzde 31.3,	B.M.C.	yüzde 28.6, M. Benz Türk yüzde 27.5, Temsa yüzde 2
2000_01	zde 33.9, Chrysler yüzde 31.3, B.M.C. yüzde 28.6,	M.	Benz Türk yüzde 27.5, Temsa yüzde 24.4, Anadolu Ho
2000_01	urnett, 20. P & R Tanıtım'dan yapılan açıklamada,	U.S.	News & World Report Dergisi'nce, 1935 yılında üç
2000_01	lmiş haliyle yeniden yayımlıyoruz.Saygılarımızla,	A.A.	-ELEKTRİKTEN PROGRAMLI KESİNTİ YAPILACAK (2) -ENER
2000_01	Ziraat Odası Yönetim Kurulu Başkanı Naci Özkara,	A.A.	muhabirine yaptığı açıklamada, hesaplamanın 1 deka
2000_01	luş: Societe Africaine du Verre Z.I. Ouled Salah,	B.P.	06, Taher Wilaya de Jijel/ALGERIA Konusu : 9.500 t
2000_01	u Pazarlama Jeneratör Bölümü sorumlusu Burak Pir,	A.A.	muhabirine yaptığı açıklamada, elektrik kesintiler
2000_01	rection de laAdministration des Services Communs,	B.P.	184 Alger Gare/Gouvernorat du Grand Alger/ALGERIA
2000_01	11 Şubat'a kadar süre verilmişti. Ali Bayramoğlu,	A.A.	muhabirine yaptığı açıklamada, ticaretin, ancak ti
2000_01	ren 10 hisse senedi ile değişim oranları şöyle: .	O.	15.69 03 - ANADOLU SİGORTA 15.56 04 - VES
2000_01	i açan kuruluştan -İHRACATCIYA İTHAL TALEPLERİ...	P.O.	Box: 18310 Nairobi/KENYA Tel: 254-2-533627-34 Fax:
2000_01	i açan kuruluştan -İHRACATÇIYA İTHAL TALEPLERİ...	P.O.	Box: 55997 21 Shlonsky Street Haifa 34987/ISRAEL T
2000_01	ve hematit taşımak için 10.000 adet büyük torba.	K.	tarihi : Belirtilmemiş Şartname : İhaleyi açan ku

500/500

Figure 3-6 – Concordance lines of abbreviation list

3.3.2 THE TEXT ANALYSIS TOOL

In this thesis, we also used the Text Analysis **[65]** tool which was developed by University of Surrey in order to carry out collocation analysis based on Smadja's work **[45]**. Actually, this tool uses spread and strength criteria defined in **[45]**. Strength eliminates infrequent collocations by assuming that the frequency of the neighboring word of a lexical item must be at least one standard deviation above the average frequency so that the neighboring word can be considered as collocation of this lexical item. The formula of strength is given Equation 3-1:

$$\boldsymbol{strength} = \frac{\boldsymbol{freq_i} - \overline{\boldsymbol{f}}}{\sigma} \geq \boldsymbol{k_0}$$

Equation 3-1 - Strength

where;

$\boldsymbol{freq_i}$ = sum of the frequency of a lexical item among *n* neighborhood

$\overline{\boldsymbol{f}}$ = average frequency of all neighboring words that appear within the neighborhood of a lexical item

σ = standard deviation of frequency of all neighboring words of a lexical item

$\boldsymbol{k_0}$ = K-score threshold

Spread is used whether to reject or accept collocations within the neighborhood. It is required that the histogram of the 10 relative frequencies of the neighboring word around a lexical item has at least one spike, which is the sharp turning point in the histogram. In other words, if the histogram does not contain any spike, these collocations will not be considered as significant collocates of this lexical item. The formula of spread is given Equation 3-2:

$$spread = \frac{\sum_{j=1}^{n} (f_i^{\,j} - \overline{f_i})^2}{n} \geq U_0$$ Equation 3-2 - Spread

where;

f_i = frequencies of neighboring word in position *j* for a lexical item in position *i*

$\overline{f}$ = average frequency neighboring word in position *j* among *n* neighborhood

n = number of neighborhood, normally *n* = 10

U_0 = U-score threshold, normally U_0 = 10

We used this text analysis tool in order to perform collocation analysis for RVs. In our study, we used default values which are 10 and 1 for strength and spread, respectively, during the collocation analysis.

3.4 METHODOLOGY

In the literature chapter, the approaches that are used in NER were described in detail. We based our approach on the bootstrap method. As aforementioned, we actually used the LG approach to find NEs in the financial news in Turkish. In this section, we will explain the steps undertaken for NEE by using the LG approach in detail.

3.4.1 PREPROCESSING THE EC2000 AND WORD FREQUENCY ANALYSIS

The first step is to compute the word frequencies in the EC2000. Figure 3-7 shows all the steps in the process of finding frequencies of words.

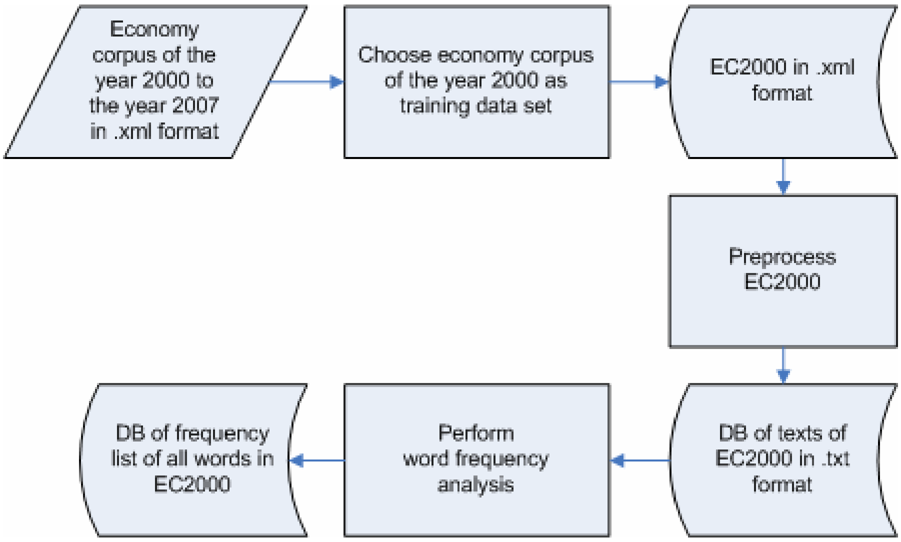

Economy corpus of the year 2000 to the year 2007 in .xml format
Choose economy corpus of the year 2000 as training data set
EC2000 in .xml format
Preprocess EC2000
DB of texts of EC2000 in .txt format
Perform word frequency analysis
DB of frequency list of all words in EC2000

Figure 3-7 – First step is to find word frequencies in the EC2000

There are three major steps in the analysis:

- ***Preprocessing the EC2000***: This sub-step has three sub-steps which are cleaning and splitting the EC2000.

    - *Converting the EC2000 from XML format into TXT format*: As the news is stored in XML format, they are required to be cleaned from tags (e.g. news_id tag, category_id tag, city_name tag, date tag, time tag, and priority tag) for accurate word frequency calculation. A simple Java filter was implemented to remove all the tags from the news and to store the untagged title along with the content of news in text format.

    - *Cleaning the EC2000*: All unwanted characters such as "&nbsp", many dashes one after another ("-----"), many full stops one after another ("….."), and spaces were removed. The method in sentence splitter suggests that the statements made by people should be enclosed with quotation marks ("). However, in the EC2000, many sentences start with two apostrophes ('') instead of quotation marks, which is clearly a mistake. Sentence splitter cannot find these mistaken characters. Therefore, these mistaken characters should be changed with quotation marks (“). We replaced these characters with quotation marks by using simple Java filter.

- *Sentence Splitting*: After removing the unwanted characters, we split the text files into sentences. The sentence splitter method is built upon LG based approach of Friburger **[42]**. It was first designed for French texts and also adapted to English texts **[1]**. In this thesis, the method was tailored for Turkish texts.

  The sentence splitter method uses a transducer that recognizes the end of a sentence through punctuation marks and puts the mark {S} between adjacent sentences. However, there are some cases in which it is very difficult to split the sentence successfully. The difficulty mainly comes from the dot when it is followed by an upper case letter because this dot can either be a full stop ending the sentence or not. In other words, there are some ambiguities in this type of cases which we explained.

  We found four types of ambiguities with the dot followed by an upper case letter in Turkish:

  - Person names can be preceded by an abbreviated title (e.g. Prof. Dr. Ural Akbulut).
  - Person names can include abbreviated word forms (e.g. A. Nejdet Sezer).
  - Organization names can be in abbreviated form (e.g. A.A. which stands for Anadolu Agency).
  - Abbreviations which are used after the name of a company (e.g. A.Ş. or T.A.Ş)

We split the texts into sentences by considering these four ambiguities. There are also other difficulties: for example the method suggests that sentences always start with capital letters. However, in the EC2000, we realized that there are sentences, which start with numeric characters. Consequently, the method was updated as to include numeric characters at the beginning of the sentences. The graph of the sentence splitter that we used is given in the Figure 3-8 below.

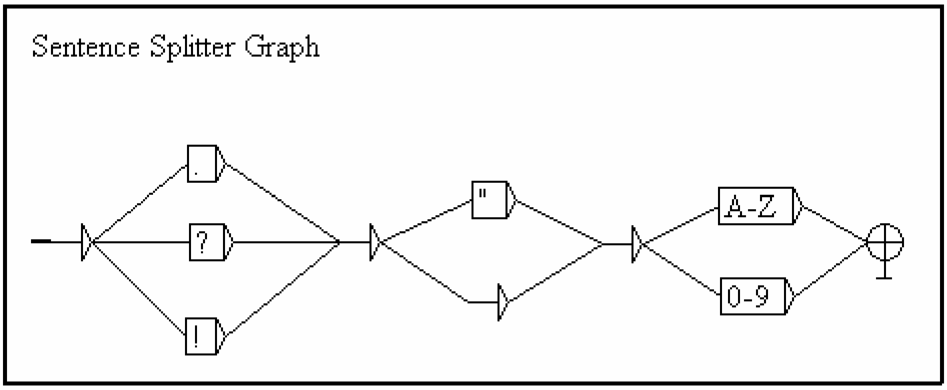


Figure 3-8 – Graph of sentence splitter

- ***Construct cleaned and split the EC2000***: Final corpus (hereafter called “FEC2000”) is 99,6 MB, which includes 368.270 lines, 9.154.458 words, and 54.579.390 characters.

- ***Word frequency analysis***: We carried out word frequency analysis by using the Nooj software on the FEC2000**.** The most frequent top 1000 words can be seen in the **Appendix A**.

3.4.2 EXTRACTION OF TURKISH REPORTING VERBS

The second step is to identify Turkish RVs. NEs can be obtained by observing their behaviors in the sentences of texts. Because RVs are verbs to report the speech of others and used for reporting what someone says, thinks or believes, person names frequently occur in the clauses of RVs such as “de–”, “belirt–”, “bildir–”, “söyle–”, “degerlendir–”, “sun–”, “sürdür–”, “anlat–”, and “hatırlat–” in Turkish. Detailed list of top 100 RVs in Turkish, which are sorted according to their frequencies in reference corpus, are given in **Appendix B**.

Figure 3-9 shows all the processes carried out for Turkish RVs identification.

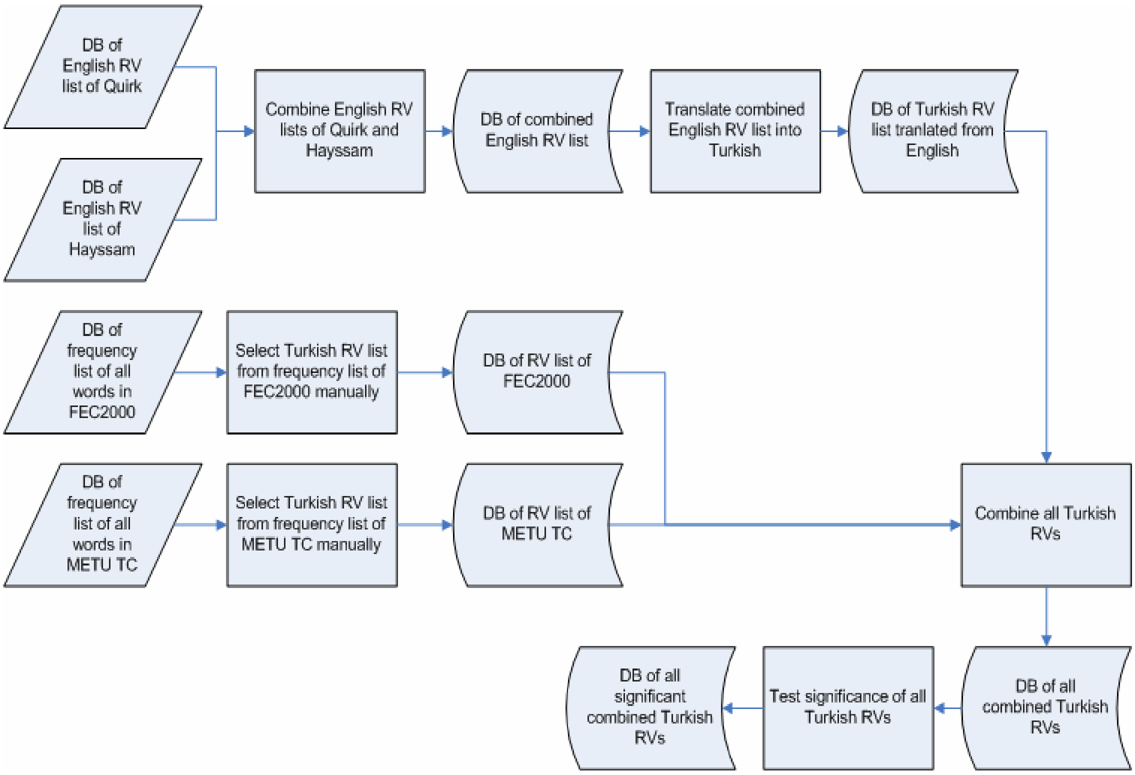

DB of English RV list of Quirk
DB of English RV list of Hayssam
Combine English RV lists of Quirk and Hayssam
DB of combined English RV list
Translate combined English RV list into Turkish
DB of Turkish RV list tranlated from English
DB of frequency list of all words in FEC2000
Select Turkish RV list from frequency list of FEC2000 manually
DB of RV list of FEC2000
DB of frequency list of all words in METU TC
Select Turkish RV list from frequency list of METU TC manually
DB of RV list of METU TC
Combine all Turkish RVs
DB of all combined Turkish RVs
Test significance of all Turkish RVs
DB of all significant combined Turkish RVs

Figure 3-9 – Second step is to find all significant Turkish reporting verbs

There are four major sub-steps:

- ***Obtain the list of Reporting Verbs in English***: Quirk et. al provides a list of RVs in **[2]**. As there is no similar study on Turkish in the literature, we followed two procedures to construct a list of Turkish RVs. First, Quirk’s RVs list is translated into Turkish by using The Red house Turkish-English dictionary **[66]**. Second, we translated RVs list of Hayssam, who used it for his thesis in **[1]**, into Turkish by using the same procedure. We also used Hayssam's list of RVs because he found different RVs that are not included in the Quirk's list of RVs but were used in financial English texts.

- ***Combine the two lists of Reporting Verbs in English****:* We combined Turkish RVs translated from the English RVs lists of Quirk et al. with that of Hayssam.

- ***Create a list of Turkish Reporting Verbs in the* FEC2000 *and the METU TC****:* Top 6000 words from both frequency list of the FEC2000 and the METU TC are examined manually to find RVs in Turkish in detail. We extracted Turkish RVs from totally 12000 words (6000 words from each corpus) and acquired a list of Turkish RVs manually.

- ***Combine all the lists of Turkish Reporting Verbs***: We combined all three lists: Turkish RVs in the FEC2000, the METU TC and Turkish RVs translated from English. We considered all word forms for the construction of the list of RVs. The frequencies of top 100 RVs extracted from train corpus and reference corpus can be seen in **Appendix B**.

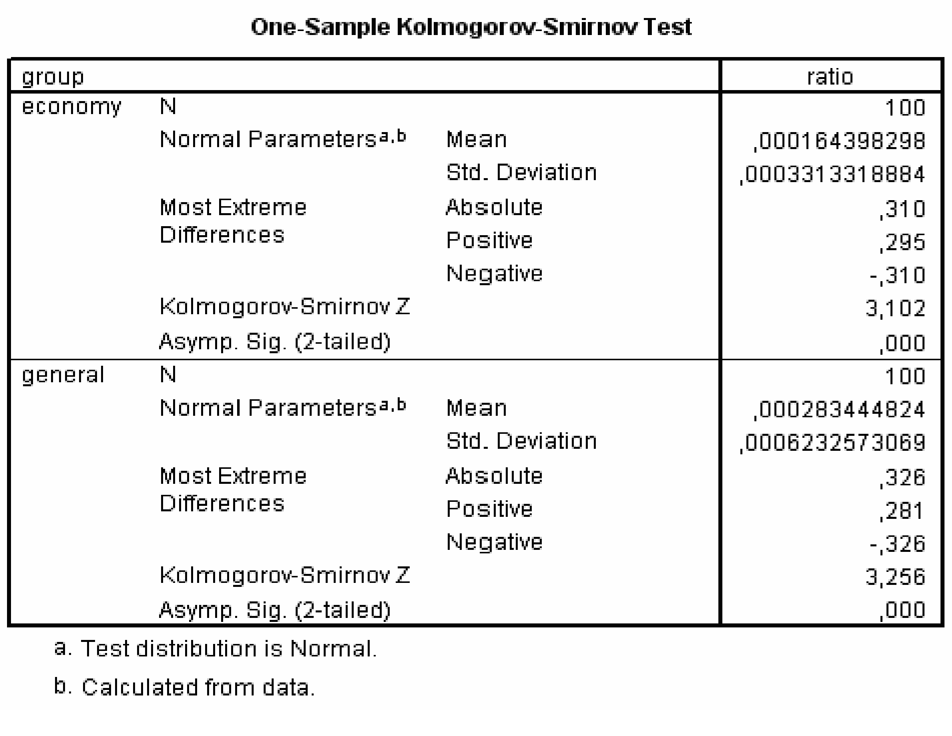
As can be seen in **Appendix B**, it is difficult to understand whether there is a significant difference between the frequencies of RVs in both corpora. Hence, a number of statistical tests should be carried out. These statistical tests will be explained in the following chapter in detail.

3.4.3 SIGNIFICANCE TEST OF TURKISH REPORTING VERBS

Every specific domain language has a set of frequently used linguistic patterns. RVs usage in these linguistic patterns is not very similar. Consequently, it can be concluded that the usage of RVs changes according to the domain in which they are used. Therefore, the differences in significance between the FEC2000 and the METU TC in terms of RVs frequency should be tested and the patterns in the FEC2000 should be found by considering the RVs that affect the significant differences between two corpora.

A number of statistical tests were carried out to understand whether there is significant difference between the frequencies of RVs in both corpora. First, the one-sample Kolmogorov Smirnov test is used to find out whether the distribution of Turkish RVs in both corpora is normally distributed or not. The result of the one-sample Kolmogorov Smirnov test can be seen in Table 3-3 below.

Table 3-3 – The one-sample Kolmogorov-Smirnov non-parametric test



One-Sample Kolmogorov-Smirnov Test

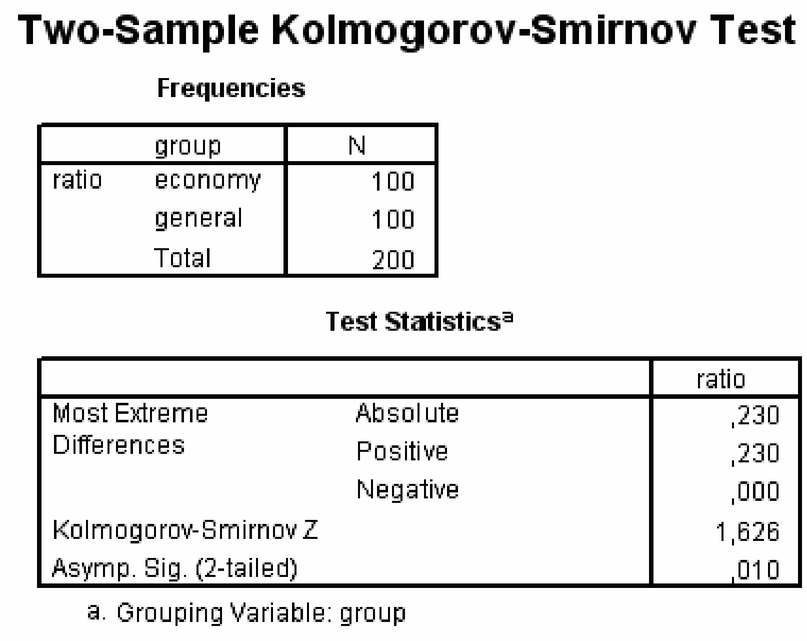
group			ratio
economy	N		100
	Normal Parameters[a,b]	Mean	,000164398298
		Std. Deviation	,0003313318884
	Most Extreme Differences	Absolute	,310
		Positive	,295
		Negative	-,310
	Kolmogorov-Smirnov Z		3,102
	Asymp. Sig. (2-tailed)		,000
general	N		100
	Normal Parameters[a,b]	Mean	,000283444824
		Std. Deviation	,0006232573069
	Most Extreme Differences	Absolute	,326
		Positive	,281
		Negative	-,326
	Kolmogorov-Smirnov Z		3,256
	Asymp. Sig. (2-tailed)		,000

a. Test distribution is Normal.

b. Calculated from data.

In the one-sample Kolmogorov Smirnov non-parametric test, our null hypothesis is that the distribution of RVs in each corpus fits the normal distribution. The decision about the hypothesis can be made according to the Kolmogorov Smirnov Z test statistic value and asymptotic significance value. We used α = 0.05 significance level in the one-sample Kolmogorov Smirnov non-parametric test. If asymptotic significance is smaller then α/2 = 0.025, we reject the null hypothesis. As seen in Table 3-3, asymptotic significance level for both the FEC2000 and the METU TC is equal to zero which is smaller than α/2 = 0.025. Therefore, it can be concluded that the distribution of RVs in each corpus does not fit with normal distribution. As a consequence, non-parametric tests such as the two-sample Kolmogorov Smirnov test and the Mann-Whitney test should be chosen to test whether RVs in both two corpora are coming from same distribution or not.

Table 3-4 – The two-sample Kolmogorov-Smirnov non-parametric test



Two-Sample Kolmogorov-Smirnov Test

Frequencies

	group	N
ratio	economy	100
	general	100
	Total	200

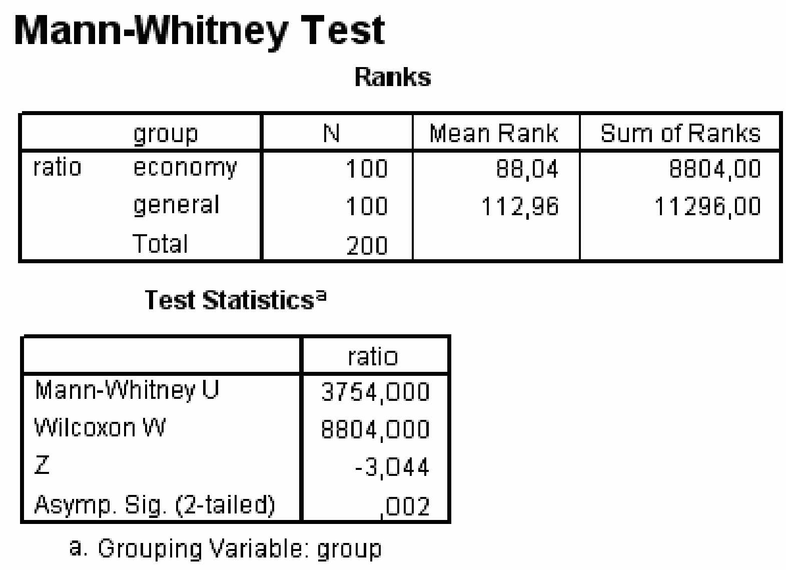
Test Statistics[a]

		ratio
Most Extreme Differences	Absolute	,230
	Positive	,230
	Negative	,000
Kolmogorov-Smirnov Z		1,626
Asymp. Sig. (2-tailed)		,010

a. Grouping Variable: group

In the two-sample Kolmogorov Smirnov non-parametric test, our null hypothesis is that the distribution of RVs in each corpus comes from the same distribution. The decision about the hypothesis can be made according to the Kolmogorov Smirnov Z test statistic value and asymptotic significance value. We used α = 0.05 significance level in the two-sample Kolmogorov Smirnov non-parametric test. If asymptotic significance is smaller then α/2 = 0.025, we reject the null hypothesis. As seen in Table 3-4, asymptotic significance level is equal to 0.010 which is smaller than α/2 = 0.025. Therefore, it can be concluded that the distribution of RVs in each corpus does not fit with the same distribution.

Table 3-5 – The Mann-Whitney non-parametric test



Mann-Whitney Test

Ranks

	group	N	Mean Rank	Sum of Ranks
ratio	economy	100	88,04	8804,00
	general	100	112,96	11296,00
	Total	200		

Test Statistics^a

	ratio
Mann-Whitney U	3754,000
Wilcoxon W	8804,000
Z	-3,044
Asymp. Sig. (2-tailed)	,002

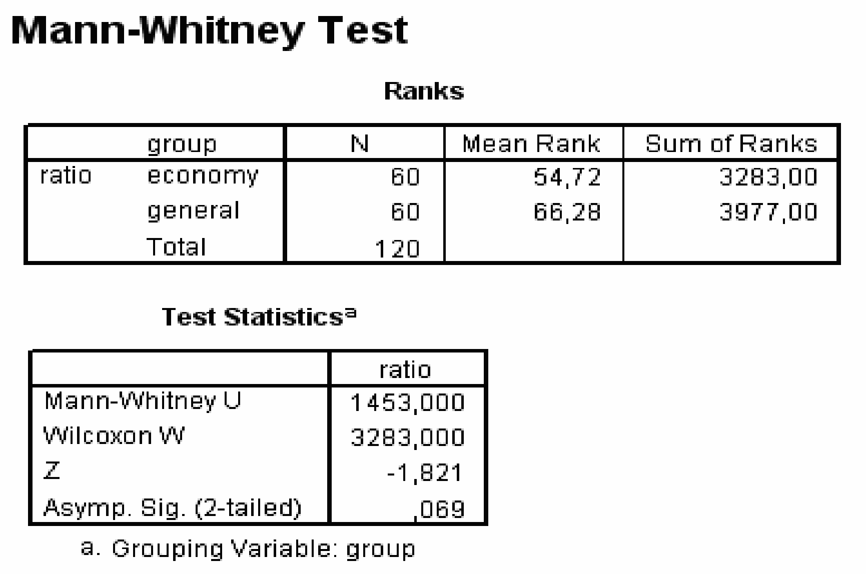
a. Grouping Variable: group

In the Mann-Whitney non-parametric test, our null hypothesis is again that the distribution of RVs in each corpus comes from the same population. Decision about the hypothesis can be made according to the same criteria used in the two-sample Kolmogorov Smirnov non-parametric test. We used α = 0.05 significance level in the Mann-Whitney non-parametric test. As seen in Table 3-5, asymptotic significance level is equal to 0.02 which is smaller than α/2 = 0.025. Therefore, it can be concluded that the distribution of RVs in each corpus does not come from the same population.

As a consequence, we concluded that the distributions of RVs in the FEC2000 and the METU TC do not have same distribution. To find which words cause the significant difference between two corpora in terms of RVs, Maximum Likelihood statistical test is applied as in **[67]**. Top 100 Turkish RVs which are sorted by their maximum likelihood values are given in **Appendix C**.

As stated in **[67]**, words at the top of the list, which is sorted by their maximum likelihood values, represent the most significant relative frequency difference between the two corpora. In other words, the words which are the most significant ones in one corpus, as compared to the other corpus, appear at the top of the list. On the other hand, words having similar frequencies in the two corpora appear at the bottom of the list. Therefore, we can focus on the list at the top with RVs which have larger maximum likelihood values in case that they can cause the significant differences between the two corpora. To find which words cause the significant difference between two corpora in terms of RVs, we iteratively remove RVs from the top of the list until the Mann-Whitney non-parametric test accepted the hypothesis which shows both corpora have the same distribution. After removing top 40 RVs from the list, the Mann-Whitney non-parametric test accepts the null hypothesis. The result of the Mann-Whitney non-parametric test after removing top 40 RVs can be seen in the Table 3-6.

Table 3-6 – The Mann-Whitney non-parametric test after removing top forty reporting verbs



Mann-Whitney Test

Ranks

	group	N	Mean Rank	Sum of Ranks
ratio	economy	60	54,72	3283,00
	general	60	66,28	3977,00
	Total	120		

Test Statistics[a]

	ratio
Mann-Whitney U	1453,000
Wilcoxon W	3283,000
Z	-1,821
Asymp. Sig. (2-tailed)	,069

a. Grouping Variable: group

In the Mann-Whitney non-parametric test, our null hypothesis is that the distributions of RVs in each corpus from which top 40 RVs are removed come from the same population. We used α = 0.05 significance level in the Mann-Whitney non-parametric test. As seen in Table 3-6, asymptotic significance level is equal to 0.69 which is greater than α/2 = 0.025. Therefore, it can be concluded that the distribution of RVs in each corpus comes from same population.

As a result, we concluded that top 40 RVs cause the significant difference between two corpora. Therefore, we took into consideration these 40 RVs to extract NEs from the texts.

3.5 NAMED ENTITY EXTRACTION

As I mentioned before, in this thesis, we tried to find person names in the Turkish financial texts and applied several steps to achieve our purpose. The third step in our approach is to find person names with the LG approach by basing on the list of Turkish RVs.

The third step contains four major processes which are;

- Concordance Analysis
- Collocation Analysis
- Pattern Generation
- Person Name Extraction

Figure 3-10 shows all the processes carried out for the recognition of person names in detail.

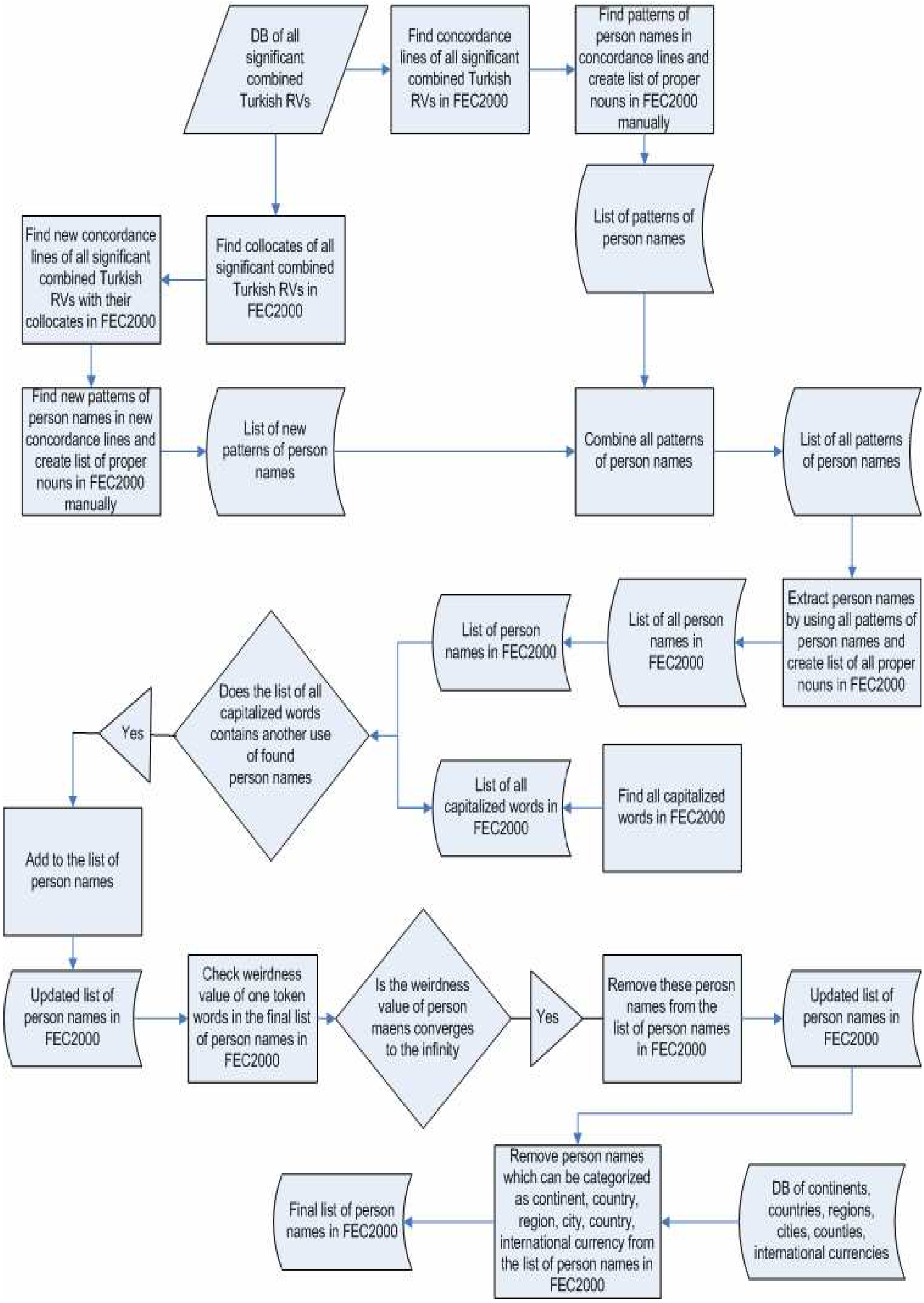

DB of all significant combined Turkish RVs
Find concordance lines of all significant combined Turkish RVs in FEC2000
Find patterns of person names in concordance lines and create list of proper nouns in FEC2000 manually
List of patterns of person names
Find collocates of all significant combined Turkish RVs in FEC2000
Find new concordance lines of all significant combined Turkish RVs with their collocates in FEC2000
Find new patterns of person names in new concordance lines and create list of proper nouns in FEC2000 manually
List of new patterns of person names
Combine all patterns of person names
List of all patterns of person names
Extract person names by using all patterns of person names and create list of all proper nouns in FEC2000
List of all person names in FEC2000
List of person names in FEC2000
Does the list of all capitalized words contains another use of found person names
Yes
List of all capitalized words in FEC2000
Find all capitalized words in FEC2000
Add to the list of person names
Updated list of person names in FEC2000
Check weirdness value of one token words in the final list of person names in FEC2000
Is the weirdness value of person maens converges to the infinity
Yes
Remove these perosn names from the list of person names in FEC2000
Updated list of person names in FEC2000
Remove person names which can be categorized as continent, country, region, city, country, international currency from the list of person names in FEC2000
DB of continents, countries, regions, cities, counties, international currencies
Final list of person names in FEC2000

Figure 3-10 – Third step is to extract person names in the FEC2000

There are five major steps to extract person names from the FEC2000. These steps are described in detail below;

3.5.1 CONCORDANCE ANALYSIS

- *Concordance Analysis:* We first carried out concordance analysis for the FEC2000 and obtained concordance lines by using the Nooj software. During the concordance analysis, we used three forms of RVs. These forms are extracted from the train set by analyzing the concordance lines around person names and reporting verbs manually. Three forms of some RVs (e.g. “de-”, “belirt-”, “kaydet-”, “açıkla-”, “ifade et-”, “söyle-”, and “bildir-”) can be seen below:

    - **Tense and person marked reporting verbs:** Examples of past tense RVs are “dedi”, “belirtti”, “kaydetti”, “açıkladı”, “ifade etti”, “söyledi”, and “bildirdi”. Hereafter, we refer tense and person marked reporting verbs as **RV_Form_1**.

    - **Subject relativized reporting verbs with suffix –An**: Examples of RVs ending with the “-en” or “-an” suffix (shown as –An) are “diyen”, “belirten”, “kaydeden”, “açıklayan”, “ifade eden”, “söyleyen”, and “bildiren”. Hereafter, we refer subject relativized reporting verbs with the suffix –An as **RV_Form_2**.

    - **Reporting verbs marked with –ArAk:** Examples of RVs ending with the “-erek” or “-arak” suffix (shown as –ArAk) are “diyerek”, “belirterek”, “kaydederek”, “açıklayarak”, “ifade ederek”, “söyleyerek”, and “bildirerek”. Hereafter, we refer reporting verbs marked with –ArAk as **RV_Form_3**.

  We will use the notations in Table 3-7 for the descriptions of the patterns that contain person names.

Table 3-7 – Notations which are used in the descriptions of patterns

PN	Person name
W	Word
W*	Word Sequence
<E>	Empty string
Title	Title of a person name
RV_Form_1	Tense and person marked reporting verbs
RV_Form_2	Subject relativized reporting verbs with suffix –An
RV_Form_3	Reporting verbs marked with –ArAk
RV_Form_4	Reporting verbs with an pronominal complement

- *Find the patterns of Person Names from the concordance lines of reporting verbs that belong to RV_Form_1:* The concordance lines of RV_Form_1 can be seen in the **Appendix E**. These lines are grouped as in Table 3-8, Table 3-9, Table 3-10, and Table 3-11.

The common pattern in Table 3-8 can be defined as:

[W*] + [PN] + [,] + [W*] + [RV_Form_1] + [.]

Table 3-8 – Pattern 1 for RV_Form_1

The Concordance line	**The material containing the person name**	**The Splitter character**	**The material preceding the reporting verb**	**The reporting verb**
1	Hükümetin uyum ve istikrar sergilediğini belirten **Arsan**	,	bu faktörün ülkenin ekonomik ve sosyal alanda gelişmesini hızlandıracağını	açıkladı.
3	Çin Devlet Kalkınma Planlaması Komitesi Başkanı **Zeng Peiyan**	,	Çin'in yakın bir gelecekte, para birimi yuanda devalüasyon uygulamasına gerek olmadığını	belirtti.
14	Sivas'ın Türkiye'nin en mamur şehirlerinden biri olduğunu belirten **Demirel**	,	Sivaslı işadamlarının il'e yatırım yapmalarını sağlayabilmek amacıyla düzenlenecek kurultaya kendisinin de katılacağını	kaydetti.

Common pattern in Table 3-9 can be defined as:

[PN] + [,] + [W*] + [RV_Form_1] + [.]

Table 3-9 – Pattern 2 for RV_Form_1

Concordance line	Person name	Splitter character	The material preceding reporting verb	Reporting verb
9	Düzel	,	kurmayı planladıkları fabrikada yılda 10 bin asansör makinesi, 100 bin kilit ve diktatör, 3 bin 500 kabin, 20 bin kapı ve 80 bin buton üretmeyi hedeflediklerini	ifade etti.
10	Bayram	,	şu ana kadar yapılan denetimlerde, mali durumu bozuk başka bir bankanın tespit edilmediğini	ifade etti.
11	Işık	,	bu satıştan elde edilen geliri, Işıklar İnşaat Malzemeleri A.Ş'nin yurtiçi tuğlaya yönelik yatırımlarının finansmanında kullanacaklarını	ifade etti.
12	Tahsin Sancak	,	A.A muhabirine yaptığı açıklamada, çay sektörünün reforma tabi tutularak, dünyaya açılması gerektiğini	kaydetti.
13	Hasan Özmen	,	pamuğun 206.4 trilyon ve yüzde 29'luk payla birinciliği koruduğuna dikkati çekerek, bitkisel yağların 96.5 trilyon liralık işlemle toplam işlem hacminin yüzde 13.6'sına sahip olduğunu	kaydetti.

Common pattern in Table 3-10 can be defined as:

[PN] + [W*] + [,] + [W*] + [RV_Form_1] + [.]

Table 3-10 – Pattern 3 for RV_Form_1

Concordance line	**The material containing person name**	**Splitter character**	**The material preceding reporting verb**	**Reporting verb**
8	**Şimşek** yaptığı açıklamada	,	"1998 yılında yaşanan sel felaketleri nedeniyle pamuk üreticisi zaten verimli ürün alamadı. Beklentimiz olan 20 centlik primin yanına dahi yaklaşılmadı"	dedi.

Common pattern in can be defined as:

[Title] + [PN] + [, | (<E>ise)|(<E>de)] + [W*] + [RV_Form_1] + [.]

Table 3-11 – Pattern 4 for RV_Form_1

Concordance line	**The material containing person name with title**	**Splitter character**	**The material preceding reporting verb**	**Reporting verb**
4	İzmir Kahveciler Odası Başkanı **Mustafa Ak**	,	yaptığı açıklamada, İzmir'deki kahvehanelerde uygul anacak fiyat listesinin 3 Ocak Pazartesi gününden itibaren geçerli olduğunu	bildirdi.
5	Trabzon Ticaret ve Sanayi Odası (TTSO) Yönetim Kurulu Başkanı **Şadan Eren**	,	konaklama tesisleri ile yörede konferans turizminin de geliştirileceğini	bildirdi.
6	Müstakil Sanayici ve İşadamları Derneği (MÜSİAD) Genel Başkanı **Ali Bayramoğlu**	,	ya pısal reformlar gerçekleştirilmeden, 2000 yılında öngörülen enflasyon hedeflerine ulaşılamayacağını	bildirdi.
7	Aksaray Valisi **Emir Durmaz**	,	"Aksaray'ın geri kalmışlıktan kurtulması, emsal iller seviyesine yükselmesi için süper teşvikli iller arasına alınması gerekir"	dedi.

Table 3-11 (continued)

Concordance line	The material containing person name with title	Splitter character	The material preceding reporting verb	Reporting verb
19	Castrol Türkiye Genel Müdürü **Ömer Dormen**	ise	bu işbirliğinin kalite, performans ve teknolojiyi ön plana çıkaracağını	ifade etti.
20	Cumhurbaşkanı **Demirel**	de	"Bu paneli yapmanız gayet iyi olur. Bakalım, benim zamanıma uyuyorsa katılırım"	dedi.

We combined all the patterns containing RV_Form_1 and created the final pattern called "RV_Form_1 final pattern" which is given as:

[Title]? + [W*]? + [PN] + [W*]? + [, | (<E>ise) | (<E>de)] + [W*] + [RV_Form_1] + [.]

where special character **?** makes the preceding token optional i.e. the pattern may contain the token followed by **?** or not.

However, the list of concordance line which is given in **Appendix E** does not contain the concordance lines of all Turkish RVs. It rather contains only a part of the concordance lines for some of Turkish RVs. After finding concordance lines of all Turkish RVs, we updated the "RV_Form_1 final pattern" for the sentence that contains RV_Form_1. It is called "RV_Form_1 updated final pattern" and is given below:

[Title]? + [W*]? + [PN] + [W*]? + [, | (<E>ise) | (<E>de) | (<E>da)] + [W*] + [RV_Form_1] + [.]

The graph of the "RV_Form_1 updated final pattern" can be seen in the Figure 3-12.

It should be stated that there are common nouns which are used instead of person names. These words are open class words (OCWs), which are nouns, adjectives and adverbs used in technical or specialist languages **[35]**. These OCWs such as those in Table 3-12, Table 3-15, and Table 3-19 will be ignored in our analysis.

Table 3-12 – Open class word pattern for RV_Form_1

Concordance line	**Person name (Open class word)**	**Splitter character**	**The material preceding reporting verb**	**Reporting verb**
18	BIS **yetkilisi**	,	bütün bunlara rağmen, 2000 Yılı Problemi'ne ilişkin asıl sonucun, gelecek hafta bankaların açılmasıyla ortaya çıkacağını	söyledi.

We tried to eliminate these OCWs based on their weirdness values in the METU TC. Each word found as a person name is checked against the words in the METU TC. As the weirdness value of the word we found goes to infinity, the word is assumed to be a person name. In other words, each initially capitalized string not matching any words in the METU TC is assumed to be a person name. We considered the cut point, which is a sharp turning point in the weirdness values, as the threshold. The words with a frequency lower than the threshold were considered as OCWs; hence they are removed from the list of person names. The formula for weirdness **[65]** is given as:

$$\boldsymbol{weirdness} = \frac{\boldsymbol{f_c} / \boldsymbol{N_c}}{\boldsymbol{f_{RC}} / \boldsymbol{N_{RC}}}$$

Equation 3-3 - Weirdness

where;

$\boldsymbol{f_c}$ = frequency of word in the Corpus

N_c = total number of words in the Corpus

f_{RC} = frequency of word in the Reference Corpus

N_{RC} = total number of words in Reference Corpus.

- *Find new patterns of Person Names from the concordance lines of reporting verbs that belong to RV_Form_2:* The concordance lines of RV_Form_2 can be seen in the **Appendix F**. These lines are grouped and given in Table 3-13 and Table 3-14.

Common pattern in Table 3-13 can be defined as:

[RV_Form_2] + [PN] + [,] + [W*]

Table 3-13 – Pattern 1 for RV_Form_2

Concordance line	**Reporting verb**	**Person name with title**	**Splitter character**	**The material succeeding person name**
1	söyleyen	Arsan	,	"2000 yılı toplam satış hedefimiz 2 milyonu aşkın hindi satmaktır. 2000 yılında piyasaların canlanmasına paralel bu satış hedefini de aşabiliriz" dedi.
2	ifade eden	Aydın	,	Bayındırlık Bakanlığı'nın depremle birlikte öne çıktığını vurguladı.
6	diyen	Chhibber	,	bu programın başarısı için özel sektör ve tüm toplumun katılımının gerekli olduğunu ifade etti.
7	kaydeden	Daloğlu	,	yarın işçilerle sözleşme imzalayacaklarını, 434 işçinin 25 Ocak Salı günü iş başı yapacağını söyledi.
9	ifade eden	Özyürek	,	"Böyle giderse hiçbir şekilde vergi toplayamayacağız. Çünkü vatandaş ödediği verginin yerine gidip gitmediğini görmek istiyor" dedi.

Table 3 – 13 (continued)

Concordance line	**Reporting verb**	**Person name with title**	**Splitter character**	**The material succeeding person name**
10	bildiren	Özyürek	,	"Enflasyonu aşağı çekebilmek vergi ile doğrudan bağlantılı. Oysa biz, bir milyar lira kazanandan da, bir milyon lira kazanandan da aynı oranda vergi alıyoruz" dedi.
11	kaydeden	Prof. Dr. Yalçın	,	"Makineleşmede traktör sayısı tek başına anlam taşımıyor. Çünkü, sadece çekici güç yaratıyor" diye konuştu.
12	belirten	Shigematsu	,	ilk çeyrekte alınacak verilerin, yeni ekonomik programın etkileri konusunda açık bir işaret vermeyeceğini kaydetti.

Common pattern in Table 3-14 can be defined as:

[RV_Form_2] + [Title] + [PN] + [, | (<E>ise)] + [W*]

Table 3-14 – Pattern 2 for RV_Form_2

Concordance line	**Reporting verb**	**Person name with title**	**Splitter character**	**The material succeeding person name**
3	açıklayan	Bakan Keçeciler	,	artık gümrük kapılarında 20 günde yapılabilen yüklemelerin, Avrupa Gümrük kapılarında olduğu gibi 3 saatte yapılabileceğini kaydetti.
4	açıklayan	Bakan Koray Aydın	,	"İkiztepe-Konak Doğanlar Otoyolu İzmir kent geçişi konusunda, topladığımız bilgiler ışığında Ankara'ya gidip değerlendirme yapacağım" dedi.
11	kaydeden	Prof. Dr. Yalçın	,	"Makineleşmede traktör sayısı tek başına anlam taşımıyor. Çünkü, sadece çekici güç yaratıyor" diye konuştu.
17	ifade eden	Mustafa Kumlu	ise	şöyle konuştu:

We combined all the patterns containing RV_Form_2 and created the final pattern called “RV_Form_2 final pattern” which is given as:

[RV_Form_2] + [Title]? + [PN] + [, | (<E>ise)] + [W*]

However, the list of concordance lines which is given in **Appendix F** does not contain concordance lines of all Turkish RVs rather it contains part of concordance lines of some Turkish RVs. After finding concordances lines of all Turkish RVs, we updated the “RV_Form_2 final pattern” for the sentence that contains RV_Form_2. It is called “RV_Form_2 updated final pattern” and is given as:

[RV_Form_2] + [Title]? + [PN] + [,|(<E>ise)|(<E>de)|(<E>da)] + [W*]

The graph of the “RV_Form_2 updated final pattern” can be seen in Figure 3-13.

It should be stated that there are common nouns which are used instead of person names such as these in concordance line **5, 8, 14, 15, and 16**. These words are OCWs given in Table 3-15 and will be ignored according to their weirdness values by using the same procedure applied to RV_Form_1.

Table 3-15 – Open class word pattern for RV_Form_2

Concordance line	**The material preceding person name**	**Person name with title**	**Splitter character**	**The material succeeding person name**
5	söyleyen	Başbakan **yardımcısı**	,	tahıl rekoltesinin önceki yıla göre 2 misli arttığını da belirtti.
8	belirten	Dünya Bankası **yetkilileri**	,	reform ile tarım sektöründekilere gerçek anlamda bir tarımsal desteğin sağlanacağını ve verimlilliğin de artacağını vurguluyorlar.

Table 3 -15 (continued)

Concordance line	The material preceding person name	Person name with title	Splitter character	The material succeeding person name
14	belirten	**üreticiler**	,	şunları söylediler: "Geçtiğimiz yıllarda teslim ettiğimiz ürüne karşılık bir miktar avans ödeniyordu. Bu yıl ise hiç para ödenmedi.İki aydan bu yana para almak için bekliyoruz”.
15	belirten	**yetkililer**	,	bu yıl ise 25 milyon metrekare fayans ve yer karosu ihraç etmeyi hedeflediklerini bildirdiler.
16	kaydeden	**yetkililer**	,	Türkiye'de ise yüksek maliyetler ve finansal sorunlar nedeniyle bu sektörlerin sorunlu bir dönem geçirdiğini dile getirdiler.

- *Find new patterns of Person Names from the concordance lines of RV_Form_3:* The concordance lines of RV_Form_3 can be seen in the **Appendix G**. These lines are grouped and given in Table 3-16, Table 3-17 and Table 3-18.

Common pattern in Table 3-16 can be defined as:

[PN] + [,] + [W*] + [RV_Form_3]

Table 3-16 – Pattern 1 for RV_Form_3

Concordance line	Person name	Splitter character	The material preceding reporting verb	Reporting verb
2	Avcı	,	en kötü ihtimalle bir kuşun yılda 30 yavrusunun olacağını	belirterek

Common pattern in Table 3-17 can be defin ed as:

[Title] + [PN] + [, | (<E>ise) | (<E>de)] + [W*] + [RV_Form_3]

Table 3-17 – Pattern 2 for RV_Form_3

Concordance line	**Person name with title**	**Splitter character**	**The material preceding reporting verb**	**Reporting verb**
1	Kadayıf üreticisi **Şevket Kılıç**	,	A.A muhabirine, kentte yazın 3-5, kış aylarında ise 200 işletmede üretim yapıldığını	açıklayarak
9	Oturum Başkanı Prof. Dr. **İlhan Özay**	ise	imtiyaz usulüne karşı olmadığını	belirterek
10	Avrupa Birliği (AB) Uzmanı **Jean François Drevet**	,	Türkiye'de bölgeler arasındaki gelişmişlik farkının çok yüksek olduğunu	belirterek
11	İstanbul Teknik Üniversitesi (İTÜ) Öğretim Üyesi Prof. Dr. **Kutsal Tülbentçi** de	,	İstanbul'da kaçak ve standart dışı yapılara yönelik çok sayıda imar affı çıktığını	belirterek
12	Prof. Dr. **Tülbentçi**	,	İstanbul'daki yapılarda beton kalitesinin düşük olduğunu	belirterek
13	İTÜ Öğretim Üyesi **Tevfik Sena Arda**	ise	çelik yapıların betonarme yapılara göre hafif olduğunu	belirterek
14	Vali **Türk**	,	Iğdır-Tuzluca güzergahında bulunan 20 bin dönümlük arazinin organize sanayi bölgesinin kurulması için tahsis edildiğini	diyerek
15	Gebze Yüksek Teknoloji Enstitüsü (GYTE) Çevre Mühendisliği Bölüm Başkanı Prof. Dr. **Mehmet Karpuzcu**	,	A.A muhabirine yaptığı açıklamada, dünyadaki su, petrol ve doğalgaz kaynaklarının sınırlı olduğunu	diyerek
18	Enerji ve Tabii Kaynaklar Bakanı ve Başbakan Yardımcısı **Cumhur Ersümer**	,	enerji meselesinin uzun yıllara sarih edecek bir mesele olduğunu	ifade ederek
19	TZOB Başkanı **Faruk Yücel**	de	TBMM'nin bu dönemde iyi çalıştığını	ifade ederek

Table 3 - 17 (continued)

Concordance line	Person name with title	Splitter character	The material preceding reporting verb	Reporting verb
21	Holding Yönetim Kurulu Başkanı **Üzeyir Garih**	ise	iş yaşamına 49 yıl önce Carrier'de başladığını	ifade ederek
22	Bakan **Aydın**	,	hükümetin attığı olumlu adımlarla Türkiye'nin yıllardır çektiği siyasi istikrarsızlığın atlatıldığını	ifade ederek
24	Doç. Dr. **İlyas Yılmazer**	ise	ovaların, depremler sonucu oluşan ulusal servet ler olduğunu	ifade ederek
25	Bakan **Önal**	,	mazot ve sınır ticareti konusundaki bir soru üzerine, İran'ın, "mazot taşıyan kamyonların mazotunu başka ülkelerden aldığını ve İran'dan transit olarak geçtiğini" iddia ettiğini	kaydederek
26	Türk-Güney Kore İş Konseyi Başkanı **Ali Kibar**	,	Türkiye ile Güney Kore'nin sanayileşme süreçleri arasında pek çok benzerlikler bulunduğunu	kaydederek
27	Renault Yönetim Kurulu Başkanı **Luis Schweitzer**	ise	ekonomide iş çevrelerinin rekabetçi bir ortamda, azami kar elde etmeye çalıştığını	kaydederek
29	KKTC Başbakanı **Derviş Eroğlu**	,	bankalar konusunda Türkiye ile temasların sürdüğünü	söyleyerek

Common pattern in Table 3-18 can be defined as:

[W*] + [PN] + [(<E>da)] + [W*] + [RV_Form_3]

Table 3-18 – Pattern 3 for RV_Form_3

Concordance line	The material containing person name	Splitter character	The material preceding reporting verb	Reporting verb
30	Mudiler adına konuşan **Namık Ramadan**	da	Yurtbankzedeler'in çok zor durumda olduğunu	söyleyerek

We combined all the patterns containing RV_Form_3 and created the final pattern called "RV_Form_3 final pattern" which is given as:

[Title] + [PN] + [, | (<E>ise) | (<E>de) | (<E>da)] + [W*] + [RV_Form_3]

However, the list of concordance lines which is given in **Appendix G** does not contain concordance lines of all Turkish RVs rather it contains part of concordance lines of some Turkish RVs. After finding concordances lines of all Turkish RVs, we updated the "RV_Form_3 final pattern" for the sentence that contains RV_Form_3. It is called "RV_Form_3 updated final pattern" and is given below:

[Title] + [PN] + [, | (<E>ise) | (<E>de) | (<E>da)] + [W*] + [RV_Form_3]

The graph of the "RV_Form_3 updated final pattern" can be seen in Figure 3-14.

It should be stated that there are common nouns which are used instead of person names such as these in concordance line **3, 4, 5, 6, 7, 8, 16, 17, 20, 23, and 28**. These words are OCWs that can be seen in Table 3-19 and will be ignored according to their weirdness values by using the same procedure.

Table 3-19 – Open class word pattern for RV_Form_3

Concordance line	The material containing person name	Splitter character	The material preceding reporting verb	Reporting verb
3	İl Tarım Müdürlüğü **yetkilileri**	,	yerli hayvan ırklarının ıslah edilmesi için her türlü olanağın en iyi şekilde değerlendirilmesine çalışıldığını	belirterek

Table 3 - 19 (continued)

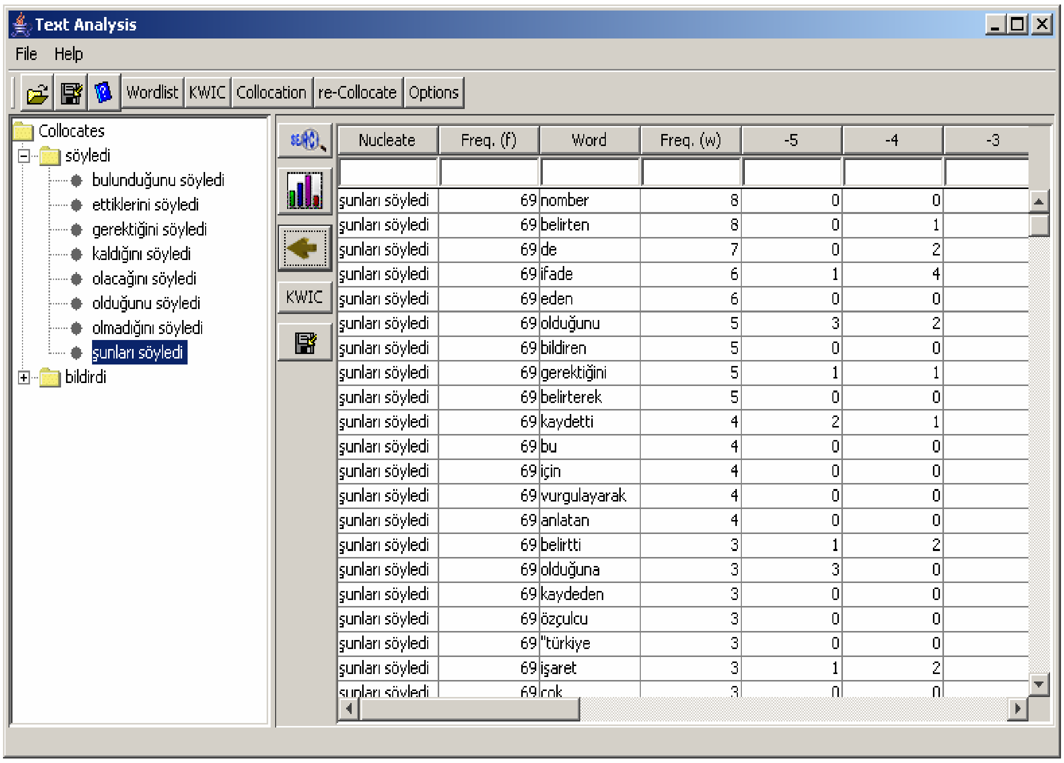
Concordance line	**The material containing person name**	**Splitter character**	**The material preceding reporting verb**	**Reporting verb**
4	A.A muhabirine bilgi veren ÇATES **yetkilileri**	,	enerji üretiminde 1998 yılına göre 1999'da 7 milyon 127 bin kilowatsaat düşüş olduğunu	belirterek
5	Borsa **uzmanları**	,	hızlı çıkışta yeni yılla birlikte faizlerdeki ani düşüşlerin önemli ölçüde etkili olduğunu	belirterek
6	**Vatandaşlar**	,	balık fiyatlarının 500 bin lira ile 2.5 milyon lira arasında olduğunu	belirterek
7	**Üreticiler**	,	Mürefte beldesindeki Zeytin Tarım Satış Kooperatifleri'ne ürünlerini teslim ettiklerini	belirterek
8	Borsa **uzmanları**	,	satış baskısına rağmen piyasanın dengeli bir seyir izlediğini	belirterek
16	DSİ **yetkilileri**	<E>	barajlarda depolanan su miktarının mevsim itibariyle uygun düzeyde bulunduğunu	diyerek
17	**Uzmanlar**	,	alternatif piyasaların faiz cephesinde dengelerin oturmasıyla birlik te Borsa'da düşüş yönünde yaratılan beklentilerin zayıflamaya başladığını	ifade ederek
20	Tansaş Gıda Şirketi **yetkilileri**	,	restoranın klasik havasının korunmasına özen gösterdiklerini	ifade ederek
23	Borsa **uzmanları**	,	işlem hacminin 18,500 direncini kırmakta yeterli olmadığını	ifade ederek
28	**Yetkililer**	,	büyükbaş hayvanların canlı olarak kilosu 1 milyon 150 bin liradan satılacağını	kaydederek

3.5.2 COLLOCATION ANALYSIS

- *Collocation Analysis:* After finding patterns from concordance lines manually, we carried out collocation analysis of RVs on the FEC2000 and obtained collocations of these RVs by using the text analysis tool.

Consequently, we acquired another form of RVs from collocation analysis, which is described below;

- **Reporting verbs with an pronominal complement**: Examples of these types of RVs are "şunları söyledi", "şunları kaydetti", "şöyle dedi", "şöyle konuştu", "sözlerini şöyle sürdürdü", and "sözlerini şöyle tamamladı". Hereafter, we refer the reporting verbs with a pronominal complement as RV_Form_4. As an example, the collocations of the reporting verb "söyledi" can be seen in Figure 3-11.



Nucleate	Freq. (f)	Word	Freq. (w)	-5	-4	-3
şunları söyledi	69	nomber	8	0	0	
şunları söyledi	69	belirten	8	0	1	
şunları söyledi	69	de	7	0	2	
şunları söyledi	69	ifade	6	1	4	
şunları söyledi	69	eden	6	0	0	
şunları söyledi	69	olduğunu	5	3	2	
şunları söyledi	69	bildiren	5	0	0	
şunları söyledi	69	gerektiğini	5	1	1	
şunları söyledi	69	belirterek	5	0	0	
şunları söyledi	69	kaydetti	4	2	1	
şunları söyledi	69	bu	4	0	0	
şunları söyledi	69	için	4	0	0	
şunları söyledi	69	vurgulayarak	4	0	0	
şunları söyledi	69	anlatan	4	0	0	
şunları söyledi	69	belirtti	3	1	2	
şunları söyledi	69	olduğuna	3	3	0	
şunları söyledi	69	kaydeden	3	0	0	
şunları söyledi	69	özçulcu	3	0	0	
şunları söyledi	69	"türkiye	3	0	0	
şunları söyledi	69	işaret	3	1	2	
şunları söyledi	69	çok	3	0	0	

Figure 3-11 – Collocation analysis of the reporting verb "söyledi"

In the Figure 3-11, we found eight collocates of the reporting verb "söyledi". These are "bulunduğunu", "ettiklerini", "gerektiğini", "kaldığını", "olacağını", "olduğunu", "olmadığını", and "şunları". We performed collocation analysis for all RVs and found their collocations.

The results of the collocation analysis revealed some interesting statistics. For example, the top 10 collocates of reporting verb “söyledi” along with their frequency distribution, spread (U-score), and strength (K-score) statistics can be seen in Table 3-20. All collocates of “söyledi” with a K-score higher than the unity are given in **Appendix I**.

Table 3-20 – Top ten collocations of reporting verb “söyledi”

Reporting Verb	Collocate	K-score	U-score	Frequency
söyledi	şunları	118	177.862,7	2.765
söyledi	olduğunu	108	108.434,5	2.499
söyledi	nomber (any numeric number)	103	11.799,5	2.239
söyledi	ve	85	9.822,5	1.831
söyledi	bir	79	12.676,9	1.787
söyledi	gerektiğini	58	36.056,5	1.349
söyledi	bu	47	2.403,5	1.034
söyledi	da	42	1.859,7	952
söyledi	de	35	1.270,3	791
söyledi	için	32	1.532,9	713

As seen in Table 3-20, there are some collocations that are used with reporting verb “söyledi” such as “şunları + söyledi”, “olduğunu + söyledi”, “gerektiğini + söyledi”. We considered these collocations and performed concordance analysis by using these new RVs.

3.5.3 CONCORDANCE ANALYSIS FOR FOURTH FORM OF REPORTING VERBS

- *Concordance analysis on RV_Form_4:* we carried out additional concordance analysis based on RV_Form_4 by using the Nooj software. Concordance lines of RV_Form_4 can be seen in the **Appendix H**.

  - *Find additional patterns of Person Names from the concordance lines of RV_Form_4:* The concordance lines of RV_Form_4 can be seen in the **Appendix H**. These lines are grouped and given in

Table 3-21, Table 3-22, Table 3-23, Table 3-24, Table 3-25, Table 3-26, and Table 3-27.

Common pattern in Table 3-21 can be defined as:

[PN] + [,] + [RV_Form_4] + [:]

Table 3-21 – Pattern 1 for RV_Form_4

Concordance line	Person name	Splitter character	The material preceding reporting verb	Reporting verb
1	Özince	,	<E>	sözlerini şöyle sürdürdü:
3	Voegele	,	<E>	sözlerini şöyle tamamladı:
8	Özince	,	<E>	şöyle dedi:

Common pattern in Table 3-22 can be defined as:

[PN] + [,] + [W*] + [,] + [RV_Form_4] + [:]

Table 3-22 – Pattern 2 for RV_Form_4

Concordance line	Person name	Splitter character	The material preceding reporting verb	Reporting verb
7	Uzunhekim	,	Dutilhlere ait denizcilik işletmesinin yanı sıra seyahat acentasının da kendilerine devredileceğini belirterek,	şöyle dedi:
13	Şadan Eren	,	yaptığı yazılı açıklamada, alınan kararları dikkatle izlediklerini ve başarılı olması için destek verdiklerini belirterek,	şunları kaydetti:

Common pattern in Table 3-23 can be defined as:

[W*] + [PN] + [,] + [RV_Form_4] + [:]

Table 3-23 – Pattern 3 for RV_Form_4

Concordance line	**The material containing person name**	**Splitter character**	**The material preceding reporting verb**	**Reporting verb**
2	Kutan'ın bu sözleri üzerine ATO Başkanı Aygün'ün "yani hükümete destek oluyorsunuz" şeklindeki sorusunu da yanıtlayan **Kutan**	,	<E>	sözlerini şöyle sürdürdü:
4	Kış aylarının tüm dünyada sert geçmesi nedeniyle ham petrol fiyatlarının yükselme eğiliminde olduğuna dikkati çeken **Turgut Bozkurt**	,	<E>	sözlerini şöyle tamamladı:
6	Ticaret merkezinin her türlü altyapısını tamamlamış olmasına karşın, tam kapasiteyle çalışamadığına işaret eden **Dalan**	,	<E>	şöyle dedi:
9	Yat Limanı inşaatı için ortaya çıkan taş ocağı sorununun da giderilmesi için çalışıldığını belirten **Ergül**	,	<E>	şöyle konuştu:
11	"Ülkemizin, geleceğin dünyasında onurlu yerini almasının olmazsa olmaz koşulu budur" diyen **Teberik**	,	<E>	şunları kaydetti:
14	Ege Serbest Bölgesi'nin yüksek teknoloji ve temiz çevre anlayışıyla üretim yapan modern bir endüstri merkezi olmasını amaçladıklarını ifade eden **Tuncer**	,	<E>	şunları söyledi:

Common pattern in Table 3-24 can be defined as:

[W*] + [Title] + [PN] + [, | (<E>ise)] + [RV_Form_4] + [:]

Table 3-24 – Pattern 4 for RV_Form_4

Concordance line	**The material containing person name with title**	**Splitter character**	**The material preceding reporting verb**	**Reporting verb**
5	Tarımın ulusal ekonominin temelini oluşturduğunu vurgulayan Prof. Dr. **Tekinel**	,	<E>	sözlerini şöyle tamamladı:
12	"Sadece üretmenin sorunu çözmediğini" ifade eden Prof. **Çakır**	,	<E>	şunları kaydetti:
17	Antepfıstığında en büyük sıkıntının dış pazarlarda rekabet edememek olduğuna dikkati çeken Güneydoğubirlik Ticaret Müdürü **Mustafa Balaban**	ise	<E>	şunları kaydetti:

Common pattern in Table 3-25 can be defined as:

[Title] + [PN] + [,] + [W*] + [,] [RV_Form_4] + [:]

Table 3-25 – Pattern 5 for RV_Form_4

Concordance line	**Person name with title**	**Splitter character**	**The material preceding reporting verb**	**Reporting verb**
16	Prof. Dr. **Karamış**	,	Taksan için Mart ayında yeniden ihale açılacağı yolunda duyumlar aldıklarını bildirerek,	şunları söyledi:

Common pattern in Table 3-26 can be defined as:

[Title] + [PN] + [(<E>da)] + [W*] + [,] [RV_Form_4] + [:]

Table 3-26 – Pattern 6 for RV_Form_4

Concordance line	Person name with title	Splitter character	The material preceding reporting verb	Reporting verb
10	Ericsson Türkiye Başkan Yardımcısı **Stefan Ofverholm**	da	Türkiye'nin Ericsson'un dünyadaki en önemli pazarlarından biri olduğunu vurgulayarak,	şöyle konuştu:

Common pattern in Table 3-27 can be defined as:

[Title] + [PN] + [(<E>de)] + [W*] + [,] [RV_Form_4] + [:]

Table 3-27 – Pattern 7 for RV_Form_4

Concordance line	Person name with title	Splitter character	The material preceding reporting verb	Reporting verb
15	Rize Çay Üreticileri Birliği Başkanı **Nurettin Kepenek**	de	üreticilerin alın terinin karşılığının yok edildiğini iddia ederek,	şunları söyledi:

We combined all patterns containing RV_Form_4 and created the final pattern called “RV_Form_4 final pattern” which is given as:

[W*]? + [Title]? + [PN] + [, | (<E>de) | (<E>da)] + [W*]? + [,] + [RV_Form_4] + [:]

However, the list of concordance lines which is given in **Appendix H** does not contain concordance lines of all Turkish RVs rather it contains part of concordance lines of some Turkish RVs. After finding concordances lines of all Turkish RVs, we updated the “RV_Form_4 final pattern” for the sentence that contains RV_Form_4. It is called “RV_Form_4 updated final pattern” and is given:

[W*]? + [Title]? + [PN] + [, | (<E>ise) | (<E>de) | (<E>da)] + [W*]? + [,] + [RV_Form_4] + [:]

The graph of the "RV_Form_4 updated final pattern" can be seen in Figure 3-15.

3.5.4 CREATING THE LIST OF ALL PATTERNS FOUND

- *Creating the list of all patterns found:* we create a list of all updated final patterns found. As a result, we obtained four updated final patterns. Regular expressions for each pattern can be seen as:

  - **PATTERN 1;**

    [Title]? + [W*]? + [PN] + [W*]? + [, | (<E>ise) | (<E>de) | (<E>da)] + [W*] + [RV_Form_1] + [.]

  - **PATTERN 2;**

    [RV_Form_2] + [Title]? + [PN] + [, | (<E>ise) | (<E>de) | (<E>da)] + [W*]

  - **PATTERN 3**;

    [Title] + [PN] + [, | (<E>ise) | (<E>de) | (<E>da)] + [W*] + [RV_Form_3]

  - **PATTERN 4;**

    [W*]? + [Title]? + [PN] + [, | (<E>ise) | (<E>de) | (<E>da)] + [W*]? + [,] + [RV_Form_4] + [:]

Consequently, we run these regular expressions in the Nooj software for the FEC2000. The graphs of four updated final patterns are given in Figure 3-12, Figure 3-13, Figure 3-14, and Figure 3-15 respectively.

- **Pattern 1**

The graphical representation of **pattern 1** can be seen Figure 3-12 below.

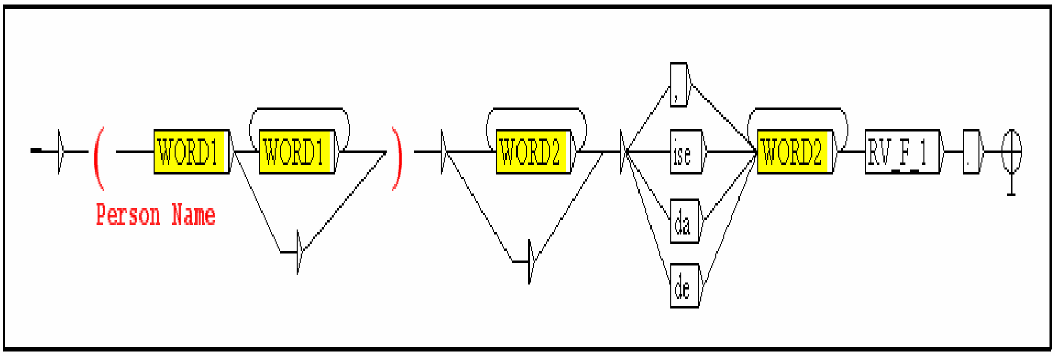


Figure 3-12 – The graphical representation of pattern 1

- **Ex:** **Şimşek** yaptığı açıklamada**,** "1998 yılında yaşanan sel felaketleri nedeniyle pamuk üreticisi zaten verimli ürün alamadı. Beklentimiz olan 20 centlik primin yanına dahi yaklaşılmadı" **dedi**.

- **Ex:** Çin Devlet Kalkınma Planlaması Komitesi Başkanı **Zeng Peiyan,** Çin'in yakın bir gelecekte, para birimi yuanda devalüasyon uygulamasına gerek olmadığını **belirtti**.

- **Pattern 2**

The graphical representation of **pattern 2** can be seen Figure 3-13 below.

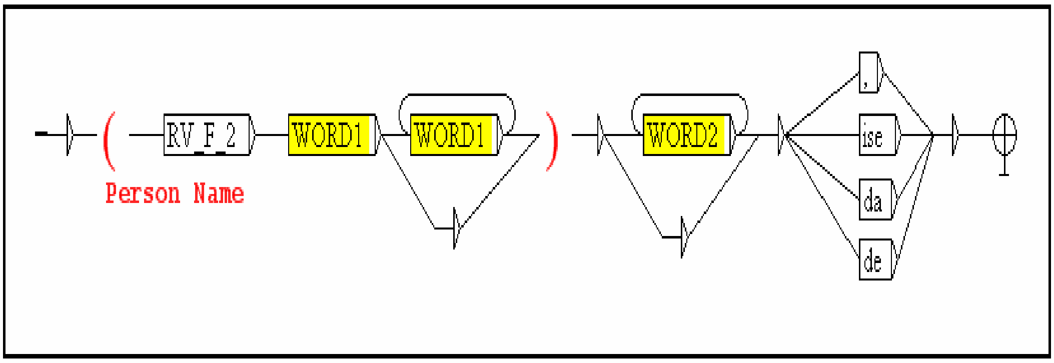


Figure 3-13 – The graphical representation of pattern 2

- **Ex: açıklayan** Bakan **Keçeciler,** artık gümrük kapılarında 20 günde yapılabilen yüklemelerin, Avrupa Gümrük kapılarında olduğu gibi 3 saatte yapılabileceğini kaydetti.

- **Ex: bildiren Özyürek,** "Enflasyonu aşağı çekebilmek vergi ile doğrudan bağlantılı. Oysa biz, bir milyar lira kazanandan da, bir milyon lira kazanandan da aynı oranda vergi alıyoruz" dedi.

➢ **Pattern 3**

The graphical representation of **pattern 3** can be seen Figure 3-14 below.

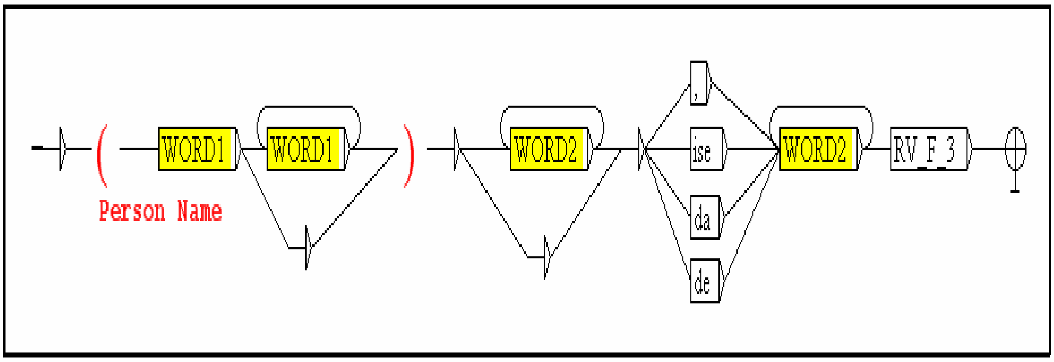


Figure 3-14 – The graphical representation pattern 3

- **Ex:** Oturum Başkanı Prof. Dr. **İlhan Özay ise** imtiyaz usulüne karşı olmadığın **belirterek**

- **Ex:** TZOB Başkanı **Faruk Yücel de** TBMM'nin bu dönemde iyi çalıştığını **ifade ederek**

➢ **Pattern 4**

The graphical representation of **pattern 4** can be seen Figure 3-15 below.

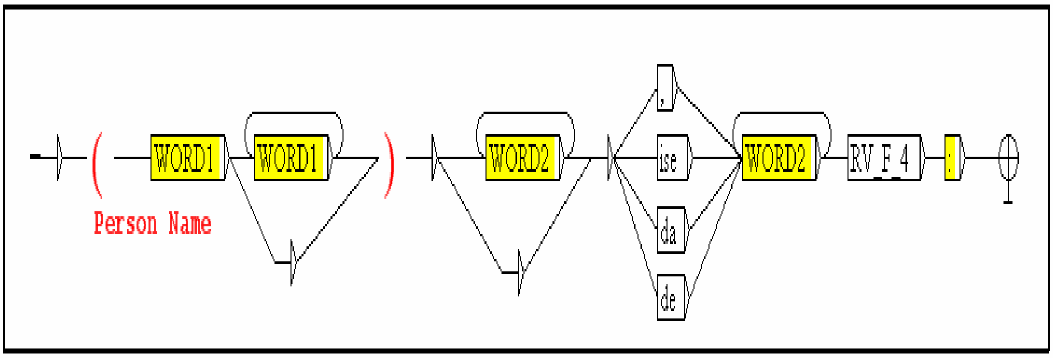


Figure 3-15 – The graphical representation of pattern 4

- Kutan'ın bu sözleri üzerine ATO Başkanı Aygün'ün, "yani hükümete destek oluyorsunuz" şeklindeki sorusunu da yanıtlayan **Kutan, sözlerini şöyle sürdürdü:**

- **Özince, sözlerini şöyle sürdürdü:**

- **Voegele, sözlerini şöyle tamamladı:**

where;

WORD1 in the Figure 3-12, Figure 3-13, Figure 3-14, and Figure 3-15 stands for any word that starts with an upper letter or a digit and continues with an upper letter, a lower letter, a digit, or a special character. The graphical representation of WORD1 is given in Figure 3-16.

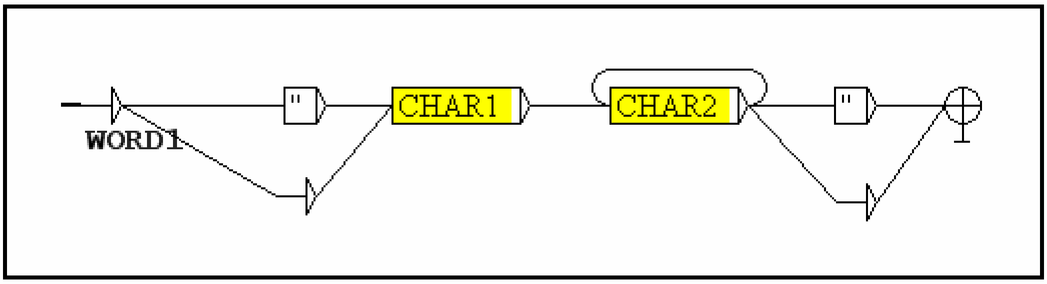


Figure 3-16 – The graphical representation of WORD1

WORD2 in the Figure 3-12, Figure 3-13, Figure 3-14, and Figure 3-15 stands for any word that starts with an upper letter, a lower letter, a digit, and a special character and continues with an upper letter, a lower letter, a digit, or a special character. The graphical representation of WORD2 is given in Figure 3-17.

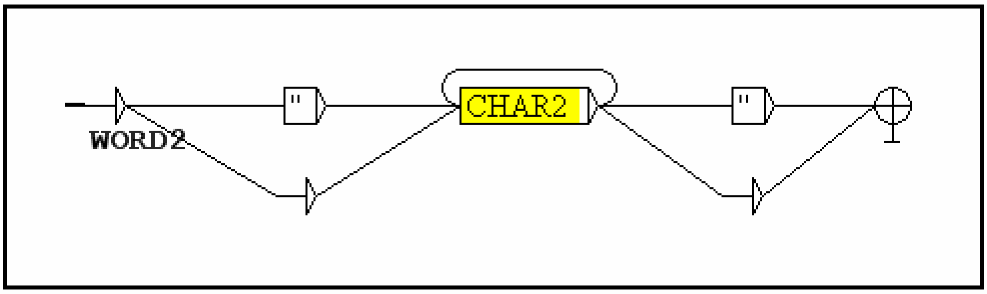


Figure 3-17 – The graphical representation of WORD2

where;

CHAR1 in Figure 3-16 stands for any upper letter or digit. The graph of CHAR1 is given in Figure 3-18.

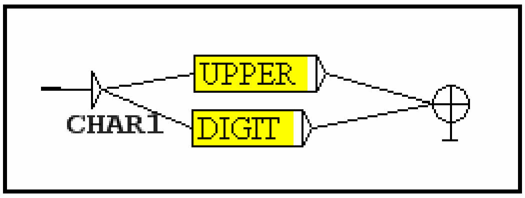


Figure 3-18 – The graphical representation of CHAR1

CHAR2 in Figure 3-16 and Figure 3-17 stands for any upper letter, lower letter, special character or digit. The graphical representation of CHAR2 is given in Figure 3-19.

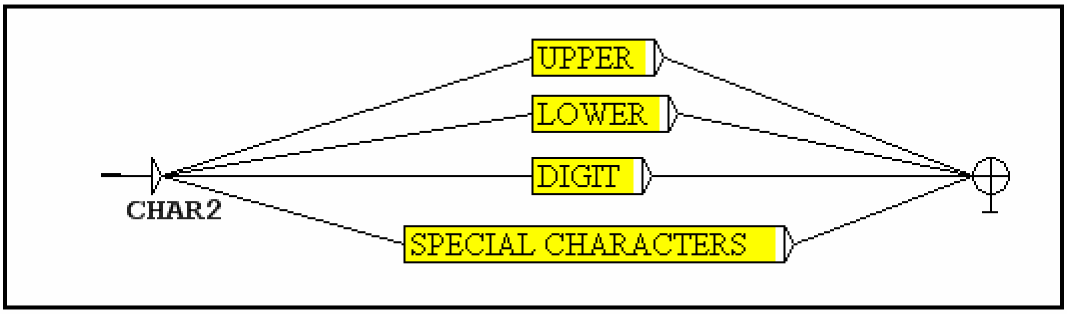


Figure 3-19 – The graphical representation of CHAR2

DIGIT in Figure 3-18 and Figure 3-19 stands for any digit. The graphical representation of DIGIT is given in Figure 3-20.

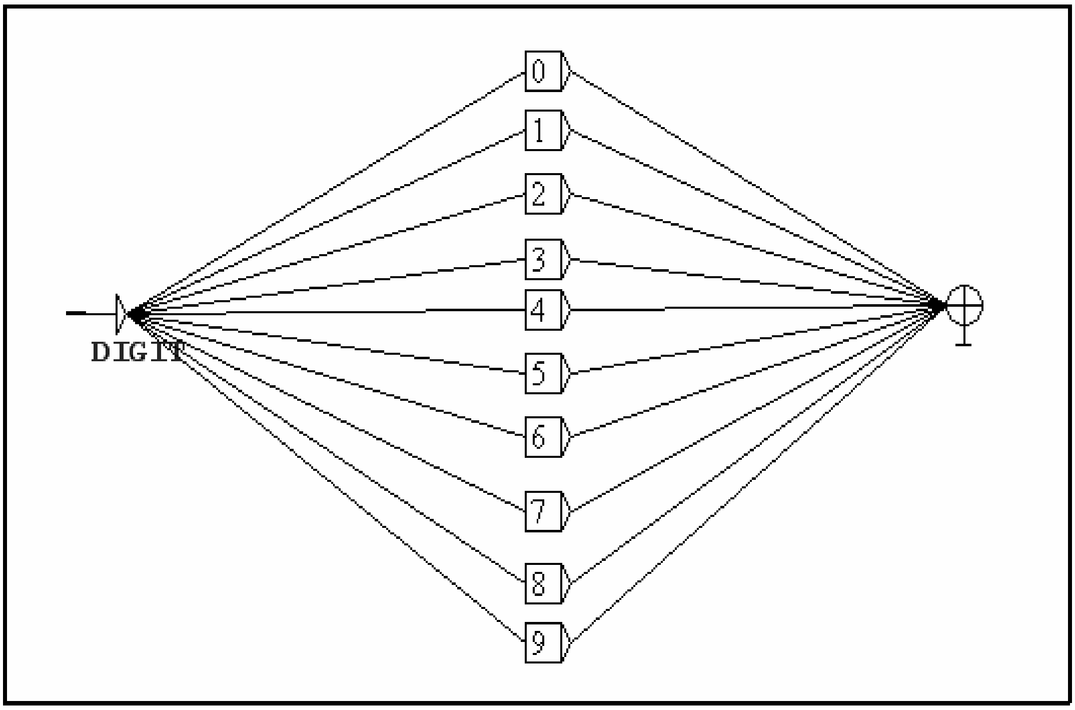


Figure 3-20 – The graphical representation of DIGIT

UPPER in Figure 3-18 and Figure 3-19 stands for any upper case letter from A to Z and Turkish upper case letters which are Ç, Ğ, İ, Ö, Ü, and Ş. The graphical representation of UPPER is given in Figure 3-21.

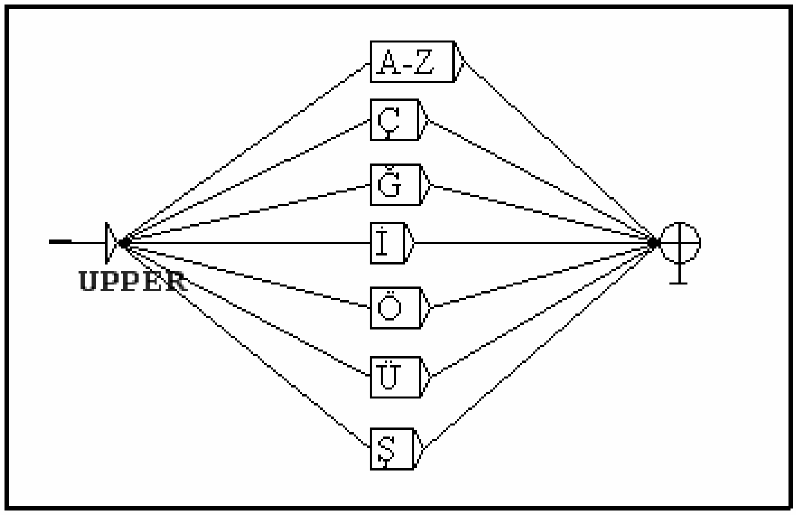


Figure 3-21 – The graphical representation of UPPER

LOWER in Figure 3-19 stands for any lower case letter from a to z and Turkish lower case letters which are ç, ğ, ı, ö, ü, and ş. The graphical representation of LOWER is given in Figure 3-22.

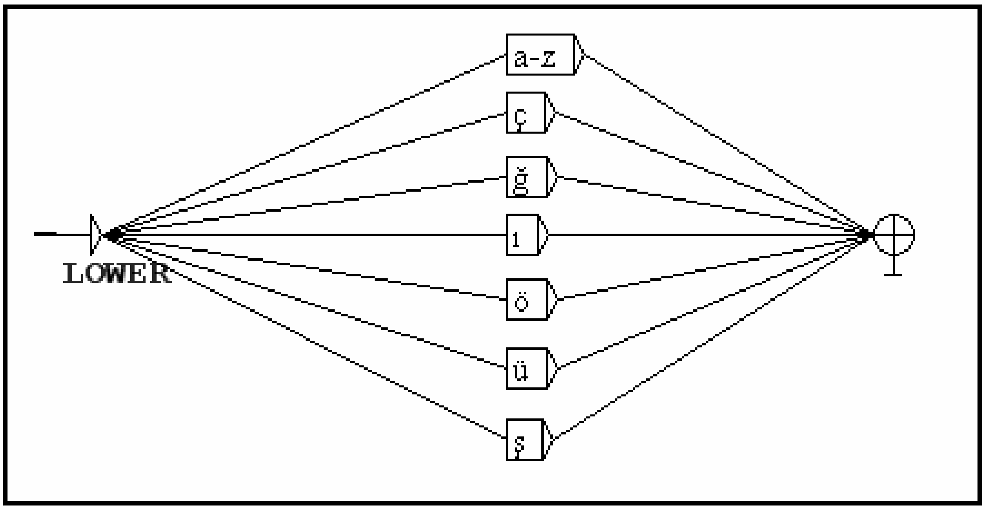


Figure 3-22 – The graphical representation of LOWER

SPECIAL CHARACTERS in Figure 3-19 stands for some special characters that can be used in sentences. The graphical representation of SPECIAL CHARACTERS is given in Figure 3-23.

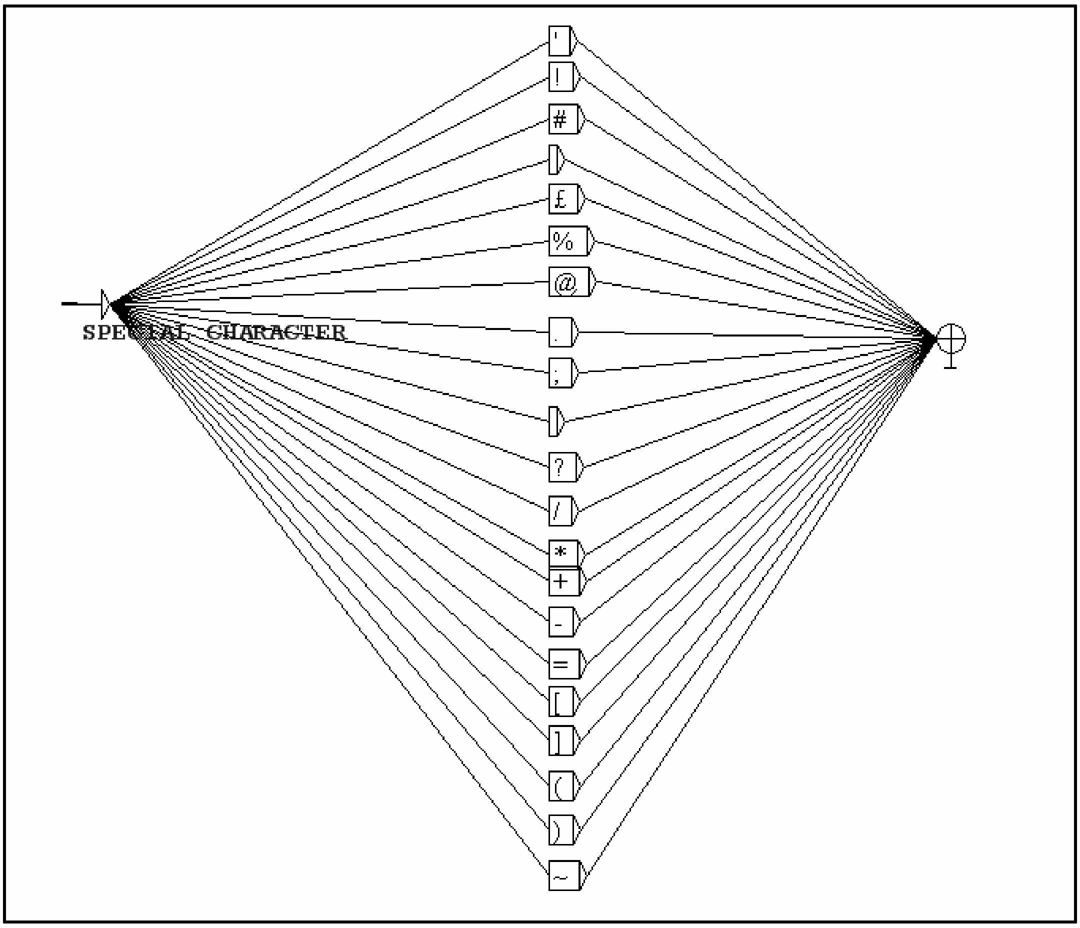


Figure 3-23 – The graphical representation of SPECIAL CHARACTERS

3.5.5 EXTRACTING PERSON NAMES

The person name extraction has five steps which are described as;

- *Extracting Person Names from the FEC2000:* We extracted person names from the FEC2000 by using the regular expressions with a simple Java tool that we developed.

- *Extracting Person Names from the list of all capitalized words and updating the list of Person Names:* Firstly, we found all the list of capitalized words from the FEC2000 by using the regular expressions with simple Java tool that we developed. Next, we compared the capitalized words extracted from a text with person names extracted from the same text by holding their position information. And then, we tried to find the new uses of person names and added them to the list of person names.

- *Remove the one-token words from updated list of Person Names according to the weirdness value:* Firstly, We found the weirdness values of all one-token person names and the one-token person names with lower weirdness value than the threshold are removed from the list of person names.

- *Remove the words that can be categorized as continent name, country name, region name, city name, county name, and international currency from updated list of Person Names:* We removed the words that are not person names but have a similar structure (in terms of being capitalized and having same order of letters) and can be categorized as continent name, country name in the world, region name in Turkey, city name in Turkey, county name in Turkey, and international currency from the list of person names by using the list of these categories. Finally, we create final list of person names from the FEC2000.

- *Constructing Final List of Person Names: A*fter removing the one token person names which have lower weirdness values and the words that can be categorized as as continent name, country name in the world, region name in Turkey, city name in Turkey, county name in Turkey, and international currency, we constructed the final list of person names.

3.6 SUMMARY

In this section, we discuss the LG approach that we used in order to obtain the NEs, specifically person names, in Turkish financial news. We describe all the steps of the LG approach in detail.

We firstly cleaned the news' contents from tags, split them into sentences by using splitter, and built a new corpus (called FEC2000) from the resulting texts. Secondly, we acquired the word frequency list of the FEC2000**.** Thirdly, by using statistical tests, we found the most significant Turkish RVs in both corpora. In this third step, we translated English RVs of Quirk et. al **[2]** and Hayssam **[1]** into Turkish by using Red House dictionary **[66]**. We identified other potential RVs by using the word frequency lists of the FEC2000 and the METU TC. We combined all RVs and created the final list of Turkish RVs. We tested the significance of these Turkish RVs, and created a list of significant Turkish RVs. Fourthly, we extracted NEs in the FEC2000 by using the LG approach. In this fourth step, we carried out concordance analysis and found concordance lines that contain these RVs. We identified three patterns in these concordance lines. After identifying these three patterns from concordance lines, we realized that there are RVs that contain other RVs in the concordance lines and we performed a collocation analysis based on these RVs containing other RVs. We acquired new Turkish RVs from collocations analysis by using the RVs containing other RVs and carried out concordance analysis by using these newly found Turkish RVs. We found one new pattern from these new concordance lines. Consequently, we combined all four patterns that are found from concordance lines and created the list of four patterns. And then, we executed regular expressions of these patterns and found materials that contain person names. After finding materials containing person names, we created a list of these materials. Then, we extracted the NEs in the FEC2000 by using these patterns. In addition, we found all capitalized words in the FEC2000 by using capitalization rules in order to find other uses of person names for NEs found before and updated the person name list. We also checked the weirdness values of person names found before and removed person names that have significantly lower values than threshold. Moreover, we removed the words that are not person names but have similar structure (in terms of being capitalized and having same order of letters) and can be categorized as continent name, country name in the world, region name, city name, and county name in Turkey, and international currency from the list of person names by using the list of these categories. Finally, we created the final list of person names.

To sum up, we believe that the LG approach is a good and an effective approach in order to acquire NEs, specifically person names, from the Turkish financial news. We will give information about our evaluation strategy and experimental results in the Evaluation chapter. Discussion and comparison will also be carried out in this chapter.

CHAPTER 4

EVALUATION

In this chapter, we will evaluate the LG approach to identify person names in the financial news published in 2000 by Anadolu Agency. We will compare our results with naive PN extraction method, which uses only capitalization rules.

4.1 THE EVALUATION STRATEGY AND THE EXPERIMENTAL RESULTS

In this chapter, we discuss the evaluation strategy that we follow and give the experimental results that we obtained in this study.

4.1.1 EVALUATION STRATEGY

In **[1]**, it was shown that person names frequently occur in the clauses that include RVs. Financial news is expected to include several RVs. This is because the statements made by company directors or presidents often appear in such texts and announcements are made about companies and stock markets. In the method chapter, we have already shown that the frequencies of RVs in the financial corpus (FEC2000) are significantly different compared to that of our Turkish reference corpus (METU TC). As a consequence, LGs which are constructed by using such verbs may help the extraction of person names.

Our testing data set comprises 200 news having 41.322 tokens, which were published on January 2001. An example paragraph which was taken from the tagged test corpus is given below:

<**ENAMEX** TYPE="PERSON">Ölçal</**ENAMEX**>, <**ENAMEX** TYPE="ORGANIZATION"> A.A</**ENAMEX**> muhabirine yaptığı açıklamada, <**ENAMEX** TYPE="ORGANIZATION"> GTO</**ENAMEX**> adına <**ENAMEX** TYPE="ORGANIZATION">Gaziantep Üniversitesi İktisadi ve İdari Bilimler Fakültesi</**ENAMEX**> Dekanı Prof. Dr. <**ENAMEX** TYPE= "PERSON">İsmail Hakkı Özsabuncuoğlu</**ENAMEX**>'nun anket yöntemiyle yaptığı "Süpermarketler" konulu çalışmanın, tüketici bilincinin hala bölgenin en gelişmiş kenti olan Gaziantep ve Türkiye'nin geldiği noktaya uygun gelişme göstermediğini ortaya koyduğunu söyledi.

The news was tagged by twenty research assistants in the Informatics Institute of Middle East Technical University by following MUC standards. MUC tags for PNs can be seen below;

- <**ENAMEX** TYPE=”PERSON”>Person Name</**ENAMEX**>
- <**ENAMEX** TYPE=”ORGANIZATION”>Organization Name</**ENAMEX**>
- <**ENAMEX** TYPE=”LOCATION”>Location Name</**ENAMEX**>

PERSON stands for named person or family, ORGANIZATION stands for named corporate, governmental, or other organizational entity and LOCATION stands for name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.) **[68]**.

A java tool was developed to extract person names from the tagged test corpus and the list of person names was constructed. And this list of person names was compared with the list of person names that were extracted from untagged test corpus by using the LG approach. It should be stated that the position of person names in the texts is very important.

Person names extracted from a text in the tagged test corpus should be compared with person names extracted from the corresponding text in the untagged test corpus. Considering this fact, we manually checked whether the positions of person names extracted from a text in the tagged corpus and positions of person names extracted from a corresponding text in the untagged corpus are the same or not.

MUCs primary measures - Precision, Recall, and F-measure - are used in the evaluation phase in this thesis. Precision is the measure of how much the response results are actually in the test set. Recall is the measure of how much of the test set are covered by the response results. These measures can be formulated as below;

$P = Correct / Found$ Equation 4-1 - Precision

$R = Correct / Exist$ Equation 4-2 - Recall

where;

P = Precision
R = Recall
Correct = Number of correctly found elements in response results i.e. number of elements actually in the test set
Found = Number of all elements in response results
Exist = Number of all elements in the test set

In **[1]**, it was stated that there are cases in which responses from a system that is adjusted for high recall can differ from a system that is adjusted for high precision. Therefore, they decided that it was better to use a combination of recall and precision in order to cover these cases such as F-measure. Consequently, they used van Rijsbergen's F-measure. The formula of van Rijsbergen's F-measure is given in Equation 4-3;

$$F = \frac{(\beta^2 + 1) * P * R}{(\beta^2 * P) + R}$$ Equation 4-3 - F-measure

where;

β is the relative importance given to the recall over precision.

This formula is derived from the formula in **[69]**, which is given below;

$$F_\alpha = \frac{1}{(\alpha/P) + (1-\alpha)/R}$$ where; $$\alpha = 1/(\beta^2 + 1)$$

If recall and precision are equally important, then β = 1. For the recall which is as twice important as precision, β = 2 is used. For the recall which is as half important as precision, β = 0.5 is used **[1][69]**.

4.1.2 EXPERIMENTAL RESULTS

Recall, Precision, and F-measure results that we achieved for the LG approach can be seen in the Table 4-1 below.

Table 4-1 – F-measure, recall, and precision results for the LG approach with C.R. (Capitalization Rule)

	Recall	**Precision**	**F-measure P&R (β = 1)**	**F-measure P&2R (β = 2)**	**F-measure 2P&R (β = 0.5)**
Local Grammar with C.R.	86.21	78.13	81.97	84.46	79.62

Although the Precision value (**78.13**) is not very high, Recall value (**86.21**) is encouragingly high. F-measure values for all β values (β = 1, β = 2, and β = 0.5) with the LG approach are 81.97, 84.46, and 79.62. They are also not very high because of the Precision value which is not very high.

We compared the LG approach with naïve PN extraction method, which takes into account only Capitalization Rules which tries to find all the capitalized words in the corpus.

Recall, Precision, and F-measure results for the LG approach and for Capitalization Rules can be seen in the Table 4-2.

Table 4-2 – F-measure, recall, and precision results for the LG approach with C.R. (Capitalization Rule) and only the Capitalization Rule

	Recall	**Precision**	**F-measure P&R (β = 1)**	**F-measure P&2R (β = 2)**	**F-measure 2P&R (β = 0.5)**
Local Grammar with C.R.	86,21	78,13	81,97	84,46	79,62
Capitalization Rule Only	99,37	6,66	12,48	26,26	8,19

As seen above, although the Recall value is better when only capitalization rules are used, Precision value of the results which are obtained by using the LG approach & C.R. together is much higher. And also F-measure values for all β values (β = 1, β = 2, and β = 0.5) by using the LG approach & C.R. together is extremely larger those with using Capitalization Rule only.

4.1.3 DISCUSSION

The precision value that we obtained by using the LG approach & C.R. together is not very high because the final list of person names contains some NEs which can be categorized as organization names rather than person names. We could not distinguish these NEs from the final list of person names during the period of this thesis. However, we believe that the final list of person names can be cleared from these kinds of NEs by constructing different LG patterns.

Materials in which we failed to identify person names are given in Table 4-3 below.

Table 4-3 – Materials containing NEs which should have been categorized as organization names

The material preceding the person name	**Person name**	**Splitting character**	**The material succeeding the person name**	**Reporting verb**
Taksan Genel Müdürü	Aşır İman	,	A.A muhabirine yaptığı açıklamada, **Taksan**'ın, Ortadoğu ve Balkanlar'ın en büyük takım tezgahı üreten fabrikalarından biri olduğunu ve ürettiği ürünlerin de dış pazarlarda beğeni topladığını	kaydetti
Bir gazetecinin "**Botaş**'ın bölünmesine neden karşı çıktınız'" şeklindeki sorusuna karşılık	Demiralp	,	şu anda petrol yas asında hüküm bulunduğunu, **Botaş**'ın tekel durumunda bulunduğunu, yasa değişmediği sürece yeni kuruluşların kurulmasına imkan olmadığını, yasa değiştikten sonra dağıtım adı altında şirketlerin kurulacağını	bildirdi

Table 4 - 3 (continued)

The material preceding the person name	**Person name**	**Splitting character**	**The material succeeding the person name**	**Reporting verb**
Taksan'da fason üretimi ve iş hacmini artırmak için çalışmalar yaptıklarını belirten	İman	,	sektörde söz sahibi olan şirketlerle görüşmelerinin devam ettiğini	söyledi

Taksan and Botaş should have been identified as organization names. Recall that we tried to extract person names from materials returned by LG patterns. However returned materials did not always include person names but also organization names. For both French and English languages **[68][69]**, a list of LG patterns, which distinguishes organization names from person names exists but not for Turkish language. We aim to construct such patterns in the future and hope to improve our results.

Another simple example that indicates why we could not distinguish the organization names from person names is given below;

- **Ex: Ankara Büyükşehir Belediyesi**, Ankara'daki metro çalışmalarının son hızla devam ettiğini **açıkladı.**

As you see from the example above, the organization name "Ankara Büyükşehir Belediyesi" has similar structures (in terms of being capitalized and having same order of letters) with person names.

Although improvements are required to increase the recall and precision values, we would like to highlight that the LG approach has already been successfully applied for two languages: the NExtract, which is a NER system used for extracting NEs from Reuters English financial news in **[1]** and CasSys, which is a NER system applied in French texts in **[42]**.

CasSys achieved 91.9 Recall and 98.9 Precision and NExtract accomplished 90.0 Recall and 88.0 Precision. In fact, these successful NER systems in the literature also faced similar problems to ours. But eventually a list of such grammars was available as there were previous studies that focused on distinguishing person names from organization names. In **[70]**, for example, it is stated that they used IBM Intelligent Miner for Text in order to detect and distinguish place names, person names and common nouns. This kind of tool can be used to solve distinguishing problem. Also we believe that we can develop a tool that can distinguish these organization names from person names by using other local grammar patterns.

Another problem we encountered is that of RVs, which are not in our list and appear frequently with person names. Hence, further analysis is required to update our list. The materials containing such RVs are as follows:

- Tarım ve Köyişleri Bakanı **Gökalp** IMF'ye verilen son ek niyet mektubunda, hububat fiyatlarının en fazla yüzde 12 artırılacağı yolunda taahhütte bulunulduğunun hatırlatılması üzerine de "hayırlı olsun" demekle ***yetindi***.
- **İdris Taş** ise "Konteyner Taşımacılığı Antalya'da" başlıklı haber röportajı çalışmasıyla mansiyon ödülü ***kazandı***.
- **Pakdil**'e 1,5 trilyon liralık çek, Milli Piyango İdaresi Muhasebe Mali İşler Daire Başkanı **Ali Çetin** tarafından ***verildi***.
- Kütahya'nın Tavşanlı ilçe Kaymakamı **Sıtkı Hanlıoğlu**, ilçede tavşan üretimi yapacak küçük işletmelerin kurulması önerisinde ***bulundu***.

As seen above, the verbs “yetindi”, “kazandı”, “verildi”, and “bulundu” appeared with person names such as Gökalp, İdris Taş, Pakdil, Ali Çetin, and Sıtkı Hanlıoğlu. Although the use of “verildi” and “bulundu” verbs may not be sufficient to identify person names alone, the use of verbs with their respective collocates “tarafindan” and “onerisinde” may give more accurate results. However, it should be investigated in detail.

Finally, we did not explore compound RVs in this thesis although we observed that the number of such verbs is considerably high in our training corpus. For example, “altını çizmek”, is used in different forms such as: “altını çizen” and “altını çizdi”. Two of the materials that contain this kind of RV and the person name are given below.

- Bölge çiftçisinin borcuna sadık olduğunun **altını çizen Davaslı**, art arda gelen icralarla uyku dahi uyuyamayan çiftçinin, geleceğe ilişkin umutlarının tükendiğini aktardı.
- Uluslararası ilişkilerde özel sektörün rolünün **altını çizen Hombach**, Türk-Yunan işadamları arasındaki ilişkilerin güçlenmesi ile başlayan gelişmeleri buna örnek gösterdi.
- **Gökalp**, Türkiye'nin, tohumculuk endüstrisinin çağdaş gelişmeleri yakalaması ve çiftçilere intikal ettirmesi gerektiğinin **altını çizdi**.
- **Hombach**, Yugoslavya'nın ''Avrupa Evi''ne geri dönmesinin ve bunun sonucunda İstikrar Paktı'na katılacak olmasının, Paktın belirlenmiş fonlarının harcama yerlerini değiştirmeyeceğinin **altını çizdi**.

Although the compound RV “altını çizmek” is not in the list of RVs that we used in our study, it is used with PNs. Because neither the frequency list obtained by using the NooJ software nor our reference corpus covers them, we were not able to obtain two token words in this thesis. If we added these kinds of verbs to our list of RVs, we believe that we would get much better results.

In our comparison, recall that the naive PN extraction method produced higher recall compared to the LG approach, whereas the highest Precision value was obtained for the LG method. The reason of high recall value can be attributed to that of organization and title names retrieved by naïve method along with person names. Naïve method is not able to categorize PNs as person or organization names.

4.2 CONCLUSION

In this chapter, the LG approach to extract person names was tested on the economy news published by Anadolu Agency in January, 2001 and scored an 81.97% F-measure. We showed the cases that degrade our performance and explained the reasons.

When we compared the use of RVs in Reuter's corpus and Turkish financial news corpus, we realized an important fact: In Reuters corpus, the reporting verb "said" is most of the time preceded by PNs, such as person names and organization names however Turkish reporting verb "söyledi" is not always preceded by PNs. Many words can occur between PNs and RVs in Turkish, therefore it makes difficult to extract the PNs because the sentences can have different word ordering rules in Turkish.

Although this technique was successfully applied in English and French texts before, we did not obtain as good results as those reported in the literature. However, this study is important as it is the first study which applied the LG approach to Turkish financial texts. Our results are encouraging and yet can be improved by acquiring additional LG patterns.

CHAPTER 5

CONCLUSION & FUTURE WORK

The aim of this thesis is to recognize person names in the financial news by using the symbolic approach. In this symbolic approach, we tried to find the local grammars which exhibit the patterns of the materials which contain person names in the sentences. Therefore, we can call our symbolic approach local grammar approach. We used this local grammar approach in order to extract PNs, specifically person names, from the financial news. Up to here, we examined the syntactic (e.g. linear ordering of words in the sentences) behaviors or characteristics of person names and used this knowledge in recognition of person names in the financial news. In other words, we generate the patterns that describe the syntactic behaviors of words in the sentences in financial texts.

The first step in our study was preprocessing of the Economy Corpus of the year 2000 (EC2000) and creating the FEC2000 which is a cleaned form of the EC2000. Then, we constructed the frequency list of the FEC2000. The second step was extracting the Turkish RVs from the frequency list of the FEC2000 manually and testing the significance of these RVs by using statistical tests. The third step was recognizing NEs from the FEC2000. At this step, we extracted the patterns of materials that contain person names by performing concordance and collocation analyses. Then, we created the list of these patterns and tried to find person names from the FEC2000 by using this list. While extracting person names from the FEC2000, we performed capitalization rules and found other uses of person names. Next, we updated the list of person names. In the fourth step, we checked the weirdness values of one-token words that are not person names but have similar structures (in terms of being capitalized and having same order of letters) in the METU TC. Then, we removed the one-token words that have significantly lower values. In the fifth step, we removed the words that are not person names but have similar structure (in terms of being capitalized and having the same order of letters) and can be categorized as one of the following: continent name, country name, region name in Turkey, city name in Turkey, county name in Turkey, and international currency. Finally, we created the final list of person names from the FEC2000.

5.1 CONTRIBUTIONS

Our first contribution has been to construct the RV list for Turkish which is one of the most important tasks in NER because RVs have not been studied yet for Turkish language.

The second contribution has been to create a frequency list of words in the financial domain. There is also no study about the frequency list of words in Turkish financial domain. This contribution can be considered as an important contribution, because there are a few studies about the frequency list of words in Turkish and these studies do not cover as many words as our corpus. For example, our reference corpus, which is METU Turkish Corpus (METU TC), contains 1.889.080 words whereas our financial corpus (FEC2000) contains 11.518.306 words. In other words, the FEC2000 is six times larger than the METU TC. Moreover, our frequency list of words is eleven times larger than another frequency list of words of Göz **[71]**, which is one of the best studies about the frequency list of words in Turkish and contains about one million words.

In our study, we also tested the significance of the RVs where most of the studies that use the LG approach do not test the significance of RVs. For example, in **[1]**, The LG approach is used and achieved highly promising results although they did not test the significance of their RVs. They only use the ratio of the frequency of RVs in RFC (Reuter's Finance Corpus) to the frequency of RVs in BNC (British National Corpus) and assume that the RVs that have a high ratio cause the difference between two corpora. However, they did not specify how they chose a threshold value for the ratios. We believe that the significance of RVs should be based on a statistical test. If the significance of RVs is tested statistically, better results would be achieved. Performing this kind of significance test in this thesis can be considered as our third main contribution to the literature.

To sum up, the LG approach that we applied to the news in the financial domain is one of the primary studies in Turkish. Considering this fact, we can say that we have obtained highly encouraging results.

5.2 FUTURE WORK

After finding person names from the FEC2000, we realized that there are some words which have similar structure (in terms of being capitalized and having same order of letters) with person names but are not actually person names, rather they can be categorized as organization names in the final list of person names. These kinds of words should be removed from the final list of person names by using another LG. We believe that if these kinds of words are removed from the final list of person names, the precision of the LG approach will increase considerably.

We should also express that some of person names in final list of person names are in the form that contains their titles. These titles should be removed from person names. Another point to consider is that we used an abbreviation list, which is constructed manually by basing on the train corpus, while splitting the sentences from the texts in the EC2000 and this abbreviation list contains limited abbreviations. If we extend the abbreviation list, we believe that the splitter will run more efficiently and we will get better results. There are also person names that have the form as follows; M. Kemal Atatürk, A. Nejdet Sezer in Turkish. If we consider this kind of person names with abbreviation and update the splitter according to this criterion, we could achieve better results.

By constructing other LGs for organization names with the LG approach, we could also extract the organization names in the news from the financial domain.

We constructed our local grammar patterns for the financial news manually by using some analysis tools. However, we believe that we can develop a tool to extract patterns automatically from an unannotated corpus like NExtract in **[1]**. We would get better results by using these patterns to extract person names from the financial news.

In summary, we applied the LG approach to the news in the financial domain and achieved highly encouraging results. If the issues described above are structured in the future, we believe that we will get much better results.

REFERENCES

[1] H. N. Traboulsi, “Named Entity Recognition: A Local Grammar-based Approach,” unpublished Ph.D. dissertation, Department of Computing School of Electronics and Physical Sciences, University of Surrey Guildford, Surrey GU2 7XH, U.K, 2006.

[2] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik, “A Comprehensive Grammar of the English Language,” *London and New York, Longman*, 1985, p. 1024.

[3] D. J. Allerton, “The linguistic and sociolinguistic status of proper names,” *in Journal of Pragmatics,* Vol 11:1. pp. 61 - 92, 1987.

[4] W. Paik, E. D. Liddy, E. Yu, M. McKenna, “Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval,” *in Corpus Processing for Lexical Acquisition*, B. Boguarev, J. Pustejovsky, Ed., Bredford, pp. 61 – 73, 1996.

[5] N. Chinchor, “MUC-7 Named Entity Retrieved Definition,” 1997, Retrieved August 20, 2007 from http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/ne_task.html.

[6] H. A. Edelstein, “Introduction to Data Mining and Knowledge Discovery,” *Two Crows Corporation*, 1999, p. 4.

[7] M. A. Hearst, “Untangling Text Data Mining,” *in Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics*, University of Maryland, 1999.

[8] M. Hearst, “What Is Text Mining?,” Essay written in October 13, 2003, Retrieved August 20, 2007 from http://people.ischool.berkeley.edu/~hearst/text-mining.html.

[9] Liddy, E.D. “Enhanced text retrieval using Natural Language Processing,” *in Bulletin of the American Society for Information Science*,

Vol. 24, pp. 14 - 16, 1998, Retrieved August 20, 2007 from http://www.asis.org/Bulletin/Apr-98/liddy.html.

[10] U. Y. Nahm, R. J. Mooney, “A Mutually Beneficial Integration of Data Mining and Information Extraction,” *in Proceedings of the Seventeenth National Conference on Artificial Intelligence*, Texas, pp. 627-632, 2001.

[11] J. Robinson, C. Klassen, “Articles with Proper Nouns,” *Douglas College Learning Centre*, 2002, Retrieved August 30, 2007 from http://www.douglas.bc.ca/services/learning-centre/pdf/gr/GR1_21_Articles_with_Proper_Nouns.pdf

[12] G. Wei, “Named Entity Recognition and an Application to Document Clustering,” unpublished M.S. Thesis, Dalhousie University, Halifax, Nova Scotia, 2004.

[13] B. M. Sundheim, “Overview of Results of the Muc-6 Evaluation,” *in Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pp. 13 – 33, 1995.

[14] J. F. McCarthy and W. G. Lehnert, “Using Decision Trees for Coreference Resolution,” *in Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, IJCAI ’95, pp. 1050 – 1055, 1995.

[15] S. Soderlan, D. Fisher, J. Aseltine, and W. Lehnert “CRYSTAL: Inducing a Conceptual Dictionary,” *in Proceedings of the Eleventh National Conference on Artificial Intelligence*, IJCAI ’95, 1995, pp. 1314 - 1319.

[16] E. Riloff, “Automatically Constructing a Dictionary for Information Extraction Tasks*,” in Proceedings of the Eleventh National Conference on Artificial Intelligence*, pp. 811 – 816, 1993.

[17] D. M. Bikel, S. Miller, R. Schwartz, and R. Weischedel, “Nymble: a High-Performance Learning Name-finder,” *in Proceedings of the fifth conference on Applied natural language processing*, pp. 194 - 201, 1997.

[18] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman, “NYU: Description of the MENE Named Entity System as Used in MUC-7,” *in Message Understanding Conference (MUC-7)*, 1998.

[19] G. R. Krupka & K. Hausman “IsoQuest, Inc.: Description of the NetOwl Extractor System as Used for MUC-7*,” IsoQuest Inc. 3900 Jermantown Ave., Suite 400 Fairfax*, VA 22030.

[20] Retrieved July 10, 2007 from NetOwl official web site http://www.netowl.com/products/extractor.html.

[21] T. Bogers, “Dutch Named Entity Recognition: Optimizing Features, Algorithms, and Output,” unpublished M.S. Thesis, University of Van Tilburg, 2004.

[22] W. J. Black, F. Rinaldi and D. Mowatt, “Facile: Description of the NE System Used For MUC-7,” *in Proceedings of the 7th Message Understanding Conference*, Department of Language Engineering, UMIST PO Box 88, Sackville Street Manchester M60 1QD, United Kingdom, 1998.

[23] H. Cunningham, Y. Wilks, and R. Gaizauskas, “Gate - a General Architecture for Text Engineering,“ *in Proceedings of the 16th Conference on Computational Linguistics (COLINC-96)*, 1996.

[24] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, and Y. Wilks, “Experience of using Gate for NLP R&D,” *in Proceedings of the Workshop on Using Toolsets and Architectures To Build NLP Systems at COLING-2000*, Luxembourg, 2000.

[25] D. Nadeau & S. Sekine, “A survey of named entity recognition and classification,“ Ed. S. Sekine and E. Ranchhod, unpublished Technical Report, 2007.

[26] J. Kim, I. Kang, K. Choi, “Unsupervised Named Entity Classification Models and their Ensembles,“ *in Proceedings of the 19th international conference on Computational linguistics*, Vol. 1: pp. 1 - 7, 2002.

[27] K. Oflazer, Ö. Çetinoğlu, B. Say, “Integrating Morphology with Multi-word Expression Processing *in Turkish,” in Second ACL Workshop on Multiword Expressions: Integrating Processing*, pp. 64 - 71, 2004.

[28] G. Gür, D. Z. Hakkani-Tür, “Name Tagging Using Lexical, Contextual, and Morphological Information,” *in Proceedings of the Workshop on Information Extraction Meets Corpus Linguistics at Second International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, 2000.

[29] O. Mason, “Automatic Processing of Local Grammar Patterns,” *in Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics*, University of Birmingham, 2004, p.166 - 17.

[30] Rune Sætre, “GeneTUC: Natural Language Understanding Automatic Information Extraction from Biomedical Texts,” *in Proc.*

Computer Science Graduate Students Conference 2004 (CSGSC-2004), Norwegian University of Science and Technology (NTNU), 2004.

[31] Z. Harris, “A Theory of Language and Information: A Mathematical Approach,” *Oxford: Clarendon Press*, 1991, p.272.

[32] M. Gross, “Local Grammars and their Representation by Finite Automata,” *in Data, Description, Discourse, Papers on English Language in honour of John McH Sinclair*, M. Hoey, Ed. London: Harper - Collins Publishers, 1993.

[33] G. Nenadić, I. Spasić, “Recognition and Acquisition of Compound Names from Corpora,” *in NLP-2000, Lecture Notes in Artificial Intelligence*, Berlin, pp. 38 - 48, 2000.

[34] M. Mohri, “Local Grammar Algorithms,” *in Festschrift in Honour of Kimmo Koskenniemi on his 60th Birthday*, Antti Arppe et al. Ed. Stanford: CSLI Publications, pp. 26 – 38, 2005.

[35] H. Traboulsi, D. Cheng and K. Ahmad, “Text Corpora, Local Grammars and Prediction,” *in Proceedings of the 4th International conference on Language Resources and Evaluation (LREC2004)*, *in memory of Antonio Zampolli*, Vol. 3, pp. 749 - 752, 2003.

[36] G. Nenadić, I. Spasić and S. Ananiadou, “Reducing Lexical Ambiguity in Serbo-Croatian by Using Genetic Algorithms,” *in Proceedings of Fourth European Conference on Formal Description of Slavic Languages*, FDSL-4, Germany, pp. 287 – 298, 2001.

[37] J. Wedekind, “On Inference-Based Procedures for Lexical Disambiguation,” *in COLLING-1996*, pp. 980 - 985, 1996.

[38] G. Nenadić, “Local Grammars and Parsing Coordination of Nouns in Serbo-Croatian,” *in Text, Speech and Dialogue (TSD 2000), Lecture Notes in Artificial Intelligence*, 2000, Vol. 1902, Springer Verlag, pp. 57 - 62.

[39] M. Gross, “A bootstrap Method for Constructing Local Grammars,” *in Proceedings of the Symposium*, Contemporary Mathematics, University of Belgrade, pp. 229 - 250, 1999.

[40] J.-s. Nam & K.-s. Choi, “A local grammar-based approach to recognizing of proper names in Korean texts,” *in The Fifth Workshop on Very Large Corpora (WVLC-5)*, Ed. J. Zhou & K. Church, pp. 273 – 278, 1997.

[41] J .Baptista, “A local grammar of proper nouns,” *in Proceedings of the Seminários de Lingustica 2*, Faro: Universidade do Algarve, pp. 21 - 37, 1998.

[42] N. Friburger and D. Maurel, “Finite-State Transducer Cascade to Extract,” *in Implementation and Application of Automata, 6th International Conference*, CIAA 2001, LNCS 2494, pp. 115 - 124, 2000.

[43] Y. Almas and K. Ahmad, “LoLo: A System based on Terminology for Multilingual Extraction,” *in Proc. of COLING/ACL’06 Workshop on Information Extraction Beyond a Document, Sydney: Association for Computational Linguistics*, pp. 56-65, 2006.

[44] S. Brin, “Extracting patterns and relations from the world wide web,” *in Proceedings of the World Wide Web and Databases, International Workshop WebDB'98, Lecture Notes in Computer Science, Springer*, pp. 172 - 183, 1998.

[45] F. Smadja, “Retrieving collocations from text: Xtract,” *Computational Linguistics* Vol. 19:1, pp. 143 - 177, 1993.

[46] K. Ahmad, L. Gillam, and D. Cheng, “Textual and Quantitative Analysis: Towards a new, e-mediated Social Science,” *in Proc. of the 1st International Conference on e-Social Science*, 2005.

[47] J. Sinclair, R. Carter, “Corpus, Concordance, Collocation”, *Oxford University Press*, 1991.

[48] Kenneth W. Church, Robert L Mercer, “Introduction to the Special Issue on Computational Linguistics Using Large Corpora,” *in the 1999 Pacific Association for Computational Linguistics Conference*, Vol. 19:1, pp. 1 - 24, 1993.

[49] P. Hanks, "Evidence and intuition in lexicography," Ed. J. Tomaszczyk and B. Lewandowska - Tomaszczyk, *John Benjamins Publishing Company*, pp. 31 – 41, 1990.

[50] C. Chekuri and M. H. Goldwasser, P. Raghavan, E. Upfal, “Web Search Using Automatic Classification,” *in Proceedings of WWW-96, 6th International Conference on the World Wide Web*, 1996.

[51] S. R. Samoilovich, "Word Frequency Analysis as a Way to Improve Writing Quality," Retrieved July 10, 2007 from http://www.usingenglish.com/articles/word-frequency-analysis-as-way-to-improve-writing-quality.html.

[52] N. Shivakumar, H. Garcia-Molina, “SCAM: A Copy Detection Mechanism for Digital Documents,” *in Proceedings of 2nd International*

Conference in Theory and Practice of Digital Libraries (DL'95), Texas, 1995.

[53] A. Kilgarriff, "Using Word Frequency Lists to Measure Corpus Homogeneity and Similarity between Corpora," *in Proceedings Fifth ACL Workshop on Very Large Corpora*, Beijing and Hong Kong, 1997.

[54] K. Cosh and P. Sawyer, "Using Natural Language Processing Tools to Assist Semiotic Analysis of Information Systems," *in Corpus Linguistics as a conference poster*, 2003.

[55] E. J. L. Bell, "Collocation Statistical Analysis Tool: An Evaluation of the Effectiveness of Extracting Domain Phrases via Collocation," unpublished B.Sc. dissertation, Lancaster University, Edward J. L. Bell, 2007, p. 7.

[56] E. König, W. Lezius, and H. Voormann, "TIGERSearch 2.1 User's Manual," *IMS, University of Stuttgart*, 2003.

[57] L. Burnard and T. Dodd, "Introducing XAIRA: An XML-aware tool for corpus indexing and searching," *Research Technology Services, OUCS*, Retrieved July 10, 2007 from http://www.tei-c.org/Talks/OUCS/2004-02/Four/xaira.pdf.

[58] S. Hoffmann, S. Evert, "BNCweb (CQP-edition): The marriage of two corpus tools," *in Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, Sabine Braun, Ed., S. Braun, K. Kohn, and J. Mukherjee, pp. 177 - 195, 2005.,

[59] P. Nijkamp and J. Spronk, "Analysis of Production and Location Decision by Means of Multi-Criteria Analysis," *Elsevier Scientific Publishing Company*, pp. 285 - 302, 1979.

[60] K. Church, W. Gale, P. Hanks, and D. Hindle, "Using Statistics in Lexical Analysis*," in Lexical Acquisition: Using On-line Resources to Build a Lexicon*, U. Zernik Ed., Lawrence Erlbaum Associates, 1991.

[61] Linguistic Data Consortium - LCTL Team, "Simple Named Entity Guidelines for Less Commonly Taught Languages Version 6.5," 2006, p.2, *Technical Repot*, Retrieved August 25, 2007 from http://crl.nmsu.edu/say/SimpleNamedEntityGuidelinesV6.5.pdf.

[62] B. Say, K. Oflazer, U. Özge, and N. B. Atalay, "METU Turkish Corpus," Retrieved from METU Informatics Institute official web site http://www.ii.metu.edu.tr/~corpus.

[63] B. Say, D. Zeyrek, K. Oflazer, and U. Özge, "Development of a Corpus and a Treebank for Present-day Written Turkish," *in Proceedings*

of the Eleventh International Conference for Turkish Linguistics, Eastern Mediterranean University, Northern Cyprus, pp. 183 - 192, 2002.

[64] M. Silberztein, “Nooj Software,” Retrieved June 28, 2007 from http://www.nooj4nlp.net.

[65] D. Cheng, “Text Analysis User Manual Version 1.1”, 2006.

[66] “Redhouse English-Turkish dictionary,” *Sev Matbaacılık ve Yayıncılık*, 2004, İstanbul, Turkey.

[67] P. Rayson and R. Garside, “Comparing Corpora using Frequency Profiling,” *in Proceedings of the Workshop on Comparing Corpora, Thirty-eighth ACL*, Hong Kong, pp. 1 – 6, 2002.

[68] A. Douthat, “The Message Understanding Conference Scoring Software User's Manual”, *in The Message Understanding Conference*, Scoring Software User's Manual, 1998, Retrieved August 25, 2007 from
http://www.itl.nist.gov/iaui/894.02/related_projects/muc/muc_sw/muc_sw_manual.html.

[69] D. W. Oard, J. Gonzalo, M. Sanderson, F. Lopez-Ostenero, and J. Wang “Interactive Cross-Language Document Selection,” *Information Retrieval*, Vol. 7:1-2, pp. 205 - 208, 2004.

[70] N. Haas, R. Bolle, N. Dimitrova, A. Janevski, and J. Zimmerman, “Personalized News Through Content Augmentation And Profiling,” *in Proceedings of International Conference on Image Processing*, ICIP’02, IEEE Press, pp. 9 - 12, 2002.

[71] İ. Göz, “Yazılı Türkçenin Kelime Sıklığı,” *Türk Dil Kurumu (TDK)*, 2003.

APPENDICES

Appendix A: Most Frequent Top 1000 Words in the FEC2000

The table shows the most frequent top 1000 words in the FEC2000.

Words	**Frequency**
ve	169331
A	153383
bir	62070
da	60247
de	54776
ile	51032
yüzde	42248
bu	41379
Tipi	41177
Fon	40737
Değişken	39518
B	37622
Türkiye	37316
için	32666
bin	31426
İMKB	31242
İSTANBUL	30305
milyon	28558
nin	28333
olarak	27651
milyar	27485
lira	24730
göre	23585
F	23417
ANKARA	23377
en	22063
olan	21126

ise	20267
olduğunu	18790
Bu	18515
söyledi	17454
trilyon	16986
şöyle	16756
Müdürlüğü	16214
nın	15638
VE	15628
daha	15455
kadar	15368
yıl	15211
dedi	15154
ABD	14881
çok	14871
Genel	14831
içinde	14749
Başkanı	14636
yapılan	14330
işlem	13935
İstanbul	13060
Türk	12951
Bankası	12898
Yat	12875
ilk	12725
büyük	12684
Hisse	12400
Y	12378
saat	12318
Holding	12163
dolar	12000
ifade	11814
hisse	11694
tarihinde	11273
O	11097
bugün	10845
nda	10746
Fiyatı	10658
Ticaret	10562
arasında	10414
kaydetti	10349
İş	10265
tarafından	10131
Karma	10083
ın	10049
yaptığı	9915
in	9813
MZF	9754
değer	9736

yapılacak	9684
yeni	9660
Ş	9651
temin	9517
Kurulu	9347
ye	9344
Doları	9269
önemli	9257
bildirdi	9252
eden	9195
fiyatları	9030
açıklamada	9021
Başkanlığı	9006
Menkul	8712
Bank	8695
Fonu	8665
Tekstil	8607
teklif	8573
bedeli	8555
TLN	8448
toplam	8401
sonra	8389
devam	8341
ilgili	8324
edilebilir	8320
MTN	8251
etti	8204
ekonomik	8199
DA	8171
ortalama	8122
İhale	8069
Ankara	8017
kapalı	7807
Tahvil	7796
e	7783
Bono	7735
Kapanış	7694
DE	7685
ne	7673
Ulusal	7650
gibi	7630
Likit	7609
yüksek	7597
AB	7539
Borsası	7501
geçen	7486
belirterek	7470
oranında	7442
Sanayi	7406

Şartnamesi	7268
üzere	7249
son	7228
Yatırım	7224
belirten	7105
Avrupa	7035
Bakanı	6983
her	6956
yer	6943
ya	6909
gören	6860
Devlet	6852
a	6754
önceki	6741
aynı	6711
BAŞKANI	6698
konuştu	6676
nedeniyle	6675
gerektiğini	6662
satın	6661
seansta	6625
En	6602
olduğu	6588
M	6528
MAB	6425
saatleri	6424
ŞAY	6406
yılında	6375
nce	6359
puan	6324
alınacak	6215
adet	6204
TÜRKİYE	6200
YÜZDE	6193
nden	6121
ilişkin	6104
nde	6025
mesai	5995
konusunda	5962
Gıda	5923
yatırım	5891
iki	5838
Ege	5805
diğer	5800
Vakıf	5794
ayında	5754
Bakanlığı	5750
diye	5667
D	5619

belirtti	5616
ihale	5616
yılı	5591
Sigorta	5558
satış	5537
kapanış	5529
ton	5512
nun	5494
YATIRIM	5481
Endeksi	5477
Tarım	5460
yılın	5458
Odası	5455
önce	5441
İzmir	5417
na	5405
faiz	5400
İŞLEM	5397
Birliği	5291
ULUSAL	5282
Finans	5246
Garanti	5236
S	5220
den	5213
EGS	5205
bulunan	5203
oldu	5202
BEN	5190
özel	5152
ALINACAK	5140
dolarlık	5136
Anadolu	5103
T	5089
SEANS	5068
Yönetim	5042
E	4999
fazla	4996
bazı	4990
Kredi	4984
ülke	4972
ndan	4929
düşük	4903
BİR	4889
usulü	4879
ay	4861
Doğan	4855
Tel	4854
seans	4847
Kıymetler	4816

SEANSTA	4796
Fax	4795
üretim	4787
liraya	4761
mukabili	4755
kredi	4684
IMF	4664
olmak	4653
Bakan	4611
bildirildi	4571
katrilyon	4557
Önceki	4556
Senedi	4543
şunları	4527
Müdürü	4506
üzerine	4501
YENİ	4499
Seans	4495
Alman	4465
bugünkü	4453
bunun	4445
konusu	4432
ancak	4422
edilecek	4414
vergi	4404
İşlem	4395
DÖVİZ	4391
dan	4380
verilecek	4376
K	4330
Ocak	4319
Yüksek	4279
birlikte	4244
liralık	4234
ŞAH	4230
senedi	4213
iyi	4201
Merkez	4194
mali	4188
TL	4188
nca	4163
fiyatı	4146
un	4126
Dünya	4105
yeniden	4098
Çim	4054
Ziraat	4051
gelen	4035
üzerinde	4025

var	4017
Bölge	4006
kamu	4005
tüm	3925
veya	3925
geçici	3909
i	3896
enerji	3892
Enerji	3884
İHALE	3861
Çimento	3857
Yapı	3856
artış	3846
Hazine	3839
itibariyle	3828
iş	3815
aylık	3809
G	3804
C	3802
amacıyla	3789
sahip	3784
para	3776
yabancı	3772
kaydeden	3766
senetleri	3762
Koç	3752
sanayi	3729
şekilde	3729
İZMİR	3717
Özel	3713
yıllık	3688
günü	3675
alan	3658
Markı	3657
etmek	3635
GAP	3628
dolara	3609
hafta	3608
bağlı	3597
Akbank	3586
ihracat	3572
petrol	3560
puana	3552
dış	3533
yönelik	3530
Demir	3520
fiyat	3517
anlatan	3513
puandan	3511

ederek	3509
tarım	3506
ithal	3506
teminatı	3495
Düşük	3483
İÇİN	3478
edilen	3466
H	3466
liradan	3463
İlgili	3457
ticaret	3454
SERBEST	3436
BAKANI	3435
net	3424
Yeni	3421
Eylül	3386
muhabirine	3385
değil	3382
Sterlini	3381
YKY	3377
yaklaşık	3362
ta	3345
İngiliz	3344
bilgi	3342
İl	3337
Hold	3333
Elektrik	3324
TEB	3318
Gökalp	3305
ikinci	3305
İLE	3304
Sabancı	3302
Dış	3293
hizmet	3291
döviz	3289
olması	3288
MENKUL	3284
döneminde	3272
senetlerinin	3262
işaret	3258
Komisyonu	3257
DUYURULARI	3245
GAZİANTEP	3241
GENEL	3235
Nisan	3235
Kasım	3222
mektupları	3222
halinde	3221
genel	3220

olacak	3219
Frangı	3216
bulunduğunu	3211
Kalkınma	3205
SATIN	3188
ihraç	3182
MİLYON	3171
Ekim	3171
istiyor	3166
Ort	3163
Mayıs	3162
oranları	3161
söz	3159
Satın	3136
Mart	3123
uluslararası	3120
Alma	3110
Adı	3107
KISA	3100
TİCARET	3075
Endeks	3075
DEĞER	3072
şu	3070
Ancak	3069
Aralık	3050
itibaren	3049
zaman	3043
yılda	3034
daki	3027
deki	3018
Adana	3001
BANKASI	2997
yazılı	2982
dünya	2975
günlük	2968
Haziran	2963
ortaya	2935
Dr	2935
düzenlenen	2927
Teklif	2890
Belediye	2862
yerine	2861
çerçevesinde	2850
Ağustos	2847
enflasyon	2844
Ata	2843
uygun	2834
BORSA	2832
Halk	2829

verdi	2816
Petrol	2815
elektrik	2808
NİN	2806
ağırlıklı	2804
dönemde	2796
Mali	2788
sadece	2780
seanstaki	2778
kendi	2761
su	2755
bilgiye	2740
ALTIN	2740
YAPTIRILACAK	2738
Temmuz	2732
dikkat	2720
devlet	2717
ortak	2717
iç	2714
değişim	2710
Toprak	2709
vadeli	2708
gelir	2703
RA	2699
açısından	2694
ardından	2691
hem	2683
karşı	2666
BU	2660
PİYASA	2654
Şirketi	2652
artışla	2651
sonunda	2648
Şubat	2646
olmadığını	2645
kazandı	2642
İhlas	2624
Global	2620
çeşitli	2610
Üniversitesi	2609
işi	2604
ürün	2601
konuda	2596
açık	2583
yaşanan	2561
serbest	2560
GÜNLÜK	2557
gerçekleşti	2555
LİRA	2553

Bursa	2552
Cam	2548
BİN	2539
Yüzde	2526
Adedi	2524
sermaye	2521
Uluslararası	2514
Merkezi	2514
kapsamında	2514
altında	2511
RESMİ	2510
Fin	2506
DAĞ	2493
açıklamaya	2479
kabul	2476
CNS	2474
Çelik	2473
Gaziantep	2471
dün	2471
oranı	2471
GMYO	2467
HOLDİNG	2465
destek	2462
Geçen	2459
gerekli	2458
gün	2455
ürünleri	2451
alınan	2450
Bölgesi	2439
belirtildi	2436
Bileşik	2435
Yardımcısı	2428
TÜRK	2420
tek	2420
KURULU	2415
grubu	2401
kısa	2400
elde	2395
muhtelif	2390
mail	2386
üretimi	2382
ayı	2379
Rusya	2364
karar	2362
dile	2353
vurguladı	2352
Alternatif	2349
Lirası	2349
DOLAR	2348

Sermaye	2347
Osmanlı	2340
suretiyle	2339
yapan	2328
Pazar	2326
süre	2322
Pınar	2319
arasındaki	2314
yükseldi	2313
ait	2310
Başı	2306
Ecz	2305
hacmi	2303
az	2293
Antalya	2289
aldığı	2288
Koçbank	2279
usulüyle	2273
Derneği	2265
HACMİ	2259
almak	2255
arada	2252
özellikle	2251
rağmen	2251
ayrıca	2250
yandan	2245
edildi	2236
sosyal	2236
olacağını	2235
sonu	2232
basın	2231
Önal	2222
hakkında	2217
bulunuyor	2216
uzun	2214
kişi	2213
Maliye	2198
İŞ	2178
yapılması	2176
KREDİ	2169
ticari	2143
MİKTARI	2136
Sümer	2135
olumlu	2133
katılma	2131
vurgulayan	2124
bütün	2122
kaydedildi	2117
toplantısında	2114

yok	2111
Kardemir	2109
Mehmet	2104
SATIŞ	2101
Tic	2097
siyasi	2091
rekabet	2091
çıktı	2088
Japon	2077
ENDEKSİ	2077
DEVLET	2073
Ç	2071
geriledi	2067
te	2063
Konya	2062
FOTOĞRAFLI	2062
ülkenin	2057
TİPİ	2053
AP	2053
İsviçre	2052
Çin	2049
gelecek	2049
hisseler	2042
adetleri	2042
Dairesi	2038
seansın	2038
şirket	2036
Ağırlıklı	2035
Ortl	2035
Başbakan	2026
TARIM	2020
Yılmaz	2010
Şeker	2010
tipi	2008
Kaynak	2007
ELEKTRİK	2006
artarak	2004
önümüzdeki	1997
hükümetin	1990
DAN	1990
İplik	1988
Prof	1986
EN	1983
alış	1982
Tofaş	1975
üç	1975
verilen	1970
Bir	1967
Almanya	1958

mümkün	1952
Bugün	1941
Kayseri	1936
geri	1935
muhabirinin	1934
MALZEME	1934
Anonim	1930
puanlık	1928
sıra	1925
Resmi	1922
işbirliği	1922
gerçekleştirilecek	1917
önem	1914
yapıldı	1913
bulunduğu	1909
Aydın	1908
sektör	1908
lirası	1900
aldı	1899
banka	1898
mark	1897
Boya	1894
YKB	1892
Oto	1892
küçük	1889
IN	1887
konuşmada	1886
düşüşle	1881
başka	1880
sektöründe	1879
doğru	1878
TOPLAM	1877
edecek	1875
ele	1874
zarf	1873
Gazete	1870
yanı	1867
ORANI	1866
Demirbank	1865
karşısında	1861
proje	1859
TRİLYON	1859
DEN	1858
sonuna	1858
Kir	1857
ihtiyacı	1853
Dışbank	1849
Giyim	1842
bankaların	1836

Bunun	1832
Alarko	1830
com	1829
Müsteşarlığı	1827
Denizbank	1825
ettiğini	1824
TOBB	1823
TEDAŞ	1818
DIŞ	1818
nedenle	1815
mevcut	1808
Turizm	1805
akşam	1803
Alfa	1801
yol	1800
İdaresi	1794
olduğuna	1794
şirketin	1791
Interbank	1789
sonucunda	1785
tam	1784
Cumhurbaşkanı	1784
Ekonomik	1782
Hastanesi	1776
İLK	1774
Makina	1773
yıllarda	1768
alanda	1765
Ereğli	1764
halka	1761
Döner	1759
kalem	1755
sektörünün	1754
sözlerine	1751
bankacılık	1748
Öte	1747
Kronu	1747
İran	1744
gıda	1741
anlattı	1740
Tahmini	1739
Fiyat	1734
PETROL	1734
talep	1733
MİLYAR	1732
çalışmaları	1732
belirlenen	1732
lik	1732
doğalgaz	1732

hale	1730
Biz	1728
BORSASI	1727
o	1726
geçerli	1726
çalışma	1716
KAYSERİ	1714
Yalova	1713
ki	1712
başına	1711
temel	1711
dikkati	1704
BUGÜN	1700
neden	1699
ADI	1699
Tüpraş	1696
alım	1695
yoğun	1692
YE	1691
veren	1688
NIN	1686
özelleştirme	1686
Borsa	1685
Grubu	1685
İç	1683
bölge	1683
Dağıtım	1681
Demirel	1678
gereken	1676
Gedik	1673
DÜNYA	1667
ciddi	1667
ekledi	1663
mücadele	1663
Londra	1662
zarar	1662
Bakanlar	1662
Et	1659
bunu	1658
ana	1658
FM	1656
Başkan	1655
kararı	1654
başladı	1651
faaliyet	1650
piyasa	1644
ek	1636
Euro	1633
KOBİ	1632

verdiği	1632
konut	1632
mal	1629
Fransız	1627
İşletme	1622
dışında	1620
Yabancı	1617
aracı	1616
olumsuz	1609
teşvik	1609
programın	1608
ARACI	1608
Gr	1607
GIDA	1606
kez	1605
bunların	1604
artık	1599
MÜDÜRÜ	1598
ALIŞ	1596
Ayrıca	1595
dünyanın	1593
kurum	1589
LEY	1586
Bayındır	1586
Ersümer	1582
eğitim	1580
konuşan	1580
Kurumu	1573
Doğu	1569
belirlendi	1568
ayrı	1567
KURUM	1566
Milli	1565
ihracatı	1559
çalışmalar	1558
Eczacıbaşı	1557
kar	1551
ı	1551
biri	1550
Hollanda	1543
toplantıda	1542
olduklarını	1540
ülkelerin	1539
verildi	1538
sektörün	1534
Servisi	1532
Oral	1530
çeken	1527
tekstil	1526

Kurul	1526
yana	1523
tüketici	1523
Su	1523
YIL	1522
Gübre	1519
sırasında	1518
miktarlarda	1514
NEW	1510
alınması	1510
Ali	1509
Japonya	1507
sektörü	1505
Müdürlük	1505
HİSSELERİN	1503
Yayın	1502
yönetim	1502
yetkilileri	1498
teknik	1498
deprem	1496
firma	1494
geldiğini	1493
dahil	1490
yakın	1489
BURSA	1487
ziyaret	1487
böyle	1487
yarın	1485
Belediyesi	1482
yılının	1482
durumunda	1480
Sabah	1478
eski	1476
SIRALAMASI	1476
istikrar	1475
sonucu	1469
yi	1469
YORK	1468
edilmesi	1468
Emek	1467
altın	1466
TOKYO	1466
KONYA	1465
hiç	1465
görev	1464
TA	1464
İN	1464
uygulanan	1462
istedi	1459

kurulu	1457
Ofisi	1456
HSBC	1455
ANTALYA	1455
üzerinden	1454
Söz	1454
Ama	1453
Irak	1453
saatlerinde	1452
grup	1450
ENERJİ	1449
Güney	1445
kaybetti	1444
içerisinde	1443
ERC	1443
hiçbir	1441
seansa	1437
devletin	1435
geçti	1434
MUHTELİF	1432
türlü	1431
Net	1431
başarılı	1428
dışı	1428
Buna	1426
pamuk	1425
şirketi	1425
GAZETE	1421
buna	1420
Açıklamada	1419
programı	1412
anda	1411
bedelsiz	1410
ediyor	1410
merkezi	1409
değişiklik	1408
Büyükşehir	1407
bundan	1407
MRY	1406
mevduat	1405
Altın	1405
Asya	1404
ULUSLARARASI	1403
açıkladı	1403
PAZAR	1402
yapıldığını	1402
Pazarlama	1401
hayvan	1399
Aksu	1397

TSKB	1396
doları	1395
Şartname	1395
durumda	1393
oluşan	1392
cins	1390
SANAYİ	1390
sayısında	1389
üyesi	1387
konulu	1386
ödeme	1385
halen	1384
HİSSE	1384
Cuma	1383
Şube	1382
DEVAM	1381
pazar	1377
ettiği	1377
Komisyon	1374
Güneydoğu	1373
çekti	1369
Eğitim	1369
Tedarik	1366
düzenlediği	1364
endeksi	1360
Böylece	1359
FİYATLARI	1357
İNŞAAT	1356
yapmak	1354
artışı	1354
Ahmet	1352
Sosyal	1351
Toskay	1351
Köyişleri	1350
Kentbank	1350
borç	1349
il	1348
Seramik	1346
TOPLANTISI	1346
Van	1344
Kanada	1342
şeklinde	1338
ÇBS	1336
İşleri	1333
payı	1333
gümrük	1332
sistemi	1330
hatırlatan	1330
Zorlu	1329

Sönmez	1327
EKONOMİDEN	1326
boru	1326
Tanrıkulu	1324
bazında	1324
istihdam	1323
CHÇ	1323
enflasyonun	1322
ama	1322
İnşaat	1322
KAZANAN	1321
et	1321
kalan	1315
değeri	1314
Ambalaj	1314
altına	1311
Sektör	1309
hızlı	1308
işlemler	1305
inşaat	1305
Kanunu	1303
düştü	1300
YA	1299
İşletmesi	1299
Aslan	1295
Borusan	1294
bile	1293
Daha	1291
ARTIŞ	1289
TEKSTİL	1286
p	1286
yeterli	1284
ayda	1284
otomobil	1284
hükümet	1283
Selçuk	1283
kazanan	1282
SON	1282
dikkate	1280
nün	1279
haline	1278
işçi	1278

Appendix B: Frequency Distribution of Top 100 Reporting Verbs

The table shows the frequency distribution of top 100 reporting verbs in Economy Corpus (EC2000) and METU Turkish Corpus (METU TC) which is taken as reference corpus. The frequency is rounded to frequency per million words token following the approach in **[1]**. Before rounding the frequencies, the EC2000 contained 11.518.306 tokens and the METU TC contained 1.889.080.

Lemma	**The EC2000**	**The METU TC**
demek	1.713	4.876
söylemek	463	2.253
istemek	695	2.162
konuşmak	881	1.494
düşünmek	206	1.380
belirtmek	1.692	1.182
yazmak	43	1.156
sormak	146	1.029
anlatmak	318	1.007
açıklamak	854	822
sağlamak	966	741
görüşmek	192	508
kabul etmek	181	502
kaydetmek	1.541	488
belirlemek	505	448
sürdürmek	298	433
tartışmak	98	413
korumak	167	395
değerlendirmek	353	385
ifade etmek	728	362
savunmak	169	339
vurgulamak	557	328
sunmak	288	322
karar vermek	92	306
dikkat çekmek / dikkati çekmek	351	250
dilemek	45	236
eklemek	162	234
dile getirmek	195	203
bitirmek	56	197

Lemma	The EC2000	The METU TC
aktarmak	95	192
bağırmak	3	170
bildirmek	426	163
yanıtlamak	93	154
uyarmak	71	139
eleştirmek	40	131
tanımlamak	28	131
önermek	63	122
hatırlatmak	257	120
karşı çıkmak	21	117
reddetmek	16	117
iddia etmek	32	113
öne sürmek	67	108
yorumlamak	15	107
değinmek	101	99
ileri sürmek	50	94
yanıt vermek	13	89
tahmin etmek	71	84
duyurmak	53	83
anımsatmak	16	82
sıralamak	59	82
tekrarlamak	13	64
açıklama yapmak	32	63
cevap vermek	18	59
kararlaştırmak	69	58
yinelemek	4	58
ilan etmek	45	57
teslim etmek	49	56
bilgi vermek	123	54
sevk etmek / sevketmek	10	50
talep etmek	45	49
doğrultmak	6	49
işaret etmek	49	41
söz vermek	8	41
haber vermek	2	39
itiraf etmek	1	35
buyurmak	1	35
dayatmak	6	33
kabullenmek	3	33
ısrar etmek	4	32
haykırmak	0	32
yalanlamak	2	29
teşvik etmek	51	26
övünmek	2	26
protesto etmek	12	26
karşılık vermek	2	26
rica etmek	2	26
beyan etmek	14	23

Lemma	The EC2000	The METU TC
temin etmek	144	17
emretmek	1	17
cevaplamak	8	16
sonuçlandırmak	22	14
teyid etmek / teyit etmek	7	13
yorum yapmak	4	13
arz etmek	29	12
görüş bildirmek	5	11
diretmek	1	10
tekrar etmek	3	8
hayret etmek	0	8
sevketmek	11	7
açıklık getirmek	9	7
taahhüt etmek	14	7
umut etmek	3	5
emir vermek	1	5
inkar etmek	2	4
sonuç çıkarmak	1	3
fikir vermek	0	3
ikaz etmek	1	2
feryat etmek	0	2
deklare etmek	1	2

Appendix C: Maximum Likelihood Values of Top 100 Reporting Verbs

The table shows the maximum likelihood values of top 100 reporting verbs for Economy Corpus (EC2000) and METU Turkish Corpus (METU TC) which is taken as reference corpus.

Reporting Verbs	Maximum Likelihood Values
yazmak	6.137,11
demek	5.887,73
söylemek	5.126,36
düşünmek	4.044,80
sormak	3.126,20
istemek	2.973,46
kaydetmek	1.661,16
anlatmak	1.419,74
bağırmak	1.030,87
tartışmak	818,44
kabul etmek	583,59
dilemek	568,80
konuşmak	554,51
görüşmek	544,65
karar vermek	460,60
ifade etmek	376,11
reddetmek	357,72
bildirmek	356,67
korumak	344,24
yorumlamak	335,66
bitirmek	320,79
temin etmek	312,29
karşı çıkmak	302,80
tanımlamak	287,02
belirtmek	284,20
yanıt vermek	270,50
yinelemek	254,82
haykırmak	206,46
savunmak	205,21
anımsatmak	198,71
eleştirmek	197,52
itiraf etmek	195,54

Reporting Verbs	Maximum Likelihood Values
buyurmak	195,35
haber vermek	183,32
iddia etmek	183,03
vurgulamak	182,72
doğrultmak	168,74
hatırlatmak	151,77
tekrarlamak	149,37
kabullenmek	137,90
yalanlamak	129,08
övünmek	122,84
rica etmek	119,54
aktarmak	119,15
sevk etmek / sevketmek	115,39
karşılık vermek	104,99
ısrar etmek	98,77
söz vermek	95,88
Sağlamak	93,63
Sürdürmek	85,22
Dayatmak	85,16
cevap vermek	83,79
bilgi vermek	81,23
Uyarmak	80,45
Emretmek	75,27
Önermek	65,80
Yanıtlamak	53,64
dikkat çekmek / dikkati çekmek	53,08
ileri sürmek	49,65
Diretmek	45,38
Eklemek	44,70
hayret etmek	37,82
açıklama yapmak	37,11
öne sürmek	33,69
teşvik etmek	23,82
Duyurmak	22,39
Duyurmak	20,66
arz etmek	20,63
protesto etmek	19,16
yorum yapmak	18,57
emir vermek	15,47
Sıralamak	12,10
Belirlemek	11,05
Cevaplamak	10,29
tekrar etmek	10,12
fikir vermek	9,88

Reporting Verbs	Maximum Likelihood Values
beyan etmek	8,14
feryat etmek	7,03
taahhüt etmek	6,96
sonuç çıkarmak	6,83
teyid etmek / teyit etmek	6,33
sunmak	6,29
görüş bildirmek	5,86
sonuçlandırmak	5,42
ilan etmek	5,03
değerlendirmek	4,61
tahmin etmek	3,51
kararlaştırmak	3,46
sevketmek	2,21
işaret etmek	1,95
açıklamak	1,91
teslim etmek	1,82
inkar etmek	1,61
ikaz etmek	1,33
umut etmek	1,32
deklare etmek	1,00
talep etmek	0,74
dile getirmek	0,51
açıklık getirmek	0,28
değinmek	0,09

Appendix D: Descriptions of Columns in Concordance Lines

Sequence Column: It contains the phrases that including the proper noun and reporting verb.

Before Column: It contains the phrases preceding the sequence column.

After Column: It contains the phrases following the sequence column.

Appendix E: Concordance Lines of First Type of Verb Form Including Seven Reporting Verbs

The table shows the concordance lines of first type of verb form including seven reporting verbs.

The concordance line	The material preceding the reporting verb	The reporting verb
1	Hükümetin uyum ve istikrar sergilediğini belirten Arsan, bu faktörün ülkenin ekonomik ve sosyal alanda gelişmesini hızlandıracağını	açıkladı.
2	Gates, korku senaryolarında canlandırılan felaketlerin olmayacağını, bununla birlikte, gelecek aylar içinde bilgisayar sistemlerinde çok sayıda küçük aksaklığın görüleceğini	belirtti.
3	Çin Devlet Kalkınma Planlaması Komitesi Başkanı Zeng Peiyan, Çin'in yakın bir gelecekte, para birimi yuanda devalüasyon uygulamasına gerek olmadığını	belirtti.
4	İzmir Kahveciler Odası Başkanı Mustafa Ak, yaptığı açıklamada, İzmir'deki kahvehanelerde uygul anacak fiyat listesinin 3 Ocak Pazartesi gününden itibaren geçerli olduğunu	bildirdi.
5	Trabzon Ticaret ve Sanayi Odası (TTSO) Yönetim Kurulu Başkanı Şadan Eren, konaklama tesisleri ile yörede konferans turizminin de geliştirileceğini	bildirdi.
6	Müstakil Sanayici ve İşadamları Derneği (MÜSİAD) Genel Başkanı Ali Bayramoğlu, ya pısal reformlar gerçekleştirilmeden, 2000 yılında öngörülen enflasyon hedeflerine ulaşılamayacağını	bildirdi.
7	Aksaray Valisi Emir Durmaz, "Aksaray'ın geri kalmışlıktan kurtulması, emsal iller seviyesine yükselmesi için süper teşvikli iller arasına alınması gerekir"	dedi.
8	Şimşek yaptığı açıklamada, "1998 yılında yaşanan sel felaketleri nedeniyle pamuk üreticisi zaten verimli ürün alamadı. Beklentimiz olan 20 centlik primin yanına dahi yaklaşılmadı"	dedi.
9	Düzel, kurmayı planladıkları fabrikada yılda 10 bin asansör makinesi, 100 bin kilit ve diktatör, 3 bin 500 kabin, 20 bin kapı ve 80 bin buton üretmeyi hedeflediklerini	ifade etti.

The concordance line	The material preceding the reporting verb	The reporting verb
10	Bayram, şu ana kadar yapılan denetimlerde, mali durumu bozuk başka bir bankanın tespit edilmediğini	ifade etti.
11	Işık, bu satıştan elde edilen geliri, Işıklar İnşaat Malzemeleri A.Ş'nin yurtiçi tuğlaya yönelik yatırımlarının finansmanında kullanacaklarını	ifade etti.
12	Tahsin Sancak, A.A muhabirine yaptığı açıklamada, çay sektörünün reforma tabi tutularak, dünyaya açılması gerektiğini	kaydetti.
13	Hasan Özmen, pamuğun 206.4 trilyon ve yüzde 29'luk payla birinciliği koruduğuna dikkati çekerek, bitkisel yağların 96.5 trilyon liralık işlemle toplam işlem hacminin yüzde 13.6'sına sahip olduğunu	kaydetti.
14	Sivas'ın Türkiye'nin en mamur şehirlerinden biri olduğunu belirten Demirel, Sivaslı işadamlarının il'e yatırım yapmalarını sağlayabilmek amacıyla düzenlenecek kurultaya kendisinin de katılacağını	kaydetti.
15	Sivas Sanayi Odası Başkanı Emin Doğan ise Sivas'ın sorunlarını anlattığı konuşmasında, Sivas'ın geri kaldığını, bundan kurtulabilmek için büyük çaba içinde olduklarını	söyledi.
16	Cumhurbaşkanı Süleyman Demirel ise Sivas'ın sorunlarını bildiklerini, Sivas'taki şevk ve gelişmenin gerçekten heyecan verici olduğunu	söyledi.
17	Sanayi ve Ticaret Bakanı Ahmet Kenan Tanrıkulu, Türkiye'nin son 10 yıldır her gün yeni bir oluşum beklentisi içinde istikrarsız bir dönem geçirdiğini	söyledi.
18	BIS yetkilisi, bütün bunlara rağmen, 2000 Yılı Problemi'ne ilişkin asıl sonucun, gelecek hafta bankaların açılmasıyla ortaya çıkacağını	söyledi.
19	Castrol Türkiye Genel Müdürü Ömer Dormen ise bu işbirliğinin kalite, performans ve teknolojiyi ön plana çıkaracağını	ifade etti.
20	Cumhurbaşkanı Demirel de "Bu paneli yapmanız gayet iyi olur. Bakalım, benim zamanıma uyuyorsa katılırım"	dedi.

First type of verb form: Examples of past tense reporting verbs are “dedi”, “belirtti”, “kaydetti”, “açıkladı”, “ifade etti”, “söyledi”, and “bildirdi”.

Appendix F: Concordance Lines of Second Type of Verb Form Including Seven Reporting Verbs

The table shows the concordance lines of second type of verb form including seven reporting verbs.

The concordance line	The reporting verb	The material succeeding the reporting verb
1	söyleyen	Arsan, "2000 yılı toplam satış hedefimiz 2 milyonu aşkın hindi satmaktır. 2000 yılında piyasaların canlanmasına paralel bu satış hedefini de aşabiliriz" dedi.
2	ifade eden	Aydın, Bayındırlık Bakanlığı'nın depremle birlikte öne çıktığını vurguladı.
3	açıklayan	Bakan Keçeciler, artık gümrük kapılarında 20 günde yapılabilen yüklemelerin, Avrupa Gümrük kapılarında olduğu gibi 3 saatte yapılabileceğini kaydetti.
4	açıklayan	Bakan Koray Aydın, "İkiztepe-Konak Doğanlar Otoyolu İzmir kent geçişi konusunda, topladığımız bilgiler ışığında Ankara'ya gidip değerlendirme yapacağım" dedi.
5	söyleyen	Başbakan yardımcısı, tahıl rekoltesinin önceki yıla göre 2 misli arttığını da belirtti.
6	diyen	Chhibber, bu programın başarısı için özel sektör ve tüm toplumun katılımının gerekli olduğunu ifade etti.
7	kaydeden	Daloğlu, yarın işçilerle sözleşme imzalayacaklarını, 434 işçinin 25 Ocak Salı günü iş başı yapacağını söyledi.
8	belirten	Dünya Bankası yetkilileri, reform ile tarım sektöründekilere gerçek anlamda bir tarımsal desteğin sağlanacağını ve verimlilliğin de artacağını vurguluyorlar.
9	ifade eden	Özyürek, "Böyle giderse hiçbir şekilde vergi toplayamayacağız. Çünkü vatandaş ödediği verginin yerine gidip gitmediğini görmek istiyor" dedi.
10	bildiren	Özyürek, "Enflasyonu aşağı çekebilmek vergi ile doğrudan bağlantılı. Oysa biz, bir milyar lira kazanandan da, bir milyon lira kazanandan da aynı oranda vergi alıyoruz" dedi.

The concordance line	The reporting verb	The material succeeding the reporting verb
11	kaydeden	Prof. Dr. Yalçın, "Makineleşmede traktör sayısı tek başına anlam taşımıyor. Çünkü, sadece çekici güç yaratıyor" diye konuştu.
12	belirten	Shigematsu, ilk çeyrekte alınacak verilerin, yeni ekonomik programın etkileri konusunda açık bir işaret vermeyeceğini kaydetti.
13	diyen	Sönmez, "eskiden çiftçi bir kilogram pamuk satarak 1 litre mazot alıyordu. Bu gün 2 kilogram pamuk satıp, 1 litre mazot alabiliyor".
14	belirten	üreticiler, şunları söylediler: "Geçtiğimiz yıllarda teslim ettiğimiz ürüne karşılık bir miktar avans ödeniyordu. Bu yıl ise hiç para ödenmedi.İki aydan bu yana para almak için bekliyoruz".
15	belirten	yetkililer, bu yıl ise 25 milyon metrekare fayans ve yer karosu ihraç etmeyi hedeflediklerini bildirdiler.
16	kaydeden	yetkililer, Türkiye'de ise yüksek maliyetler ve finansal sorunlar nedeniyle bu sektörlerin sorunlu bir dönem geçirdiğini dile getirdiler.
17	ifade eden	Tes-İş Sendikası Başkanı Mustafa Kumlu ise şöyle konuştu:

Second type of verb form: Examples of reporting verbs ending with "-en" suffix are "diyen", "belirten", "kaydeden", "açıklayan", "ifade eden", "söyleyen", and "bildiren".

Appendix G: Concordance Lines of Third Type of Verb Form Including Seven Reporting Verbs

The table shows the concordance lines of third type of verb form including seven reporting verbs.

The concordance line	The material preceding the reporting verb	The reporting verb
1	Kadayıf üreticisi Şevket Kılıç, A.A muhabirine, kentte yazın 3-5, kış aylarında ise 200 işletmede üretim yapıldığını	açıklayarak
2	Avcı, en kötü ihtimalle bir kuşun yılda 30 yavrusunun olacağını	belirterek
3	İl Tarım Müdürlüğü yetkilileri, yerli hayvan ırklarının ıslah edilmesi için her türlü olanağın en iyi şekilde değerlendirilmesine çalışıldığını	belirterek
4	A.A muhabirine bilgi veren ÇATES yetkilileri, enerji üretiminde 1998 yılına göre 1999'da 7 milyon 127 bin kilowatsaat düşüş olduğunu	belirterek
5	Borsa uzmanları, hızlı çıkışta yeni yılla birlikte faizlerdeki ani düşüşlerin önemli ölçüde etkili olduğunu	belirterek
6	Vatandaşlar, balık fiyatlarının 500 bin lira ile 2.5 milyon lira arasında olduğunu	belirterek
7	Üreticiler, Mürefte beldesindeki Zeytin Tarım Satış Kooperatifleri'ne ürünlerini teslim ettiklerini	belirterek
8	Borsa uzmanları, satış baskısına rağmen piyasanın dengeli bir seyir izlediğini	belirterek
9	Oturum Başkanı Prof. Dr. İlhan Özay ise imtiyaz usulüne karşı olmadığını	belirterek
10	Avrupa Birliği (AB) Uzmanı Jean François Drevet, Türkiye'de bölgeler arasındaki gelişmişlik farkının çok yüksek olduğunu	belirterek
11	İstanbul Teknik Üniversitesi (İTÜ) Öğretim Üyesi Prof. Dr. Kutsal Tülbentçi de, İstanbul'da kaçak ve standart dışı yapılara yönelik çok sayıda imar affı çıktığını	belirterek
12	Prof. Dr. Tülbentçi, İstanbul'daki yapılarda beton kalitesinin düşük olduğunu	belirterek
13	İTÜ Öğretim Üyesi Tevfik Sena Arda ise çelik yapıların betonarme yapılara göre hafif olduğunu	belirterek
14	Vali Türk, Iğdır-Tuzluca güzergahında bulunan 20 bin dönümlük arazinin organize sanayi bölgesinin kurulması için tahsis edildiğini	diyerek

The concordance line	The material preceding the reporting verb	The reporting verb
15	Gebze Yüksek Teknoloji Enstitüsü (GYTE) Çevre Mühendisliği Bölüm Başkanı Prof. Dr. Mehmet Karpuzcu, A.A muhabirine yaptığı açıklamada, dünyadaki su, petrol ve doğalgaz kaynaklarının sınırlı olduğunu	diyerek
16	DSİ yetkilileri barajlarda depolanan su miktarının mevsim itibariyle uygun düzeyde bulunduğunu	diyerek
17	Uzmanlar, alternatif piyasaların faiz cephesinde dengelerin oturmasıyla birlik te Borsa'da düşüş yönünde yaratılan beklentilerin zayıflamaya başladığını	ifade ederek
18	Enerji ve Tabii Kaynaklar Bakanı ve Başbakan Yardımcısı Cumhur Ersümer, enerji meselesinin uzun yıllara sarih edecek bir mesele olduğunu	ifade ederek
19	TZOB Başkanı Faruk Yücel de TBMM'nin bu dönemde iyi çalıştığını	ifade ederek
20	Tansaş Gıda Şirketi yetkilileri, restoranın klasik havasının korunmasına özen gösterdiklerini	ifade ederek
21	Holding Yönetim Kurulu Başkanı Üzeyir Garih ise iş yaşamına 49 yıl önce Carrier'de başladığını	ifade ederek
22	Bakan Aydın, hükümetin attığı olumlu adımlarla Türkiye'nin yıllardır çektiği siyasi istikrarsızlığın atlatıldığını	ifade ederek
23	Borsa uzmanları, işlem hacminin 18,500 direncini kırmakta yeterli olmadığını	ifade ederek
24	Doç. Dr. İlyas Yılmazer ise ovaların, depremler sonucu oluşan ulusal servet ler olduğunu	ifade ederek
25	Bakan Önal, mazot ve sınır ticareti konusundaki bir soru üzerine, İran'ın, "mazot taşıyan kamyonların mazotunu başka ülkelerden aldığını ve İran'dan transit olarak geçtiğini" iddia ettiğini	kaydederek
26	Türk-Güney Kore İş Konseyi Başkanı Ali Kibar, Türkiye ile Güney Kore'nin sanayileşme süreçleri arasında pek çok benzerlikler bulunduğunu	kaydederek
27	Renault Yönetim Kurulu Başkanı Luis Schweitzer ise ekonomide iş çevrelerinin rekabetçi bir ortamda, azami kar elde etmeye çalıştığını	kaydederek
28	Yetkililer, büyükbaş hayvanların canlı olarak kilosu 1 milyon 150 bin liradan satılacağını	kaydederek
29	KKTC Başbakanı Derviş Eroğlu, bankalar konusunda Türkiye ile temasların sürdüğünü	söyleyerek
30	Mudiler adına konuşan Namık Ramadan da Yurtbankzedeler'in çok zor durumda olduğunu	söyleyerek

Third type of verb form: Examples of reporting verbs ending with “-erek” suffix are “diyerek”, “belirterek”, “kaydederek”, “açıklayarak”, “ifade ederek”, “söyleyerek”, and “bildirerek”.

Appendix H: Concordance Lines of Fourth Type of Verb Form Including Six Reporting Verbs

The table shows the concordance lines of third type of verb form including six reporting verbs.

The concordance line	The material preceding the reporting verb	The reporting verb
1	Özince,	sözlerini şöyle sürdürdü:
2	Kutan'ın bu sözleri üzerine ATO Başkanı Aygün'ün, "yani hükümete destek oluyorsunuz" şeklindeki sorusunu da yanıtlayan Kutan,	sözlerini şöyle sürdürdü:
3	Voegele,	sözlerini şöyle tamamladı:
4	Kış aylarının tüm dünyada sert geçmesi nedeniyle ham petrol fiyatlarının yükselme eğiliminde olduğuna dikkati çeken Turgut Bozkurt,	sözlerini şöyle tamamladı:
5	Tarımın ulusal ekonominin temelini oluşturduğunu vurgulayan Prof. Dr. Tekinel,	sözlerini şöyle tamamladı:
6	Ticaret merkezinin her türlü altyapısını tamamlamış olmasına karşın, tam kapasiteyle çalışamadığına işaret eden Dalan,	şöyle dedi:
7	Uzunhekim, Dutilhlere ait denizcilik işletmesinin yanı sıra seyahat acentasının da kendilerine devredileceğini belirterek,	şöyle dedi:
8	Özince,	şöyle dedi:
9	Yat Limanı inşaatı için ortaya çıkan taş ocağı sorununun da giderilmesi için çalışıldığını belirten Ergül,	şöyle konuştu:
10	Ericsson Türkiye Başkan Yardımcısı Stefan Ofverholm da Türkiye'nin Ericsson'un dünyadaki en önemli pazarlarından biri olduğunu vurgulayarak,	şöyle konuştu:
11	"Ülkemizin, geleceğin dünyasında onurlu yerini almasının olmazsa olmaz koşulu budur" diyen Teberik,	şunları kaydetti:
12	"Sadece üretmenin sorunu çözmediğini" ifade eden Prof. Çakır,	şunları kaydetti:

The concordance line	The material preceding the reporting verb	The reporting verb
13	Şadan Eren, yaptığı yazılı açıklamada, alınan kararları dikkatle izlediklerini ve başarılı olması için destek verdiklerini belirterek,	şunları kaydetti:
14	Ege Serbest Bölgesi'nin yüksek teknoloji ve temiz çevre anlayışıyla üretim yapan modern bir endüstri merkezi olmasını amaçladıklarını ifade eden Tuncer,	şunları söyledi:
15	Rize Çay Üreticileri Birliği Başkanı Nurettin Kepenek de üreticilerin alın terinin karşılığının yok edildiğini iddia ederek,	şunları söyledi:
16	Prof. Dr. Karamış, Taksan için Mart ayında yeniden ihale açılacağı yolunda duyumlar aldıklarını bildirerek,	şunları söyledi:
17	Antepfıstığında en büyük sıkıntının dış pazarlarda rekabet edememek olduğuna dikkati çeken Güneydoğubirlik Ticaret Müdürü Mustafa Balaban ise	şunları kaydetti:

Fourth type of verb form: Examples of fourth type of reporting verbs are "şunları söyledi", "şunları kaydetti", "şöyle dedi", "şöyle konuştu", "sözlerini şöyle sürdürdü", and "sözlerini şöyle tamamladı".

Appendix I: Collocations of Reporting Verb “söyledi”

The table shows the collocations of reporting verb “söyledi” together with their frequency distribution, spread (U-score), and strength (K-score) statistics. Only the collocations that have K-score higher than unity are given in the table below.

Reporting verb	Collocate	K-score	U-score	Frequency
söyledi	şunları	118	177.862,7	2.765
söyledi	olduğunu	108	108.434,5	2.499
söyledi	nomber (any numeric number)	103	11.799,5	2.239
söyledi	ve	85	9.822,5	1.831
söyledi	bir	79	12.676,9	1.787
söyledi	gerektiğini	58	36.056,5	1.349
söyledi	bu	47	2.403,5	1.034
söyledi	da	42	1.859,7	952
söyledi	de	35	1.270,3	791
söyledi	için	32	1.532,9	713
söyledi	ile	23	564,4	518
söyledi	olarak	22	531,9	484
söyledi	büyük	20	1.011,5	461
söyledi	daha	20	645,4	457
söyledi	çok	20	757,3	453
söyledi	a,a	21	2.375,4	402
söyledi	önemli	17	462,4	389
söyledi	olmadığını	15	2.020,9	356
söyledi	bin	15	634,3	351
söyledi	bulunduğunu	15	1.992,9	349
söyledi	türkiye'nin	15	678,8	348
söyledi	muhabirine	15	1.890,8	344
söyledi	belirterek	14	1.738,4	339
söyledi	olacağını	14	2.239,8	339
söyledi	ifade	14	902,1	329
söyledi	yüzde	14	327,6	311
söyledi	devam	12	601,4	285
söyledi	yıl	12	241,6	266
söyledi	bakan	12	879,2	265
söyledi	belirten	11	1.307,2	265
söyledi	eden	11	1.311,4	263
söyledi	ettiğini	11	1.339,7	260

Reporting verb	Collocate	K-score	U-score	Frequency
söyledi	milyon	11	307,8	253
söyledi	en	11	262,3	249
söyledi	ekonomik	10	117,0	228
söyledi	türkiye	10	211,5	224
söyledi	bunun	9	190,6	223
söyledi	olduklarını	9	904,9	223
söyledi	lira	9	333,9	221
söyledi	geldiğini	9	927,0	221
söyledi	kadar	9	118,5	219
söyledi	milyar	9	235,2	216
söyledi	konusunda	9	147,6	213
söyledi	içinde	9	162,5	206
söyledi	yeni	9	101,5	206
söyledi	türk	9	236,7	188
söyledi	yaptığı	9	395,8	187
söyledi	türkiye'de	8	236,4	187
söyledi	ticaret	8	198,7	183
söyledi	gökalp	8	326,2	182
söyledi	ise	7	103,9	181
söyledi	iyi	7	99,9	181
söyledi	edildiğini	7	556,0	171
söyledi	ederek	7	413,5	169
söyledi	ettiklerini	7	559,3	162
söyledi	ilgili	7	54,8	161
söyledi	ankara	7	504,0	159
söyledi	"bu	7	625,1	159
söyledi	istanbul	7	598,1	158
söyledi	tarım	7	97,2	156
söyledi	sanayi	7	87,1	156
söyledi	istediklerini	6	522,8	156
söyledi	kaydeden	6	397,2	150
söyledi	sonra	6	58,0	142
söyledi	konuda	6	57,5	139
söyledi	son	6	47,3	138
söyledi	şöyle	6	298,8	134
söyledi	tarafından	6	69,5	129
söyledi	genel	6	73,3	125
söyledi	dünya	6	89,9	124
söyledi	başkanı	6	146,9	122
söyledi	önal	6	161,0	122
söyledi	bakanı	6	175,5	120
söyledi	devlet	6	128,4	116
söyledi	geçen	6	74,5	115
söyledi	üzerine	5	75,4	115
söyledi	anlatan	5	338,1	113
söyledi	dolar	5	78,1	113

Reporting verb	Collocate	K-score	U-score	Frequency
söyledi	şekilde	5	48,4	113
söyledi	açıklamada	5	282,1	112
söyledi	olan	5	54,8	112
söyledi	ancak	5	77,8	112
söyledi	yılında	5	65,8	100
söyledi	önce	4	33,2	99
söyledi	iki	4	34,3	88
söyledi	yüksek	4	19,0	84
söyledi	ülke	3	26,5	83
söyledi	yapıldığını	3	206,9	83
söyledi	arasında	3	26,3	77
söyledi	ortaya	3	99,0	75
söyledi	hedeflediklerini	3	212,0	68
söyledi	sahip	3	63,2	67
söyledi	kredi	3	21,9	65
söyledi	nedeniyle	3	20,4	63
söyledi	demirel	3	61,6	62
söyledi	yönelik	3	24,3	59
söyledi	izmir	3	63,3	55
söyledi	enerji	3	13,89	55
söyledi	yeniden	2	27,5	55
söyledi	gibi	2	20,1	55
söyledi	olumlu	2	47,60	50
söyledi	toskay	2	66,56	48
söyledi	hükümetin	2	34,61	47
söyledi	iş	2	13,21	47
söyledi	petrol	2	11,04	46
söyledi	vergi	2	14,44	46
söyledi	dış	2	19,24	46
söyledi	ilk	2	17,24	46
söyledi	cumhurbaşkanı	2	41,85	45
söyledi	yalova	2	35,84	44
söyledi	beklediklerini	2	174,24	44
söyledi	göre	2	10,89	39
söyledi	bugün	2	16,36	38
söyledi	yaptıklarını	1	93,3	37
söyledi	önem	1	55,86	36
söyledi	yatırım	1	18,37	35