YILDIRIM BEYAZIT UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

OTTOMAN-TURKISH OPTICAL CHARACTER RECOGNITION AND LATIN TRANSCRIPTION

M.Sc. Thesis by

MUSTAFA DOĞRU

Department of Computer Engineering

January, 2016

ANKARA

OTTOMAN-TURKISH OPTICAL CHARACTER RECOGNITION AND LATIN TRANSCRIPTION

A Thesis Submitted to

the Graduate School of Natural and Applied Sciences of Yıldırım Beyazıt University

In Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Engineering, Department of Computer Engineering

by

Mustafa DOĞRU

January, 2016

ANKARA

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **“OTTOMAN-TURKISH OPTICAL CHARACTER RECOGNITION AND LATIN TRANSCRIPTION”** completed by Mustafa DOĞRU under supervision of **Assoc. Prof.Dr. Fatih KOYUNCU** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof.Dr. Fatih KOYUNCU

(Supervisor)

Asst. Prof.Dr. Mehmet DEMIRER	Asst. Prof.Dr. Baha ŞEN
(Jury Member)	**(Jury Member)**

Prof.Dr. Fatih V. ÇELEBİ

Director

Graduate School of Natural and Applied Sciences

ETHICAL DECLERATION

I have prepared this dissertation study in accordance with the Rules of Writing Thesis of Yıldırım Beyazıt University of Science and Technology Institute;

- Data I have presented in the thesis, information and documents that I obtained in the framework of academic and ethical rules,
- All information, documentation, assessment and results that I presented in accordance with scientific ethics and morals,
- I have gave references all the works that I were benefited in this dissertation by appropriate reference,
- I would not make any changes in the data that I were used,
- The work presented in this dissertation I would agree that the original,

I state, in the contrary case I declare that I accept the all rights losses that may arise against me.

OTTOMAN-TURKISH OPTICAL CHARACTER RECOGNITION AND LATIN TRANSCRIPTION

ABSTRACT

There are numerous documents in Ottoman-Turkish on the archives or online resources. Unfortunately these documents could not be understood by the people who cannot read Ottoman-Turkish alphabet. Ottoman-Turkish optical character recognition and Latin transcription could be the solution of this problem. In this thesis, Tesseract optical character recognition engine is used in order to recognize Ottoman-Turkish characters. Also, various methods are developed for the transcription from Ottoman Turkish to Latin. Characters on some Ottoman-Turkish images could not be recognized by optical character recognition methods. So, Ottoman-Turkish keyboard was developed for writing unrecognized characters with Ottoman-Turkish alphabet. Dictionary tables are used for transcription process. So enrichment data in the dictionary tables will increase of transcription success. Thus, an application was developed for enrichment data in the dictionary tables.

Keywords: Optical Character Recognition (OCR), Tesseract, Transcription, Ottoman Turkish, Latin, Regular Expression, Database

OSMANLI TÜRKÇESİ OPTİK KARAKTER TANIMA VE LATİNCE TRANSKRİPSİYONU

ÖZET

Arşivlerde veya çevrim içi kaynaklarda sayısız Osmanlıca belgeler vardır. Bu belgeler maalesef Osmanlıca okuyamayan kişiler tarafından anlaşılamamaktadır. Osmanlı Türkçesi optik karakter tanıma ve Latince transkripsiyonu bu problemin çözümü olabilir. Bu tezde Tesseract optik karakter tanıma motoru Osmanlıca karakterleri tanımak için kullanılmıştır. Ayrıca Osmanlı Türkçesinden Latinceye transkripsiyon için çeşitli metotlar geliştirilmiştir. Bazı Osmanlıca resimlerdeki karakterler optik karakter tanıma metotları ile tanınamamaktadır. Tanınamayan bu karakterleri Osmanlıca alfabesi ile yazmak için Osmanlıca klavye geliştirilmiştir. Transkripsiyon işlemi için sözlük tabloları kullanılmaktadır. Sözlük tablolarındaki veriyi zenginleştirmek transkripsiyon başarısını artıracağından dolayı sözlük tablolarını geliştirmek için bir uygulama geliştirilmiştir.

Anahtar sözcükler: Optik Karakter Tanıma, Tesseract, Transkripsiyon, Osmanlıca Türkçesi, Latince, Kurallı İfadeler, Veritabanı

ACKNOWLEDGMENTS

First of all I want to thank my supervisor Assoc. Prof.Dr. Fatih KOYUNCU for supporting me to achieve my long way. It has been an honor and a pleasure for me to work with him.

Finally, I would like to thank my wife and my lovely daughter Hatice Erva for their love, patience, and understanding,

Without you, I wouldn’t be here.

2016, 19 January **Mustafa DOĞRU**

CONTENTS

Page

M.Sc THESIS EXAMINATION RESULT FORM ii

ETHICAL DECLERATION iii

ABSTRACT iv

ÖZET v

ACKNOWLEDGMENTS vi

CONTENTS vii

ABBREVATION x

LIST OF TABLES xi

LIST OF FIGURES xii

CHAPTER ONE - INTRODUCTION 1

1.1. Related Works; **2**

1.2. Aim of the Work **3**

CHAPTER TWO - BASIC INFORMATION ABOUT THE OTTOMAN-TURKISH LANGUAGE 5

CHAPTER THREE - OPTICAL CHARACTER RECOGNATION 10

3.1. The Need of Ottoman-Turkish OCR **10**

3.2. OCR Engines **11**

3.3. Ottoman-Turkish Optical Character Recognition with Tesseract **12**

3.4. Document Image Enhancement Techniques before the Optical Character Recognition **13**

3.4.1. Skew Correction **13**

3.4.2. Simple Image Processing Operations **14**

3.4.3. Document Image Restoration and Enhancement **15**

3.5. Ottoman - Turkish Character Recognition on the First Two Pages of the Nutuk **18**

CHAPTER FOUR - TRANSCRIPTION 22

4.1. Transcription Phases **22**

4.1.1. Preprocessing Operation **22**

4.1.2. One To One Transcription Method **23**

4.1.2.1. Transcription Operation on the First Quatrain of Turkish National Anthem by One to One Transcription Method **24**

4.1.3. Transcription using Regular Expressions Method **25**

4.1.3.1. Creating Convenient Regular Expressions **26**

4.1.3.2. Look at Modern Turkish-Latin Table **32**

4.1.3.3. *Transcription Operation on the Third Quatrain of Turkish National Anthem by Transcription Using Regular Expressions Method* **33**

4.2. Comparison of One To One Transcription and Transcription Using Regular Expressions **34**

CHAPTER FIVE - OTTOMAN-TURKISH KEYBOARD 36

5.1. The Transcription of an Inscription Using the Ottoman-Turkish Keyboard.. **37**

5.2. The Transcription of an Ottoman Tombstone Using the Ottoman-Turkish Keyboard **39**

CHAPTER SIX - DATABASE and SOFTWARE 43

6.1. Ottoman Optical Character Recognition and Transcription Software **43**

6.2. Ottoman Dictionary Database **45**

CHAPTER SEVEN CONCLUSION 46

REFERENCES 47

APPENDIX A - OPTICAL CHARACTER RECOGNITION ON THE FIRST PAGE OF THE NUTUK 51

APPENDIX B - OPTICAL CHARACTER RECOGNITION ON THE SECOND PAGE OF THE NUTUK 55

APPENDIX C - MAIN GRAPHICAL USER INTERFACE 61

APPENDIX D - OTTOMAN-TURKİSH KEYBOARD TOOL 62

APPENDIX E - TRANSCRIPTION TOOL 63

APPENDIX F - ENRICHMENT DATABASE TOOL 64

RESUME 65

ABBREVATION

OCR	Optical Character Recognition
HMM	Hidden Markov Model
CBR	Content-Based Retrieval
ANN	Artificial Neural Network
RE	Regular Expression
GUI	Graphical User Interface

LIST OF TABLES

Table 2.1 Ottoman-Turkish alphabet.. 6
Table 2.2 Ottoman-Turkish numerals... 8
Table 3.1 Character accuracy for Arabic packages [17]. 11
Table 4.1 Suffixes... 23
Table 4.2 Regular expression generating rules 1.. 27
Table 4.3 Regular expression generating rules 2.. 29
Table 4.4 Regular expression generating rules 3.. 30
Table 5.1 Transcription of the tombstone... 41

LIST OF FIGURES

Figure 1.1 Overall Block Diagram of Ottoman - Turkish optical character recognition and Latin transcription 2

Figure 3.1 Sample an Ottoman-Turkish image 10

Figure 3.2 Character accuracy for Arabic packages [17] 11

Figure 3.3 a) Tesseract architecture [23], b) Block level diagram [24] 12

Figure 3.4 Skew correction on Ottoman-Turkish page 14

Figure 3.5 a) Sample original Ottoman-Turkish image, b) Grayscale image, c) Monochrome image, d)Inverted image, e) Sharpening filter on an image f) Smoothing filter on an image 15

Figure 3.6 a) Sample original Ottoman-Turkish image, b) Grayscale image, c) Normalization using Imagemagick library, d) Adaptive blur with value 50 using Imagemagick library 16

Figure 3.7 a) Sample original Ottoman-Turkish image, b) Monochrome + adaptive-blur with value 50 using Imagemagick library 17

Figure 3.8 First page of the Nutuk 18

Figure 3.9 Second page of the Nutuk 20

Figure 4.1 Diagram of one to one transcription method 24

Figure 4.2 Turkish national anthem's first quatrain with Ottoman-Turkish and transcription 25

Figure 4.3 Diagram of transcription using regular expressions method 26

Figure 4.4 Detailed generating regular expression using rules tables for each letter 31

Figure 4.5 The query for موستافه word 33

Figure 4.6 Transcription using regular expressions on the third quatrain of Turkish national anthem 33

Figure 5.1 Ottoman-Turkish Keyboard 36

Figure 5.2 An Ottoman Inscription 37

Figure 5.3 The writing of the inscription with Ottoman-Turkish Alphabet 37

Figure 5.4 Transcription of the inscription 38

Figure 5.5 Ottoman tombstone 39

Figure 5.6 The writing of the Ottoman tombstone with Ottoman-Turkish alphabet 40

Figure 6.1 Implemented java classes 44

Figure 6.2 Database tables 45

Figure C.1 Main graphical user interface (GUI). 61

Figure D.1 Ottoman-Turkish keyboard 62

Figure E.1 Transcription tool 63
Figure F.1 Enrichment database tool 64

CHAPTER ONE

INTRODUCTION

There are numerous documents in the archives, which had been written with Ottoman-Turkish Alphabet which is a version of Persian-Arabic Alphabet. These documents include valuable information about various fields. But unfortunately these documents could not be understood by those who could not read Ottoman-Turkish Alphabet. Transcription from Ottoman Turkish alphabet to modern Turkish alphabet is very helpful in case of to understand Ottoman-Turkish documents.

Figure 1.1 shows overall block diagram of Ottoman-Turkish optical character recognition and Latin transcription. According to this Figure, firstly, Ottoman-Turkish Text needs to extract from an Ottoman-Turkish image. Ottoman-Turkish optical character recognition or retyping text by Ottoman Keyboard can be use in order to extract the text from an Ottoman-Turkish image. After then, Latin text converts by using some transcription operations in order to recognized words in the text.

Based on what is covered above, this thesis is comprised of the following chapters.

Chapter 1 is an introduction about this thesis. Chapter 2 is briefly information of Ottoman-Turkish Language; Chapter 3 is describing in detail optical character recognition and practical examples on some pages. Chapter 4 presents the implementation of transcription from Ottoman-Turkish to Latin-based Turkish and practical examples on some Ottoman-Turkish texts. Chapter 5 is about Ottoman-Turkish Keyboard in order to write with Ottoman-Turkish alphabet. Chapter 6 gives the information about our database. Chapter 7 concludes the thesis and suggests future work.

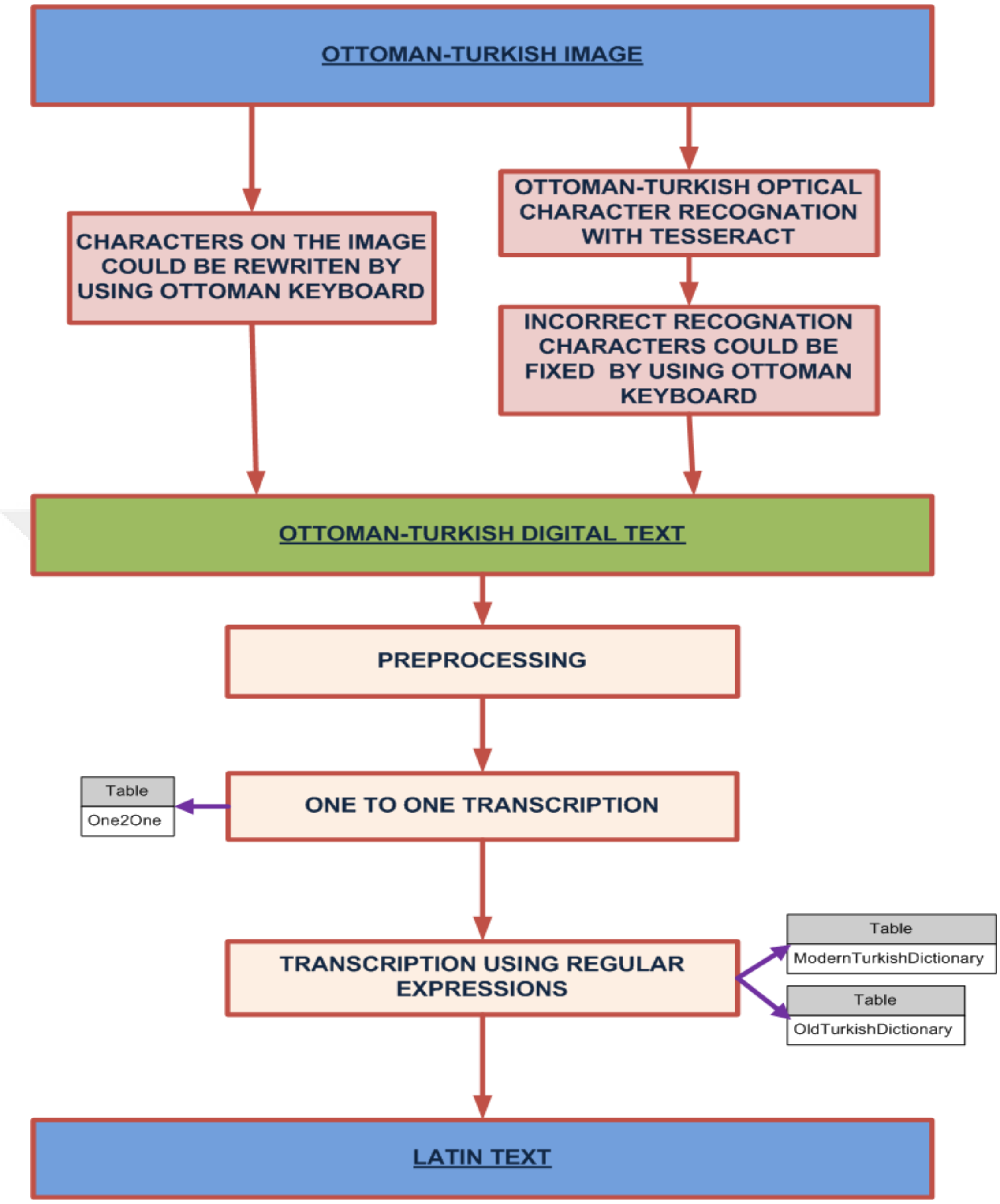

OTTOMAN-TURKISH IMAGE
CHARACTERS ON THE IMAGE COULD BE REWRITEN BY USING OTTOMAN KEYBOARD
OTTOMAN-TURKISH OPTICAL CHARACTER RECOGNATION WITH TESSERACT
INCORRECT RECOGNATION CHARACTERS COULD BE FIXED BY USING OTTOMAN KEYBOARD
OTTOMAN-TURKISH DIGITAL TEXT
PREPROCESSING
Table
One2One
ONE TO ONE TRANSCRIPTION
Table
ModernTurkishDictionary
TRANSCRIPTION USING REGULAR EXPRESSIONS
Table
OldTurkishDictionary
LATIN TEXT

Figure 1.1 Overall block diagram of Ottoman - Turkish optical character recognition and Latin transcription

1.1. Related Works;

There are many works on optical character recognition. Most of these works are generally focused character recognition for Latin or Arabic. Studies of optical character recognition for Urdu, Persian or Arabic languages could also benefit in order to recognize characters on Ottoman-Turkish images, because Ottoman-Turkish letters are same or similar in the Urdu, Arabic and Persian Letters.

Works of character recognition on the Ottoman-Turkish images have recently increased in the literature. Öztürk et al. (2000) [1] are development a program using supervised feedforward neural network with backpropagation for the character recognition of Ottoman-Turkish. Kurt et al. (2009) [2] presents a Linear Discriminant Analysis based automatic Ottoman Alphabet Character Recognition System. Onat et al. (2006) [3] developed a system by using Hidden Markov Model (HMM) techniques for Ottoman script recognition. Another work is that Content-Based Retrieval (CBR) System for Ottoman Archives [4]. In this study, "the symbols extracted from the documents are matched with the most similar one in the symbol library, which is created in a supervised manner" (p.1). Başar et al. (2007) [5] have recognized Ottoman characters using Artificial Neural Network (ANN). They pointed out "The system has achieved 85.5% classification accuracy" (p.4). Ataer and Duygulu (2007) [6] proposed a method for retrieval of Ottoman documents without requiring character recognition. Yalnız (2008) [7] propose context-sensitive segmentation and recognition method for connected letters in Ottoman script is proposed. Also Yalnız et al. (2009) [8] investigated several methods for character segmentation and recognition stages for printed and handwritten historical documents.

There are few works on transcription from Ottoman-Turkish language to the Latin. Andrews et al. (2010) [9] developed a system for Latin based transcription from Ottoman-Turkish texts.

1.2. Aim of the Work

The main goal of this thesis is characters recognition Ottoman-Turkish images and then the Latin-based Turkish transcription from ottoman documents. In this way, Ottoman-Turkish documents could be automatically converted into the Latin-based documents. In this manner, people can easily understand these converted documents with the Latin alphabet. Firstly, optical character recognition (OCR) with Tesseract Engine could be used to extract characters from an image that is written Ottoman-

Turkish Alphabet. Also Ottoman-Turkish text could be rewritten by Ottoman-Turkish keyboard. Ottoman-Turkish text could be transcribed to Latin-based Turkish by different approaches after handling Ottoman-Turkish text.

CHAPTER TWO

BASIC INFORMATION ABOUT THE OTTOMAN-TURKISH LANGUAGE

“The Ottoman-Turkish Language was used in Ottoman Empire. The Ottoman-Turkish vocabulary generally consists of one of native Turkish, Arabic and Persian words. Persian and Arabic highly influenced Ottoman-Turkish Language. In the 17th and 18th centuries, Persian and Arabic vocabulary amounted for up to 88% of its vocabulary”. [10]

“The letters of the Ottoman-Turkish Alphabet are 32 in number, and consist of 28 Arabic letters together with some which the Persians have added (ژ چ گ پ). Ottoman-Turkish Language reads and writes from right to left. Capital letters are unknown”. [11]

Generally, vowels letters are “ا, و, ي, ه” but “و, ي, ه” letters could also be consonant letters. It is depended the position of letter in the word. The letters of the Ottoman-Turkish alphabet divided into two parts. These are connected and unconnected letters. The connected letters are also known as cursive letter. These letters’ shapes are different form according to the position of word. The unconnected letters are also known as isolated letters. The unconnected letters never joined to other letters. Table 2.1 and Table 2.2 show the letters and numbers.

Orthographic signs could be used in the some words. These signs are changed the transcription and the pronunciation of the words. These are also known "hareke". These signs are Ustun, Esre, Otre, Jezma, Shedda, Medda and Nunation. Ustun is marked with the sign (◌َ) and put over the letters. The pronounced of Ustun is "e" or "a". Esre is marked with the sign (◌ِ) and put under the letters. The pronounced of Esre is "i" or "ı". Otre is marked with (◌ُ) and put over the letters. The pronounced of Otre is "u" or "ü". Jezma sign is (◌ْ) and put over the letters. The pronounced of Jezma is connected two consonant with a vowel letter. Shedda sign is (◌ّ) and put over the letters. The letter with Shedda is to be doubled without the interposition of a vowel. Medda sign is (◌ٓ) which means long. Nunation sign are (◌ٌ, ◌ً, ◌ٍ) and put over or under the letters. The pronounced of Nunation is "en, "in" or "ın".

Table 2.1 Ottoman-Turkish alphabet.

Isolated	Final	Medial	Initial	Modern Turkish Name	Name
ا	ـا	—		Elif	Alif
ب	ـب	ـبـ	بـ	Be	Baa
پ	ـپ	ـپـ	پـ	Pe	Pe
ت	ـت	ـتـ	تـ	Te	Taa
ث	ـث	ـثـ	ثـ	Se	Thaa
ج	ـج	ـجـ	جـ	Cim	Jeem
چ	ـچ	ـچـ	چـ	Çim	Chim
ح	ـح	ـحـ	حـ	Ha	Haa
خ	ـخ	ـخـ	خـ	Hı	KHaa
د	ـد	—		Dal	Daal
ذ	ـذ	—		Zel	Dhal

Table 2.1 (Continued) Ottoman-Turkish alphabet.

Isolated	Final	Medial	Initial	Modern Turkish Name	Name
ر	ـر	—		Rı	Raa
ز	ـز	—		Ze	Zaa
ژ	ـژ	—		Je	Zhe
س	ـس	ـسـ	سـ	Sin	Seen
ش	ـش	ـشـ	شـ	Şın	Sheen
ص	ـص	ـصـ	صـ	Sad	Saad
ض	ـض	ـضـ	ضـ	Dad	Dhad
ط	ـط	ـطـ	طـ	Tı	Tta
ظ	ـظ	ـظـ	ظـ	Zı	Dha
ع	ـع	ـعـ	عـ	Ayın	Ain
غ	ـغ	ـغـ	غـ	Gayın	Ghain
ف	ـف	ـفـ	فـ	Fe	Faa
ق	ـق	ـقـ	قـ	Kaf	Qaf
ك	ـك	ـكـ	كـ	Kef	Kaaf

Table 2.1 (Continued) Ottoman-Turkish alphabet.

Isolated	Final	Medial	Initial	Modern Turkish Name	Name
گ	ـگ	ـگـ	گـ	Gef	Geef
ل	ـل	ـلـ	لـ	Lam	Laam
م	ـم	ـمـ	مـ	Mim	Meem
ن	ـن	ـنـ	نـ	Nun	Noon
و	ـو	—		Vav	Waaw
ه	ـه	ـهـ	هـ	He	Haa
ی	ـی	ـیـ	یـ	Ye	Yaa

Table 2.2 Ottoman-Turkish numerals.

Numerals	Numbers	Ottoman-Turkish Name	Modern Turkish Name	Name
٠	0	صفر	Sıfır	Zero
١	1	بر	Bir	One
٢	2	ایکی	İki	Two
٣	3	اوچ	Üç	Three
٤	4	درت	Dört	Four

Table 2.2 (Continued) Ottoman-Turkish numerals.

Numerals	Numbers	Ottoman-Turkish Name	Modern Turkish Name	Name
٥	5	بش	Beş	Five
٦	6	آلتى	Altı	Six
٧	7	يدى	Yedi	Seven
٨	8	سكز	Sekiz	Eight
٩	9	طقوز	Dokuz	Nine
١٠	10	اون	On	Ten

The grammar of Ottoman-Turkish Language is similar to the grammar of modern Turkish.

For instance;

Ottoman – Turkish:

اكلير	**ياش ايكن**	**اغاج**
Verb	Adverb	Subject

Turkish in Turkey:

Ağaç	**yaş iken**	**eğilir**
Subject	Adverb	Verb

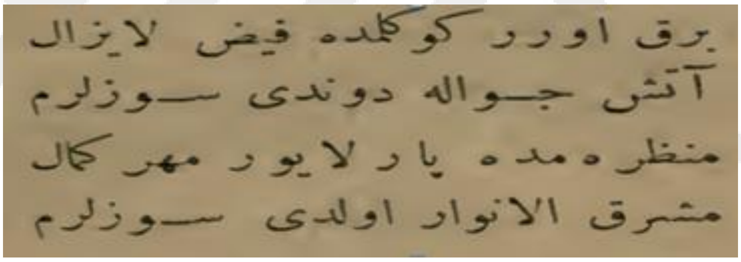
CHAPTER THREE

OPTICAL CHARACTER RECOGNATION

Optical character recognition (OCR) is the process of converting scanned images into machine readable character streams, plain (e.g. text files) or formatted (e.g. HTML files). OCR is a successful branch of Pattern Recognition. [12]

Type of OCR could be Offline or Online, Printed or Handwritten, Isolated or Cursive, Single font or Omni-font. The General OCR Process is image acquisition, pre-processing, segmentation, feature extraction and classification & recognition. [13].

For instance,



برق اورر کوکلمده فیض لایزال
آتش جواله دوندی سوزلرم
منظره مده یار لایور مهر کمال
مشرق الانوار اولدی سوزلرم

Figure 3.1 Sample an Ottoman-Turkish image

For this image, the correct result must be as below, after OCR process runs.

ب ر ق   ا و ر ر ک و ک ل م د ه   ف ی ض   لا ی ز ا

آ ت ش   ج و ا ل ه   د و ن د ی   س و ز ل ر م

م ن ظ ر ه   م د ه   پ ا ر لا ی و ر   م ه ر   ک م ا ل

م ش ر ق   ا لا ن و ا ر   ا و ل د ی   س و ز ل ر م

3.1. The Need of Ottoman-Turkish OCR

Ottoman-Turkish was widely used in the Ottoman territories over three continents during Ottoman Empire. Archives and online resources include documents with ottoman-Turkish Alphabet. Some of online resources are Turkish Grand National Assembly Library and Documentation Center webpage [14], Turkey Ministry of Culture and Tourism webpage (15) and Farabi Digital Library [16]. These Ottoman-Turkish documents contain invaluable information about various fields.

3.2. OCR Engines

Many OCR Engines could be made use of optical character recognition. For Arabic language, some commercial OCR products are ABBYY FineReader, OmniPage Professional, Readiris Pro and Sakhr software. Modawi (2005) [17] shows the comparison of some Arabic OCR software products. The following figure and table show the performances of OCR for commercial Arabic software products.

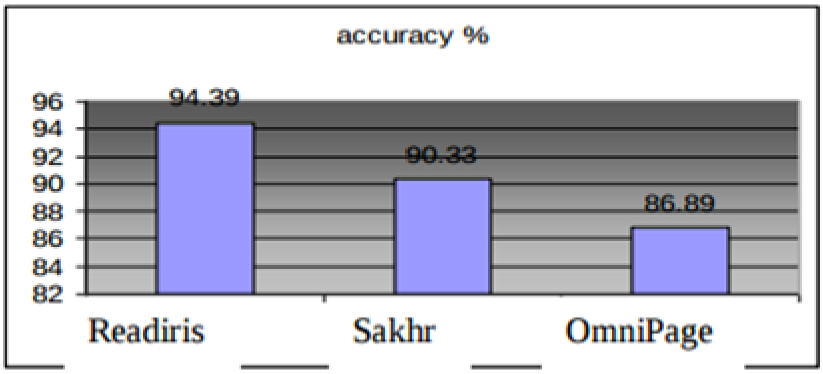
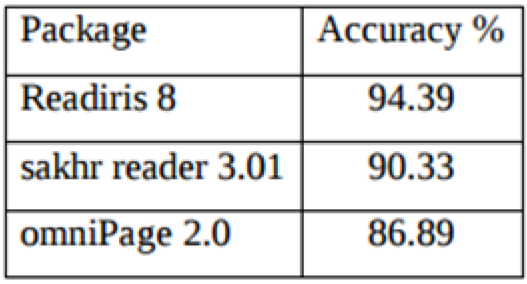


Figure 3.2 Character accuracy for Arabic packages [17]

Table 3.1 Character accuracy for Arabic packages [17].



Package	Accuracy %
Readiris 8	94.39
sakhr reader 3.01	90.33
omniPage 2.0	86.89

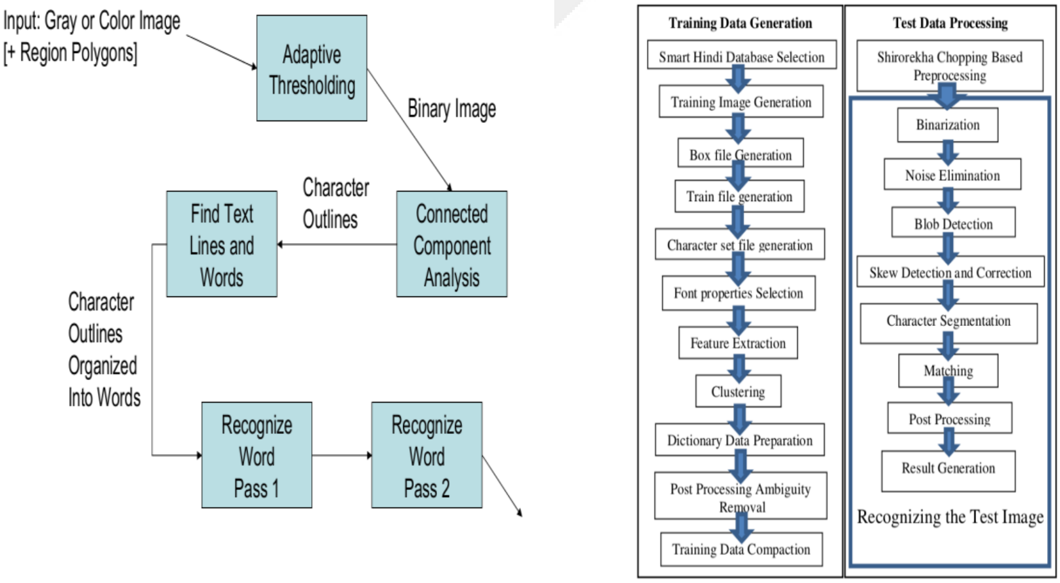
Tesseract, GOCR and OCRopus are an open source OCR projects. Tesseract only supports Arabic language of these and also the uses most widely open source OCR tool. So in this study, Tesseract software is benefited for recognition of Ottoman-

Turkish characters. Also Heliński et al. (2012) show comparison Tesseract and ABBYY FineReader tools [18] also give information about performance values.

3.3. Ottoman-Turkish Optical Character Recognition with Tesseract

Ottoman-Turkish optical character recognition (OCR) is the recognition characters with Ottoman-Turkish alphabet of input Ottoman-Turkish images. Tesseract has been improved an open source project since 2006 and written in C and C++.

Tesseract engine has a lot of methods and algorithms. Some of them are layout analysis, line and word finding, word recognition, static character classifier, adaptive classifier. Tesseract architecture and its methods and algorithms are discussed in researches [19-24]. Tesseract architecture and block level diagram could also see in the Figure 3.3.


Input: Gray or Color Image
[+ Region Polygons]
Adaptive Thresholding
Binary Image
Connected Component Analysis
Character Outlines
Find Text Lines and Words
Character Outlines Organized Into Words
Recognize Word Pass 1
Recognize Word Pass 2
Training Data Generation
Smart Hindi Database Selection
Training Image Generation
Box file Generation
Train file generation
Character set file generation
Font properties Selection
Feature Extraction
Clustering
Dictionary Data Preparation
Post Processing Ambiguity Removal
Training Data Compaction
Test Data Processing
Shirorekha Chopping Based Preprocessing
Binarization
Noise Elimination
Blob Detection
Skew Detection and Correction
Character Segmentation
Matching
Post Processing
Result Generation
Recognizing the Test Image

(a) (b)

Figure 3.3 a) Tesseract architecture [23], b) Block level diagram [24]

Tesseract is used Leptonica [25] library. This library is an image processing library for image processing operations. Tesseract is internally used this library before doing actual optical character recognition. Leptonica library does image processing operations like binarization, noise removing, skew and orientation detection etc.

Tesseract could recognize Arabic images after Tesseract version 3. Tesseract version 3 Engine with Arabic language dataset could be utilized for recognition Ottoman-Turkish characters. Arabic language dataset could be suitable for Ottoman documents, because Ottoman-Turkish letters consist of 28 Arabic letters. Ottoman and Arabic scripts have many common characteristics.

There are three main problems for optical character recognition (OCR) on Ottoman-Turkish images with Tesseract. First problem is that Arabic dataset is not suitable for 4 Persian letters. For instance, چ letter is generally recognized ج letter. Second problem is that the performance of optical character recognition (OCR) is directly dependent on quality of input documents. Last problem is that handwriting and Ottoman calligraphy affect the performance of optical character recognition (OCR).

3.4. Document Image Enhancement Techniques before the Optical Character Recognition

Some image enhancement techniques could be used before optical character recognition. Also these techniques could improve reading performance. Although Tesseract does most of this image processing operation internally, these operations could be benefited before optical character recognition operation. These operations are skew correction, image processing operations and historical document restoration and enhancement.

3.4.1. Skew Correction

Skew correction can be done before optical character recognition. For the skew correction for an image, skew angle is firstly determined using Hough Transform and then the image is rotate accordingly. Hough transform is a feature extraction technique used in image analysis, computer vision and digital image processing [26].

It is commonly used for solution to difference problems that are object, line and circle detection, skew correction etc. More detail about Hough Transform can be found in [27, 28]. In the Figure 3.4 shows an example of skew correction on Ottoman-Turkish page. For this page, skew angle is -4.77 degrees.

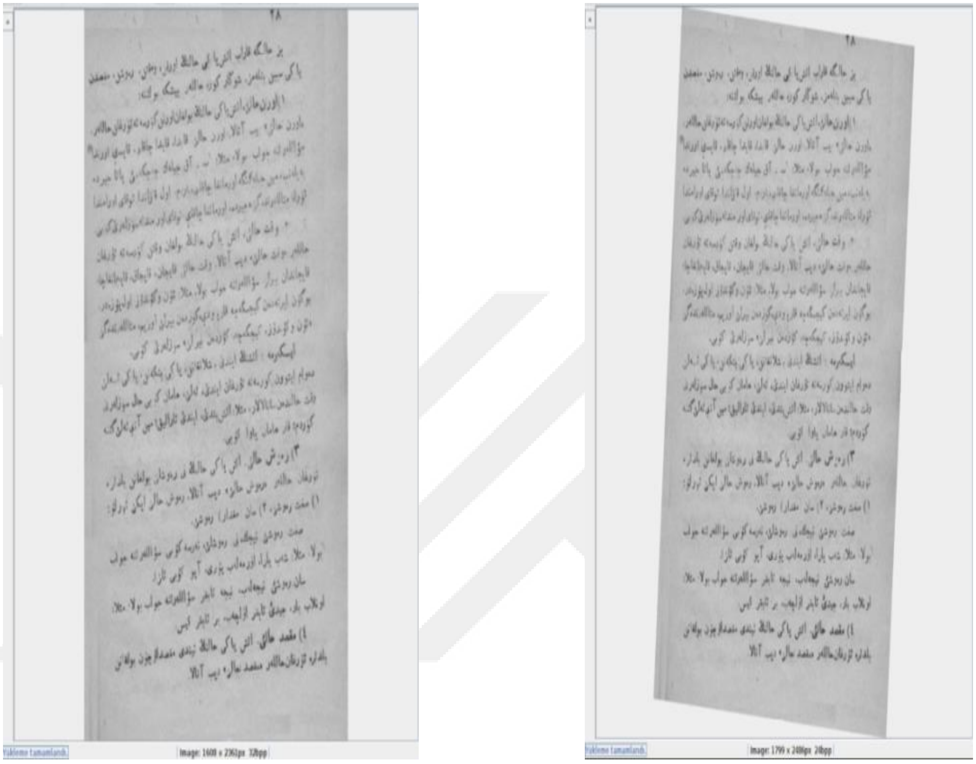
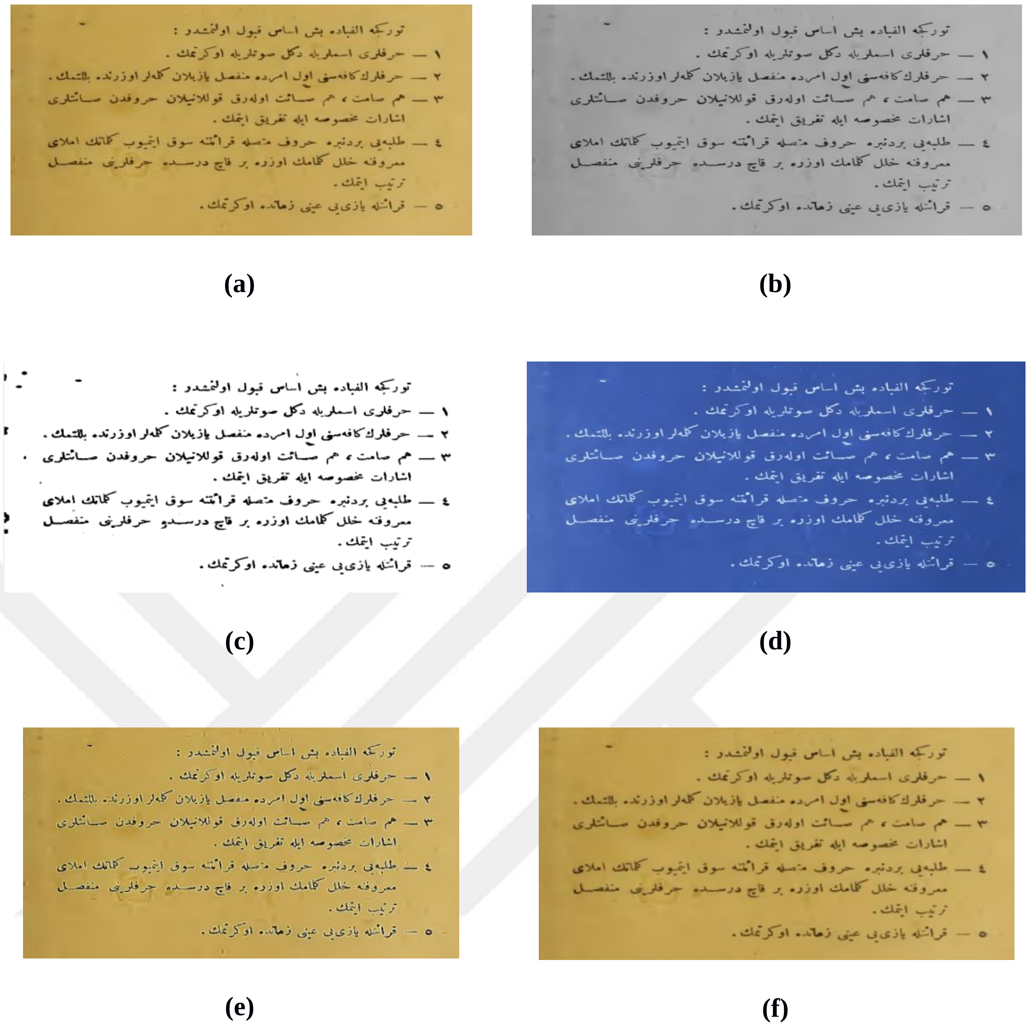


Figure 3.4 Skew correction on Ottoman-Turkish page

3.4.2. Simple Image Processing Operations

Image processing operations could be used before the recognition. These are brightness, contrast, grayscale, monochrome, invert, sharpen filter, smooth filter. The performance of optical character recognition could be small change but generally does not affect optical character results after using these operations. Also the results of the recognition could be negatively impacted by these operations. In Figure 3.5 shows the results of these operations on a sample Ottoman-Turkish image.



تورکجه الفباده بش اساس قبول اولنمشدر :

١ — حرفلری اسملریله دکل سوتلریله اوکرتمك .

٢ — حرفلرك کافه‌سنی اول امرده منفصل یازیلان کلمه‌لر اوزرنده بلتمك .

٣ — هم صامت ، هم صائت اوله‌رق قوللانیلان حروفدن صائتلری
اشارات مخصوصه ایله تفریق ایتمك .

٤ — طلبه‌یی بردنبره حروف متصله قرائته سوق ایتمیوب کلماتك املای
معروفنه خلل کلامك اوزره بر قاچ درسده حرفلرینی منفصل
ترتیب ایتمك .

٥ — قرائتله یازی‌یی عینی زمانده اوکرتمك .

(a) (b)

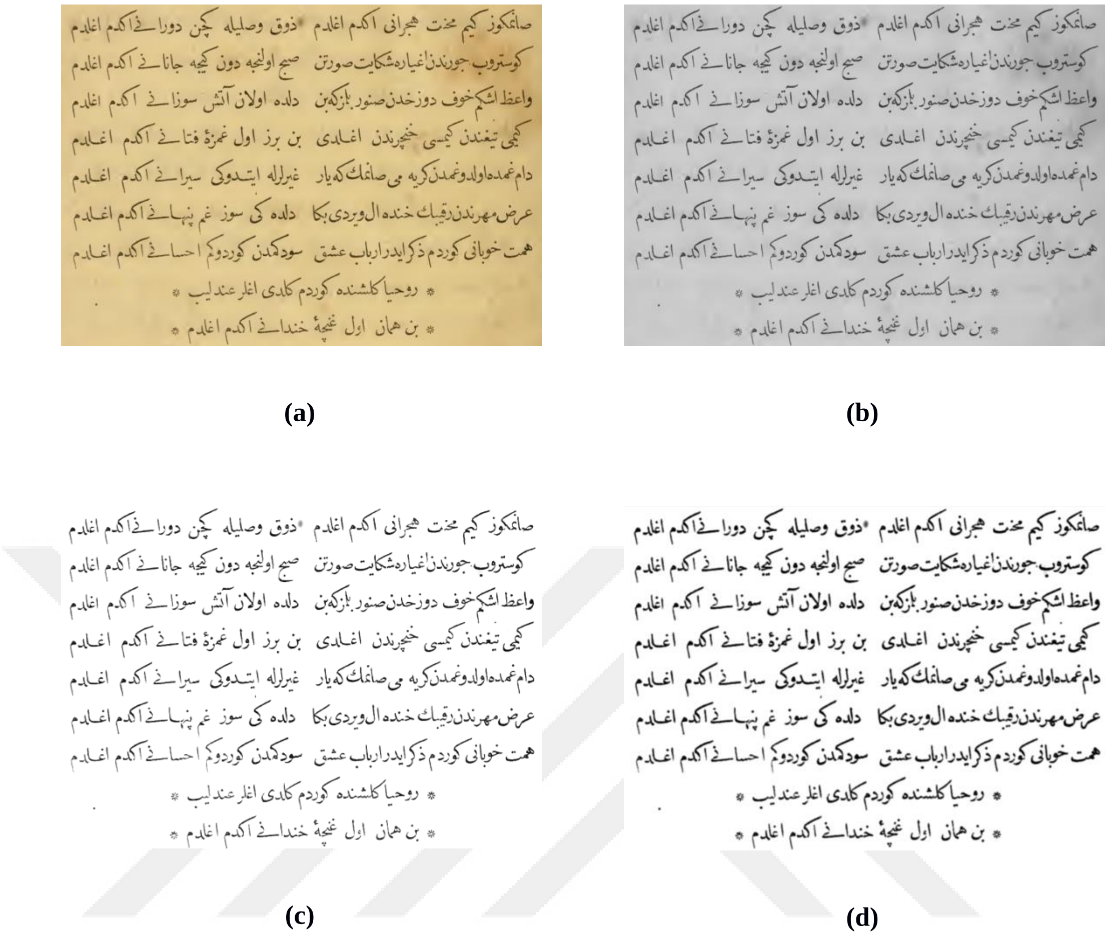
(c) (d)

(e) (f)

Figure 3.5 a) Sample original Ottoman-Turkish image, b) Grayscale image, c) Monochrome image, d)Inverted image, e) Sharpening filter on an image f) Smoothing filter on an image

3.4.3. Document Image Restoration and Enhancement

Many methods have implemented for image restoration and enhancement. Imagemagick library [29] methods are very useful library for document image restoration and enhancement. Some examples results are shown in Figure 3.6 and Figure 3.7 by using Imagemagick Library methods.

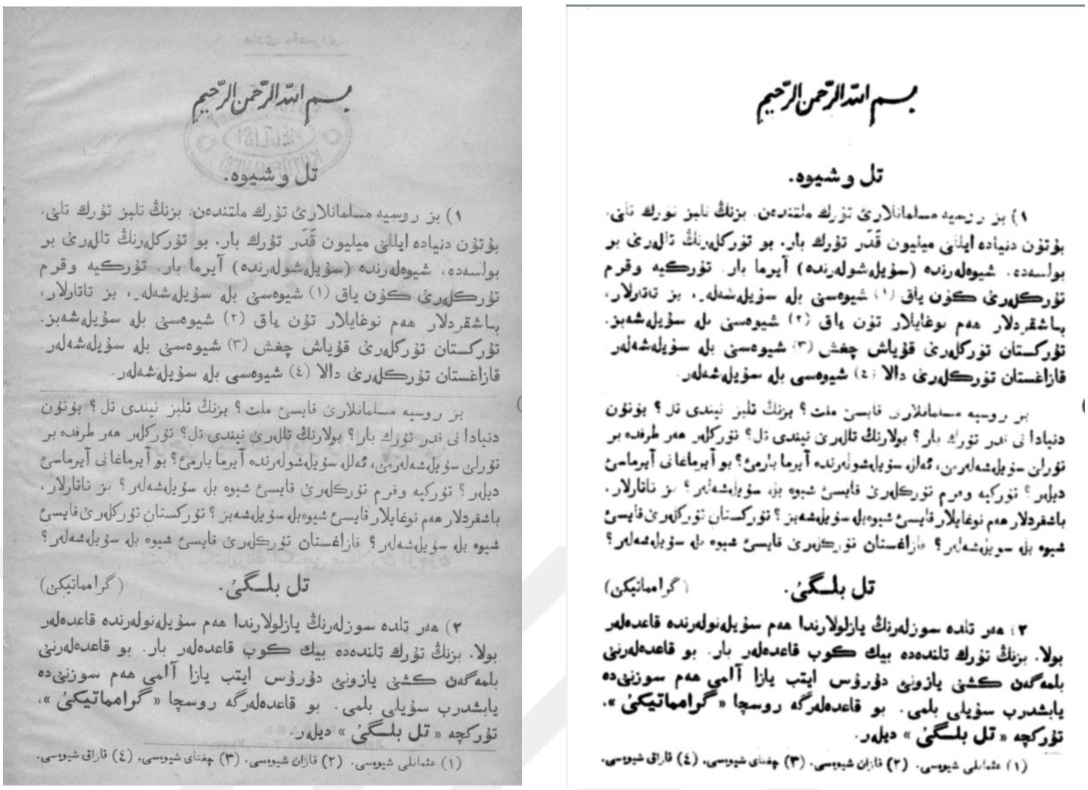

(a)

(b)

(c)

(d)

Figure 3.6 a) Sample original Ottoman-Turkish image, b) Grayscale image, c) Normalization using Imagemagick library, d) Adaptive blur with value 50 using Imagemagick library



(a)

بسم الله الرحمن الرحيم

تل و شیوه.

۱) بز روسیه مسلمانلاری تۆرك ملتندهن. بزنڭ تلبز تۆرك تلی.
بۆتۆن دنیاده ایللی میلیون قدر تۆرك بار. بو تۆركلهرنڭ تللهری بر
بولسهده، شیوهلهرنده (سۆیلهشولهرنده) آیرما بار. تۆركیه وقرم
تۆركلهری كۆن یاق (۱) شیوهسی بله سۆیلهشهلهر، بز تاتارلار،
باشقردلار ههم نوغایلار تۆن یاق (۲) شیوهسی بله سۆیلهشهبز.
تۆركستان تۆركلهری قۆیاش چغش (۳) شیوهسی بله سۆیلهشهلهر.
قازاغستان تۆركلهری دالا (٤) شیوهسی بله سۆیلهشهلهر.

بز روسیه مسلمانلاری قایسی ملت؟ بزنڭ تلبز نیندی تل؟ بۆتۆن
دنیادا نی قدر تۆرك بار؟ بولارنڭ تللهری نیندی تل؟ تۆركلهر ههر طرفده بر
تۆرلی سۆیلهشهلهرمی، ئهلله سۆیلهشولهرنده آیرما بارمی؟ بو آیرماغانی آیرماسی
دیلهر؟ تۆركیه وقرم تۆركلهری قایسی شیوه بله سۆیلهشهلهر؟ بز تاتارلار،
باشقردلار ههم نوغایلار قایسی شیوه بله سۆیلهشهبز؟ تۆركستان تۆركلهری قایسی
شیوه بله سۆیلهشهلهر؟ قازاغستان تۆركلهری قایسی شیوه بله سۆیلهشهلهر؟

تل بلگی. (گراممانیكی)

۲) ههر تلده سوزلهرنڭ یازلولارندا ههم سۆیلهنولهرنده قاعدهلهر
بولا. بزنڭ تۆرك تلندهده بیك كوب قاعدهلهر بار. بو قاعدهلهرنی
بلمهگهن كشی یازونی دۆرۆس ایتب یازا آلمی ههم سوزنیده
یابشدرب سۆیلی بلمی. بو قاعدهلهرگه روسچا «گراممانیكی»،
تۆركچه «تل بلگی» دیلهر.

(۱) عثمانلی شیوهسی. (۲) قازان شیوهسی. (۳) چغتای شیوهسی. (٤) قازاق شیوهسی.

(b)

بسم الله الرحمن الرحيم

تل و شیوه.

۱) بز روسیه مسلمانلاری تۆرك ملتندهن. بزنڭ تلبز تۆرك تلی.
بۆتۆن دنیاده ایللی میلیون قدر تۆرك بار. بو تۆركلهرنڭ تللهری بر
بولسهده، شیوهلهرنده (سۆیلهشولهرنده) آیرما بار. تۆركیه وقرم
تۆركلهری كۆن یاق (۱) شیوهسی بله سۆیلهشهلهر، بز تاتارلار،
باشقردلار ههم نوغایلار تۆن یاق (۲) شیوهسی بله سۆیلهشهبز.
تۆركستان تۆركلهری قۆیاش چغش (۳) شیوهسی بله سۆیلهشهلهر.
قازاغستان تۆركلهری دالا (٤) شیوهسی بله سۆیلهشهلهر.

بز روسیه مسلمانلاری قایسی ملت؟ بزنڭ تلبز نیندی تل؟ بۆتۆن
دنیادا نی قدر تۆرك بار؟ بولارنڭ تللهری نیندی تل؟ تۆركلهر ههر طرفده بر
تۆرلی سۆیلهشهلهرمی، ئهلله سۆیلهشولهرنده آیرما بارمی؟ بو آیرماغانی آیرماسی
دیلهر؟ تۆركیه وقرم تۆركلهری قایسی شیوه بله سۆیلهشهلهر؟ بز تاتارلار،
باشقردلار ههم نوغایلار قایسی شیوه بله سۆیلهشهبز؟ تۆركستان تۆركلهری قایسی
شیوه بله سۆیلهشهلهر؟ قازاغستان تۆركلهری قایسی شیوه بله سۆیلهشهلهر؟

تل بلگی. (گراممانیكی)

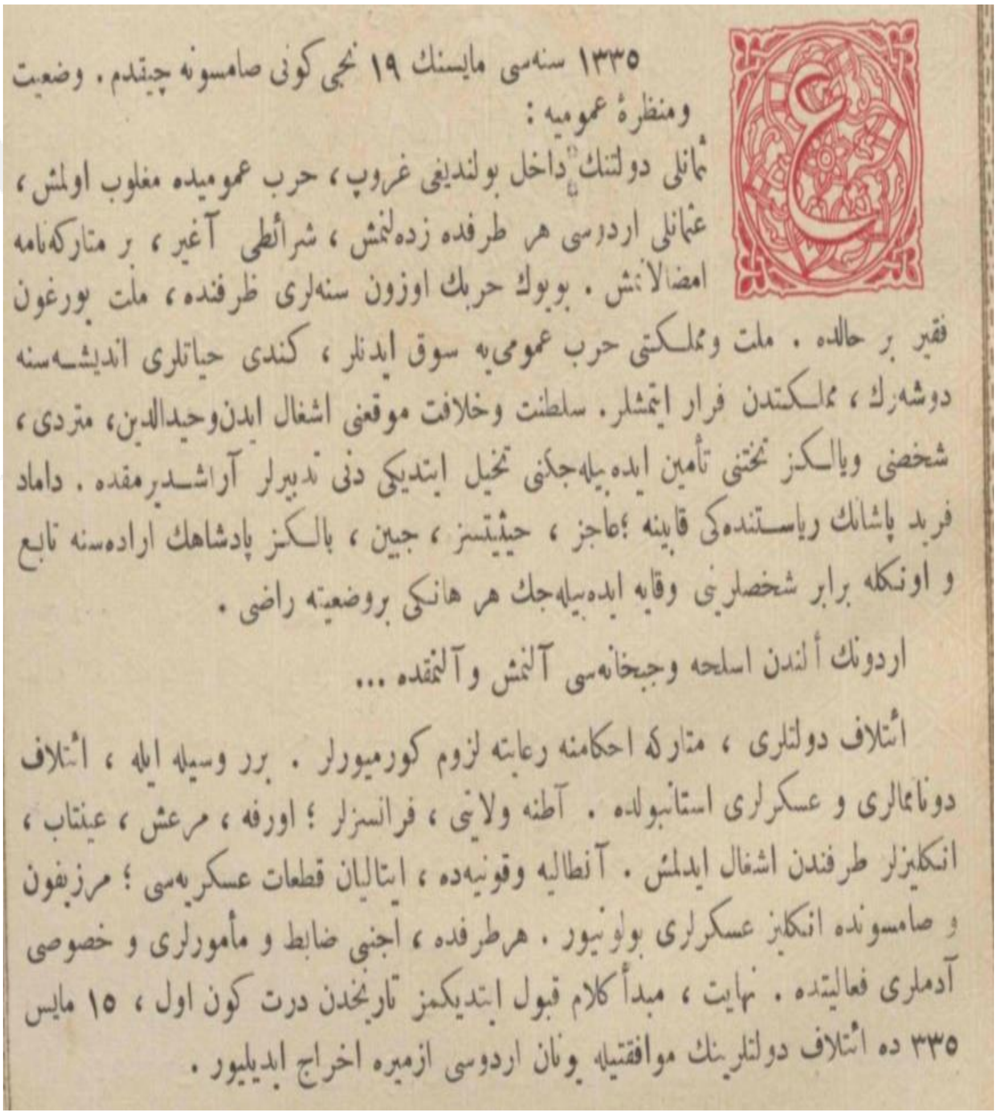
۲) ههر تلده سوزلهرنڭ یازلولارندا ههم سۆیلهنولهرنده قاعدهلهر
بولا. بزنڭ تۆرك تلندهده بیك كوب قاعدهلهر بار. بو قاعدهلهرنی
بلمهگهن كشی یازونی دۆرۆس ایتب یازا آلمی ههم سوزنیده
یابشدرب سۆیلی بلمی. بو قاعدهلهرگه روسچا «گراممانیكی»،
تۆركچه «تل بلگی» دیلهر.

(۱) عثمانلی شیوهسی. (۲) قازان شیوهسی. (۳) چغتای شیوهسی. (٤) قازاق شیوهسی.

Figure 3.7 a) Sample original Ottoman-Turkish image, b) Monochrome + adaptive-blur with value 50 using Imagemagick library

3.5. Ottoman-Turkish Character Recognition on the First Two Pages of the Nutuk

The Nutuk was a speech delivered by Mustafa Kemal Atatürk. Mustafa Kemal Ataturk was a revolutionary who helped establish the Republic of Turkey. He was Turkey's first president, and his reforms modernized the country. The Nutuks' first and second pages are applied to optical character recognition for Ottoman-Turkish. The first page of the Nutuk shows in the Figure 3.8.



١٣٣٥ سنه‌سی مایسنك ١٩ نجی كونی صامسونه چیقدم . وضعیت
ومنظرهٔ عمومیه :
ثمانلی دولتنك داخل بولندیغی غروپ ، حرب عمومیده مغلوب اولمش ،
عثمانلی اردوسی هر طرفده زده‌لنمش ، شرائطی آغیر ، بر متارکه‌نامه
امضالانمش . بویوك حربك اوزون سنه‌لری ظرفنده ، ملت یورغون
فقیر بر حالده . ملت ومملكتی حرب عمومی‌یه سوق ایدنلر ، كندی حیاتلری اندیشه‌سنه
دوشه‌رك ، مملكتدن فرار ایتمشلر . سلطنت وخلافت موقعنی اشغال ایدن وحیدالدین ، مترددی ،
شخصنی ویالكز تختنی تأمین ایده‌بیله‌جكنی تخیل ایتدیكی دنی تدبیرلر آراشدیرمقده . داماد
فرید پاشانك ریاستنده‌كی قابینه ؛ عاجز ، حیثیتسز ، جبین ، یالكز پادشاهك اراده‌سنه تابع
و اونكله برابر شخصلرینی وقایه ایده‌بیله‌جك هر هانكی بروضعیته راضی .

اردونك ألندن اسلحه وجبخانه‌سی آلنمش وآلنمقده ...

ائتلاف دولتلری ، متاركه احكامنه رعایته لزوم كورمیورلر . برر وسیله ایله ، ائتلاف
دونانمالری و عسكرلری استانبولده . آطنه ولایتی ، فرانسزلر ؛ اورفه ، مرعش ، عینتاب ،
انكلیزلر طرفندن اشغال ایدلمش . آنطالیه وقونیه‌ده ، ایتالیان قطعات عسكریه‌سی ؛ مرزیفون
و صامسونده انكلیز عسكرلری بولونیور . هرطرفده ، اجنبی ضابط و مأمورلری و خصوصی
آدملری فعالیتده . نهایت ، مبدأ كلام قبول ایتدیكمز تاریخدن درت كون اول ، ١٥ مایس
٣٣٥ ده ائتلاف دولتلرینك موافقتیله یونان اردوسی ازمیره اخراج ایدیلیور .

Figure 3.8 First page of the Nutuk

The result of optical character recognition for this page with Tesseract V3 shows in the appendix A. Also the following text is the ocr result for two lines.

سنهس مایسنك ثمتدكولى صامسوتح جيقدم ٠ وضعيث

نانلى دولتنك،إدأخل يرلنديني غروب ، حرب نموميده مغلوب اولمرء

The correct result must be for this part as follows:

١٣٣٥ سنه سي مايسنك ١٩ نجي كوني صامسونه جيقدم ٠ وضعيت

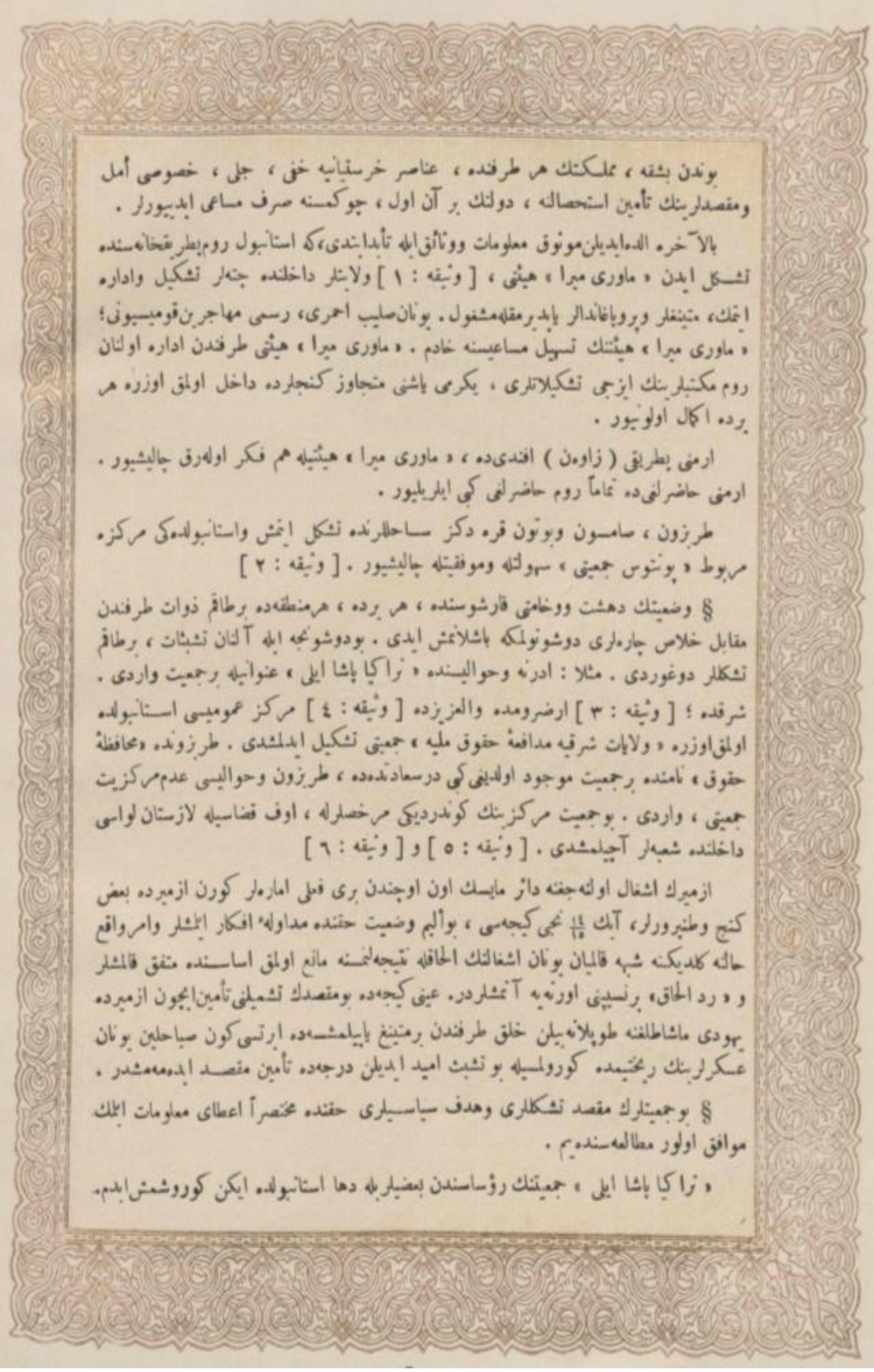
: و منظرۀ عموميه

، عثمانلى دولتنك داخل بولنديغى غروپ، حرب عموميده مغلوب اولمش

For these sentences, the green characters have been recognized correctly but the red characters have not been recognized after optical character recognition operation. For instance, ١٣٣٥ is not recognized after optical character recognition operation. مايسنك word is correctly recognized word.

For this page, the total number of characters is 958. The number of correctly founded characters is 727 and incorrectly founded characters are 231. According to these results recognition rate is 75.88%. The appendix A is shown correct and incorrect recognized characters in more detail.

The second page of the Nutuk shows in the Figure 3.9.



بوندن بشقه ، مملكتك هر طرفنده ، عناصر خرستيانيه خفى ، جلى ، خصوصى أمل ومقصدلرينك تأمين استحصاله ، دولتك بر آن اول ، چوكمسنه صرف مساعى ايديورلر .

بالآخره الده ايديلن موثوق معلومات ووثائق ايله تأييد ايتدى، كه استانبول روم بطريقخانه‌سنده تشكل ايدن « ماورى ميرا » هيئتى ، [وثيقه : ١] ولايتلر داخلنده چته‌لر تشكيل واداره ايتمك ، ميتينغلر وپروپاغاندالر ياپديرمقله مشغول . يونان صليب احمرى ، رسمى مهاجرين قوميسيونى؛ « ماورى ميرا » هيئتنك تسهيل مساعيسنه خادم . « ماورى ميرا » هيئتى طرفندن اداره اولنان روم مكتبلرينك ايزجى تشكيلاتلرى ، يكرمى ياشنى متجاوز كنجلرده داخل اولمق اوزره هر يرده اكمال اولونيور .

ارمنى بطريقى (زاوه‌ن) افندى‌ده ، « ماورى ميرا » هيئتيله هم فكر اوله‌رق چاليشيور . ارمنى حاضرلغى‌ده تماماً روم حاضرلغى كبى ايلرليور .

طربزون ، سامسون وبوتون قره دكز ساحللرنده تشكل ايتمش واستانبولده‌كى مركزه مربوط « پونتوس جمعيتى » سهولتله وموفقيتله چاليشيور . [وثيقه : ٢]

§ وضعيتك دهشت ووخامتى قارشوسنده ، هر يرده ، هر منطقه‌ده برطاقم ذوات طرفندن مقابل خلاص چاره‌لرى دوشونولمكه باشلانمش ايدى . بودوشونجه ايله آلنان تشبثات ، برطاقم تشكللر دوغوردى . مثلا : ادرنه وحواليسنده « تراكيا پاشا ايلى » عنوانيله برجمعيت واردى . شرقده ؛ [وثيقه : ٣] ارضرومده والعزيزده [وثيقه : ٤] مركز عموميسى استانبولده اولمق اوزره « ولايات شرقيه مدافعهٔ حقوق مليه » جمعيتى تشكيل ايدلمشدى . طربزونده «محافظهٔ حقوق » نامنده برجمعيت موجود اولديغى كبى درسعادتده‌ده ، طربزون وحواليسى عدم مركزيت جمعيتى ، واردى . بوجمعيت مركزينك كوندرديكى مرخصلرله ، اوف قضاسيله لازستان لواسى داخلنده شعبه‌لر آچيلمشدى . [وثيقه : ٥] و [وثيقه : ٦]

ازميرك اشغال اولنه‌جغنه دائر مايسك اون اوچندن برى فعلى اماره‌لر كورن ازميرده بعض كنج وطنپرورلر، آيك ١٤/١٥ نجى كيجه‌سى ، بو أليم وضعيت حقنده مداولهٔ افكار ايتمشلر وامرواقع حالنه كلديكنه شبهه قالميان يونان اشغالنك الحاقله نتيجه‌لنمسنه مانع اولمق اساسنده متفق قالمشلر و « رد الحاق » پرنسيپنى اورته‌يه آتمشلردر . عينى كيجه‌ده بومقصدك تشميلى تأمين ايچون ازميرده يهودى ماشاطلغنه طوپلانه‌بيلن خلق طرفندن برميتينغ ياپيلمشسه‌ده ايرتسى كون صباحلين يونان عسكرلرينك رختيمده كورولمسيله بو تشبث اميد ايديلن درجه‌ده تأمين مقصد ايده‌مه‌مشدر .

§ بوجمعيتلرك مقصد تشكللرى وهدف سياسيلرى حقنده مختصراً اعطاى معلومات ايتمك موافق اولور مطالعه‌سنده‌يم .

« تراكيا پاشا ايلى » جمعيتنك رؤساسندن بعضيلريله دها استانبولده ايكن كوروشمش ايدم.

Figure 3.9 Second page of the Nutuk

The result of optical character recognition for this page with Tesseract V3 shows in the appendix B. Also the following text is the ocr result for two lines.

برندن هيقه ، مملكتك ص طرفنده م عناصر خرسقيانيه غنى ، جلى ، خصوس أمل

ومقصدلرينك امين استحصاله ، دولتك بر ان اول ، جهكمسنه صرف مسار ابدكورلر ٠

The correct result must be for this part as follows:

بوندن بشقه ، مملكتك هر طرفنده , عناصر خرستيانيه خفى ، جلى ، خصوصي أمل

ومقصدلرينك تأمين استحصالنه ، دولتك بر آن اول ، چوكمسنه صرف مساعي ايدييورلر

٠

In this text, the green characters have also been recognized correctly and the red characters have not been recognized after optical character recognition operation. For instance, “و” character is not recognized in the “بوندن” word. Others characters in this word are recognized correctly.

For this page, the total number of characters is 1804 for the Nutuk’s second page. The number of correctly recognized characters is 1396 and the number of incorrectly recognized characters is 409. According to these results recognition rate is 77.38% for this page. The appendix B is shown correct and incorrect recognized characters in more detail.

CHAPTER FOUR

TRANSCRIPTION

“Transcription records or represents the sound or pronunciation of words in one alphabet using the letters of another alphabet” [28]. For Instance;

“يلديريم بيازيت كليه سي”

The transcription of the this text is “Yıldırım Beyazıt Külliyesi”

Transliteration is simply converts the letters of one alphabet to those of another alphabet [28].

The transliteration of the above text is “yldirim byazit klyesi”

Ottoman Text Archive Project (OTAP) [28], have worked about transcription techniques since 2002. This group works to order to an implementation of transcription systems.

4.1. Transcription Phases

There are three main phases for transcription operation. First phase is preprocessing operation. Second phase is used one to one transcription method. Third phase is transcription using regular expressions method.

4.1.1. Preprocessing Operation

In this phase, the Ottoman words are arranged before the transcription. Firstly, some special signs or characters remove from words. These signs are Ustun (◌َ), Esre (◌ِ), Otre (◌ُ), Jezma (◌ْ), Shedda (◌ّ), Medda (◌ٓ) signs. Also, (-) character removes from the words. This character also known as Arabic Tatweel character. For example, "هـر" word includes “ـ” character. Before transcription these characters are removed from words. Other removed characters are bidirectional control characters**.** These characters behave like a letter in a right to left or left to right script. These characters are also invisible characters.

Secondly, one of suffixes in table 4.1 is combined with the root word, if the suffix is unconnected with the root word.

Table 4.1 Suffixes.

كى	لر	سي
كي	له	يه
رك	دن	جك
سين	در	سنده
ملى	ك	سندن
سيز	لان	لك
جى		

Lastly, “ن" character is appending for Nunation (◌ٌ, ◌ً, ◌ٍ) signs before the database operations.

4.1.2. One To One Transcription Method

The Ottoman word is queried from the database table. If the Ottoman word and its Latin equivalent exist in database, transcription operation is successfully finished. This method is simple and correction ratio is high. But, if the Latin equivalent of Ottoman word does not exist in the database table, One to one transcription method is ineffective.

One2One table is used for this method. 65391 ottoman words and Latin equivalents exist in this table in the present. Figure 4.1 shows the one to one transcription method diagram.

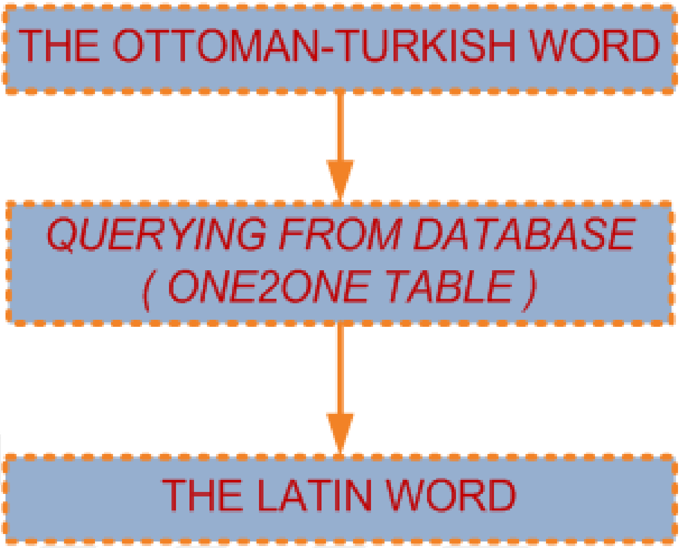
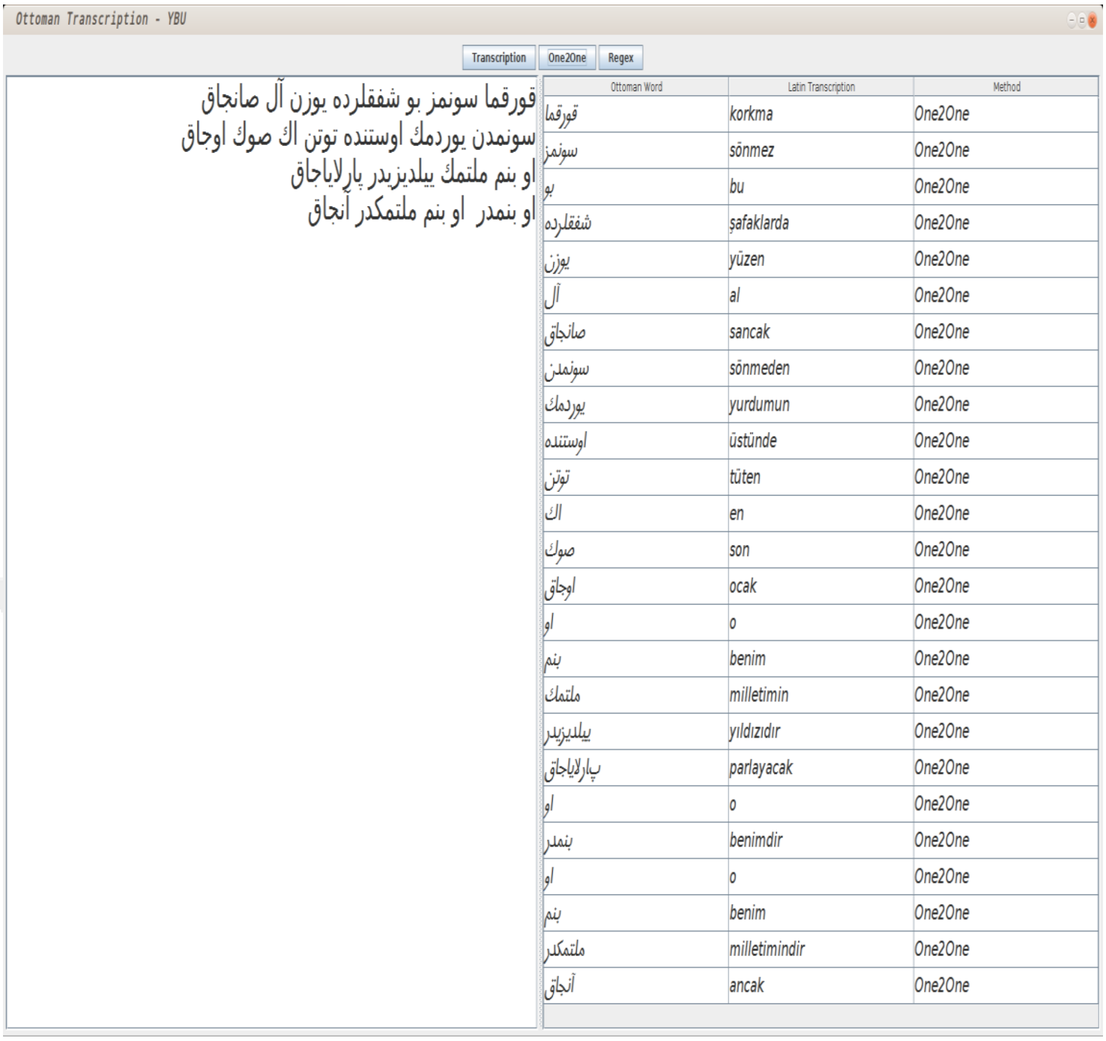


Figure 4.1 Diagram of one to one transcription method

4.1.2.1. Transcription Operation on the First Quatrain of Turkish National Anthem by One to One Transcription Method

Transcription is 100% successful for Turkish National Anthem because each of Ottoman-Turkish words and its Latin equivalents already contain in the One2one table.

Figure 4.2 shows result of the transcription result; According to this figure, the first quatrain of Turkish national anthem is in the text area with Ottoman-Turkish alphabet. After one2one button clicked, the table occurred in the figure. This table includes each of Ottoman-Turkish words, the Latin equivalent of the Ottoman-Turkish word and method name.



قورقما سونمز بو شفقلرده يوزن آل صانجاق
سونمدن يوردمك اوستنده توتن اك صوك اوجاق
او بنم ملتمك ييلديزيدر پارلاياجاق
او بنمدر او بنم ملتمكدر آنجاق

Ottoman Word	Latin Transcription	Method
قورقما	korkma	One2One
سونمز	sönmez	One2One
بو	bu	One2One
شفقلرده	şafaklarda	One2One
يوزن	yüzen	One2One
آل	al	One2One
صانجاق	sancak	One2One
سونمدن	sönmeden	One2One
يوردمك	yurdumun	One2One
اوستنده	üstünde	One2One
توتن	tüten	One2One
اك	en	One2One
صوك	son	One2One
اوجاق	ocak	One2One
او	o	One2One
بنم	benim	One2One
ملتمك	milletimin	One2One
ييلديزيدر	yıldızıdır	One2One
پارلاياجاق	parlayacak	One2One
او	o	One2One
بنمدر	benimdir	One2One
او	o	One2One
بنم	benim	One2One
ملتمكدر	milletimindir	One2One
آنجاق	ancak	One2One

Figure 4.2 Turkish national anthem's first quatrain with Ottoman-Turkish and transcription

4.1.3. Transcription using Regular Expressions Method

In this phase, firstly the last form of convenient regular expressions [RE] is created for Ottoman-Turkish word. Then the Latin word is tried to find from the modern ModernTurkishDictionary table by last form of convenient regular expressions. Currently, ModernTurkishDictionary table contains 1,146,560 modern Turkish-Latin words. This table is populated by using Zemberek Natural Language Processing Project's words [29]. If the word is not found in the ModernTurkishDictionary table, at this time, the word is tried to find OldTurkishDictionary Table. This table includes the old Turkish words which is not haven in modern Turkish-Latin table. Figure 4.3 shows the transcription using regular expressions method diagram.

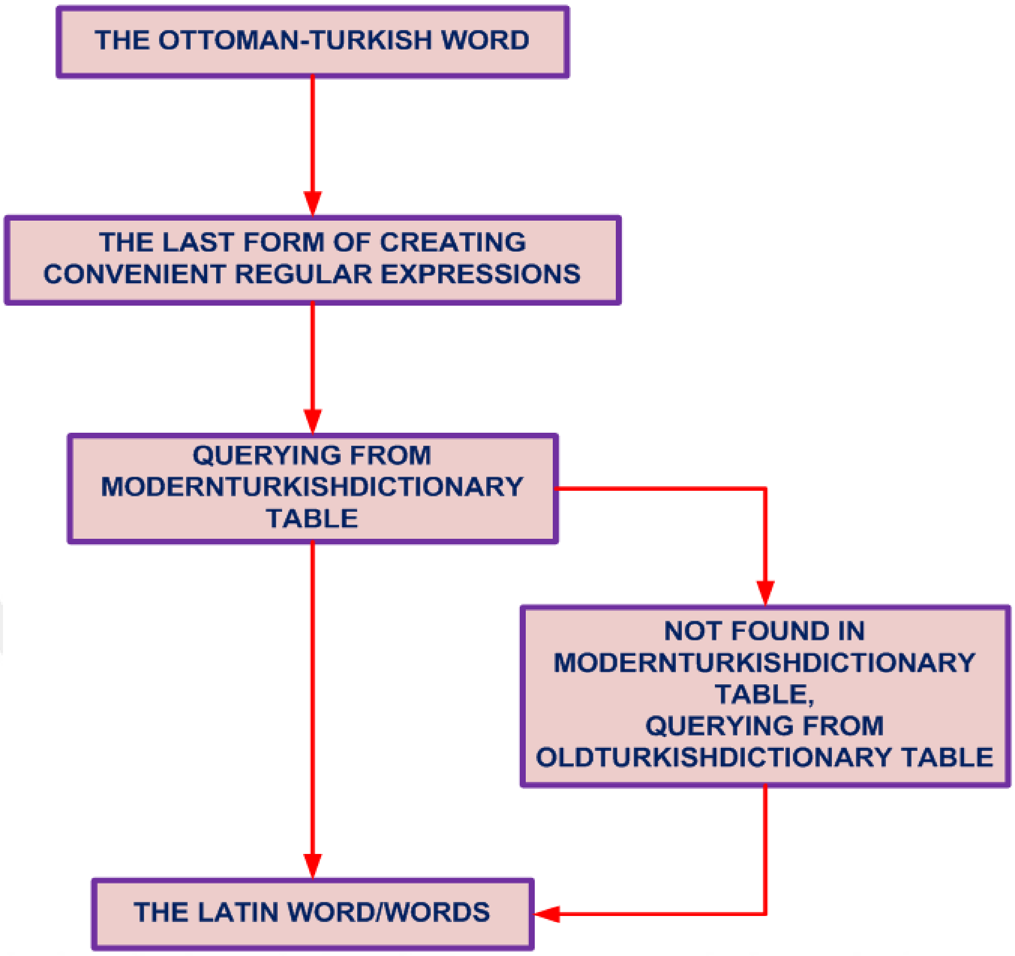

THE OTTOMAN-TURKISH WORD
THE LAST FORM OF CREATING CONVENIENT REGULAR EXPRESSIONS
QUERYING FROM MODERNTURKISHDICTIONARY TABLE
NOT FOUND IN MODERNTURKISHDICTIONARY TABLE, QUERYING FROM OLDTURKISHDICTIONARY TABLE
THE LATIN WORD/WORDS

Figure 4.3 Diagram of transcription using regular expressions method

4.1.3.1. Creating Convenient Regular Expressions

Regular expressions for each the Ottoman-Turkish word is generated according to the following regular expression generating rules tables. One of these tables is used for each letters in the word. Firstly, table 4.2 is used. If convenient sequential letter/letters do not exist, table 4.3 is used. If not, table 4.4 is used for generating regular expressions. Furthermore, (a|e|ı|i|u|ü|o|ö||) expression is inserted between consonant letters, except from ع, غ, ء,ى ,ي , و, ا, آ, أ, ه, ۀ letters. Because a vowel letter generally exists between two sequential consonants letters for transaction operation.

Table 4.2 Regular expression generating rules 1.

LOCATION	LETTERS	RE
The first three letters	ا و ی	oy|öy|uy|üy
The first two letters	ا و **or** أ و	av|ev|o|ö|u|ü
The first letter	ا	a|e|i|o
The first letter	و	v
The first letter	ه	h
The first letter	ی	y
The first letter	ي	y
The last three letters	د ا ش	daş|taş
The last three letters	نجی	nci|ncı|ncü|ncu
The last three letters	مسی	msi|msı
The last three letters	مسه	msa|mse
The last two letters	یه	ya|ye
The last two letters	مق	mak
The last two letters	مك	mek|mın
The last two letters	له	le|la

Table 4.2 (continued) Regular expression generating rules 1.

LOCATION	LETTERS	RE
The last two letters	لی	li|lı|lu|lü
The last two letters	کی	ki
The last two letters	سز	siz|sız|suz|süz
The last two letters	لك	lik|lık|luk|lük
The last two letters	ده	de|da|te|ta
The last two letters	دن	den|dan|ten|tan
The last two letters	لر	lar|ler
The last two letters	جی	cı|ci|cu|cü|çı|çi|çu|çü
The last two letters	جه	ce|ca|çe|ça
The last letter	ه	(a|e|i|ı|h)
The last letter	ی	i|ı|u|ü

Table 4.3 Regular expression generating rules 2.

LETTERS	RE	LETTERS	RE
ییو	yo|yö	اغ	a|ağ
عیه	iye	ای	e|i|ay|ey
لای	lay	اي	e|i|ay|ey
اوغ	uğ|oğ	وه	e|ve|vh
یا	ya	یو	yu|yü|yo|yö|iv|ıv
وی	v|y	او	av|ev|o|ö|u|ü
وو	v	أو	av|ev|o|ö|u|ü
يي	y	عه	a|e
وا	v|va	یه	ye|ya
ئو	o|u	لا	la|lı
عا	a	اع	a|ag

Table 4.4 Regular expression generating rules 3.

LETTER	RE	LETTER	RE	LETTER	RE
ب	b|p	ض	d|z	ى	y|e|ı|i|u|ü
ت	t	ط	t	ی	y|e|ı|i|u|ü
ث	s	ظ	z	ئ	y|e|ı|i|u|ü
پ	p	ف	f	ي	y|e|ı|i|u|ü
ج	c|ç	ق	k|g|ğ	ع	a|o|ö|u|ı|i||
چ	ç	ل	l	ء	i||
ح	h	م	m	غ	g|ğ|k|a|o
خ	h	ن	n	ك	k|g|ğ|n
د	d|t	ا	a|o	ڭ	k|g|ğ|n
ذ	z	آ	a	گ	k|g|ğ|n
ر	r	أ	a|e|o	ک	k|g|ğ|n
ز	z	إ	ı|i	ش	ş
ژ	j	و	v|u|ü|o|ö|ı|i	ه	a|e|i|ı|h
س	s	ؤ	v|u|ü|o|ö|ı|i		

For instance, "موستافه" word is an Ottoman-Turkish word and Latin equivalent is **Mustafa**. For this word, the last form of creating convenient regular expressions is as below;

“(m)(v|u|ü|o|ö|ı|i)(a|e|ı|i|u|ü|o|ö||)(s)(a|e|ı|i|u|ü|o|ö||)(t)(a|o)(a|e|ı|i|u|ü|o|ö||)(f)(a|e|i|ı|h)”

The Figure 4.4 shows how the last form of creating convenient regular expressions is generated by using rules tables in details.

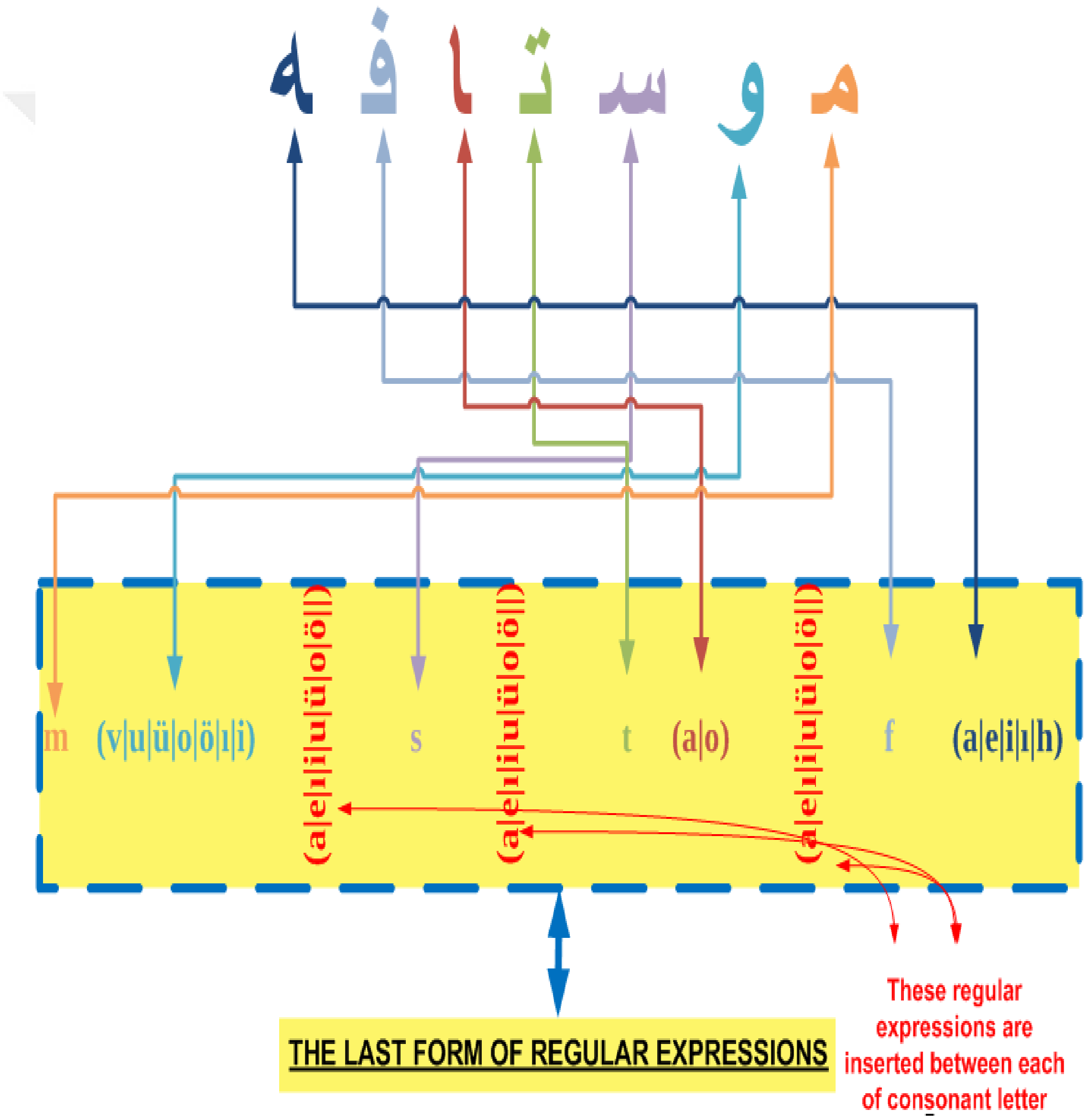

م
و
س
ت
ا
ف
ه
m
(v|u|ü|o|ö|ı|i)
(a|e|ı|i|u|ü|o|ö||)
s
(a|e|ı|i|u|ü|o|ö||)
t
(a|o)
(a|e|ı|i|u|ü|o|ö||)
f
(a|e|i|ı|h)
THE LAST FORM OF REGULAR EXPRESSIONS
These regular expressions are inserted between each of consonant letter

Figure 4.4 Detailed generating regular expression using rules tables for each letter

The regular expression of each of Ottoman-Turkish letters is as below for this word:

“م”→ (m)

“و” → (v|u|ü|o|ö|ı|i)

“س” → (s)

“ت” → (t)

“ا” → (a|o)

“ف” → (f)

“ه” → (e|a)

Additionally, “(a|e|ı|i|u|ü|o|ö||)” expression is inserted between the consonant letters.

Another example:

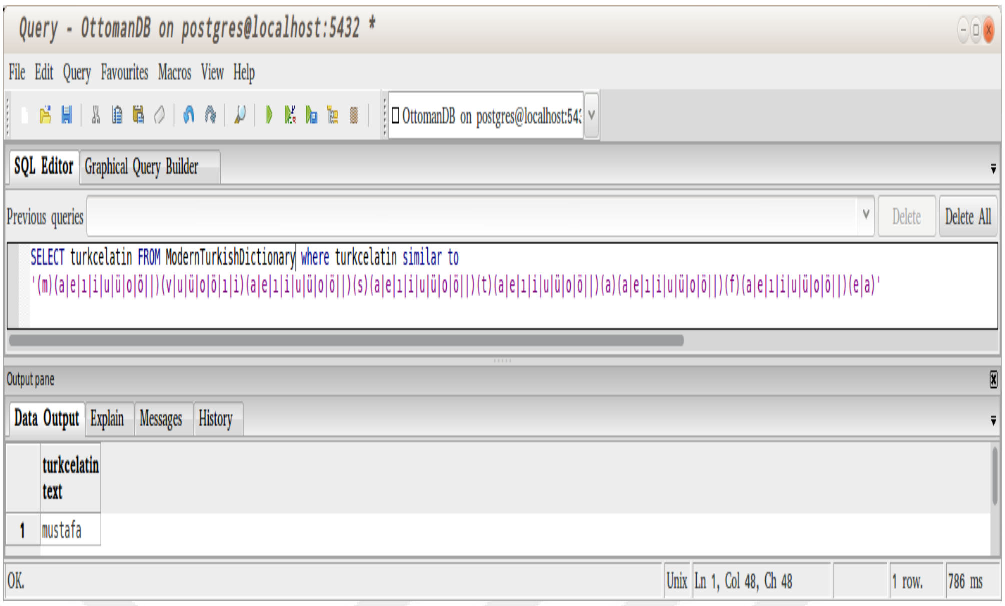
“**اويالا**” word translates “**oyala**”. For this word, regular expression is generated as below;

(oy|öy|uy|üy)(a|e|ı|i|u|ü|o|ö||)(a)(a|e|ı|i|u|ü|o|ö||)(la|lı)

(oy|öy|uy|üy) expression is generated that because of first two letters are “اوى" as shown in Table 4.2. (la|lı) expression is generated that because of last letter are “لا" as shown in Table 4.3.

4.1.3.2. Look at Modern Turkish-Latin Table

All records are scanned in the ModernTurkishDictionary by the last form of the created regular expressions. If convenient Turkish-Latin word/words find, transcription operation is successfully finished. For instance, the select query in the Figure 4.5 runs for “**موستافه**” word.


Query - OttomanDB on postgres@localhost:5432 *
File Edit Query Favourites Macros View Help
OttomanDB on postgres@localhost:543
SQL Editor
Graphical Query Builder
Previous queries
Delete
Delete All
Output pane
Data Output
Explain
Messages
History

```
SELECT turkcelatin FROM ModernTurkishDictionary where turkcelatin similar to
'(m)(a|e|ı|i|u|ü|o|ö||)(v|u|ü|o|ö|ı|i)(a|e|ı|i|u|ü|o|ö||)(s)(a|e|ı|i|u|ü|o|ö||)(t)(a|e|ı|i|u|ü|o|ö||)(a)(a|e|ı|i|u|ü|o|ö||)(f)(a|e|ı|i|u|ü|o|ö||)(e|a)'
```

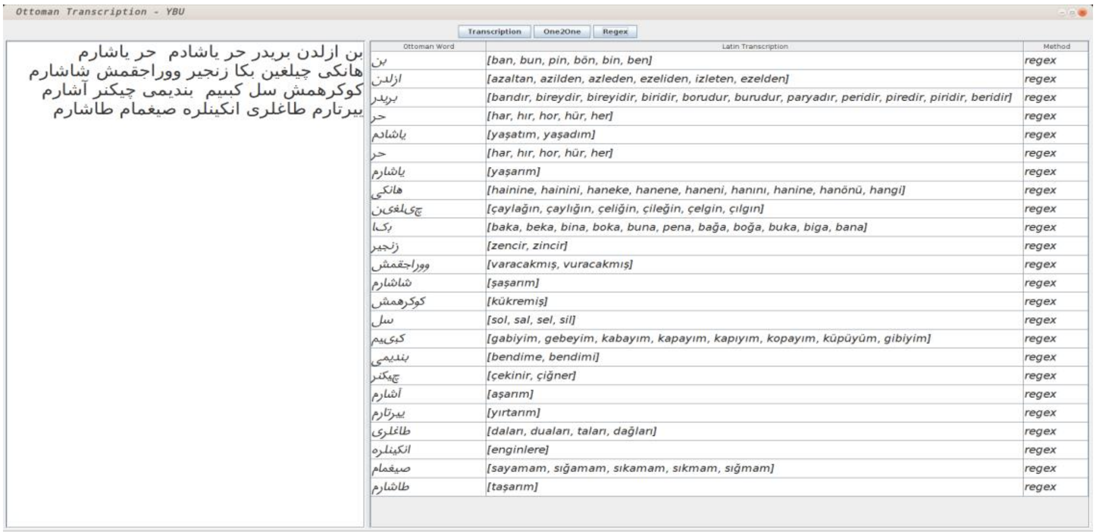
	turkcelatin text
1	mustafa

```
OK.    Unix  Ln 1, Col 48, Ch 48    1 row.  786 ms
```

Figure 4.5 The query for موستافه word

4.1.3.3. Transcription Operation on the Third Quatrain of Turkish National Anthem by Transcription Using Regular Expressions Method

Transcription with regular expressions on the third quatrain of Turkish national anthem shows in the Figure 4.6.



بن ازلدن بریدر حر یاشادم حر یاشارم
هانکی چیلغین بکا زنجیر ووراجقمش شاشارم
کوکرهمش سل کبیم بندیمی چیکنر آشارم
ییرتارم طاغلری انکینلره صیغمام طاشارم

Ottoman Word	Latin Transcription	Method
بن	[ban, bun, pin, bön, bin, ben]	regex
ازلدن	[azaltan, azilden, azleden, ezeliden, izleten, ezelden]	regex
بریدر	[bandır, bireydir, bireyidir, biridir, borudur, burudur, paryadır, peridir, piredir, piridir, beridir]	regex
حر	[har, hır, hor, hür, her]	regex
یاشادم	[yaşatım, yaşadım]	regex
حر	[har, hır, hor, hür, her]	regex
یاشارم	[yaşarım]	regex
هانکی	[hainine, hainini, haneke, hanene, haneni, hanını, hanine, hanönü, hangi]	regex
چیلغین	[çaylağın, çaylığın, çeliğin, çileğin, çelgin, çılgın]	regex
بکا	[baka, beka, bina, boka, buna, pena, bağa, boğa, buka, biga, bana]	regex
زنجیر	[zencir, zincir]	regex
ووراجقمش	[varacakmış, vuracakmış]	regex
شاشارم	[şaşarım]	regex
کوکرهمش	[kükremiş]	regex
سل	[sol, sal, sel, sil]	regex
کبیم	[gabiyim, gebeyim, kabayım, kapayım, kapıyım, kopayım, küpüyüm, gibiyim]	regex
بندیمی	[bendime, bendimi]	regex
چیکنر	[çekinir, çiğner]	regex
آشارم	[aşarım]	regex
ییرتارم	[yırtarım]	regex
طاغلری	[dalanı, duaları, taları, dağları]	regex
انکینلره	[enginlere]	regex
صیغمام	[sayamam, sığamam, sıkamam, sıkmam, sığmam]	regex
طاشارم	[taşarım]	regex

Figure 4.6 Transcription using regular expressions on the third quatrain of Turkish national anthem

The third quatrain of Turkish national anthem is in the text area with Ottoman-Turkish Alphabet. After regex button clicked, the table occurred in the Figure. This table contains each of Ottoman-Turkish words, the Latin equivalent of the Ottoman-Turkish word and method name. According to this figure, the following words' Latin transcription is found correctly, but incorrect words are found beside the correct word.

بن , ازلدن , بريدر , حر , ياشادم , هانكى , چيلغين , بكا , زنجير , ووراجقمش ,

سل , كبديم , بنديمى , چيكنر , طاغلرى , صيغمام

For instance, "طاغلرى" is correctly transcription of “dağları”. But “daları”, “duaları”, “taları” are incorrect transcription of this word.

The following words' Latin transcription is found accurately,

, طاشارم , انكينلره , ييرتارم , آشارم كوكرهمش , شاشارم , ياشارم

4.2. Comparison of One To One Transcription and Transcription Using Regular Expressions

- One to one Transcription method is simple, Transcription using regular expressions method is more complex,
- One to one transcription method is more successful than transcription using regular expressions method.
- If the Latin equivalent of Ottoman word does not exist in the One2One Table, One to one transcription method is ineffective.
- Transcription using regular expressions method is feasible for all ottoman words. But one to one transcription method is not feasible.
- Transcription using regular expressions method could be found incorrect words.

- The success rate of transcription using regular expressions method depends on the length of Ottoman words. If the length of word increases, then the success rate of transcription. also increases.

Because of above reasons, one to one transcription method primarily is used when transcription operation is done. If transcription is not success, then transcription using regular expressions method is used. In transcription using regular expressions method, firstly, modern-Turkish table is used, If the Latin record is not exist in modern-Turkish table, old-Turkish table is used for transcription.

CHAPTER FIVE

OTTOMAN-TURKISH KEYBOARD

Handwriting or most of calligraphic Ottoman documents are not recognized with optical character recognition techniques. So, Ottoman-Turkish Keyboard is designed for writing for these kinds of documents. Wrongly words or characters could also be edited after optical character recognition process through this keyboard. Ottoman-Turkish Keyboard is shown in the Figure 5.1. Ottoman Text could be written with keyboard action and bottom buttons using this keyboard.

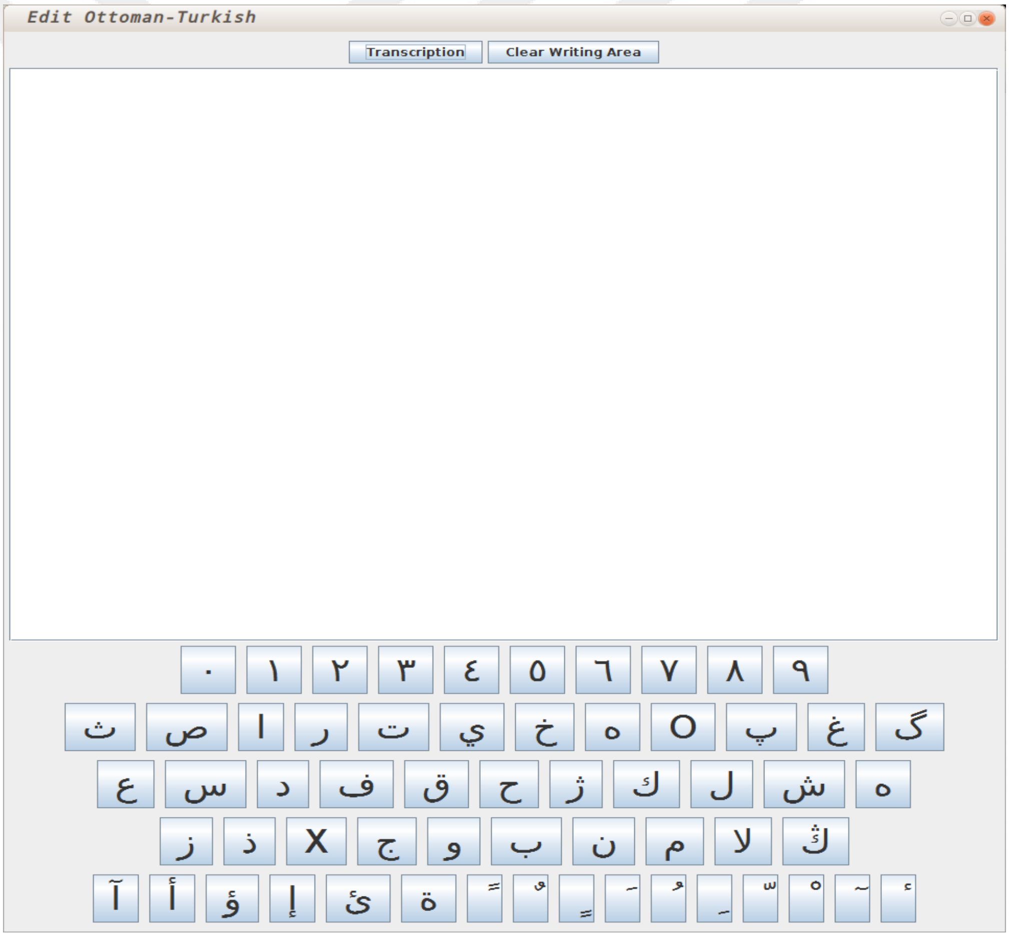


Figure 5.1 Ottoman-Turkish Keyboard

5.1. The Transcription of an Inscription Using the Ottoman-Turkish Keyboard

The Figure 5.2 shows an inscription with Ottoman-Turkish alphabet. This inscription was used Sultan Ahmet Prison until 1970.

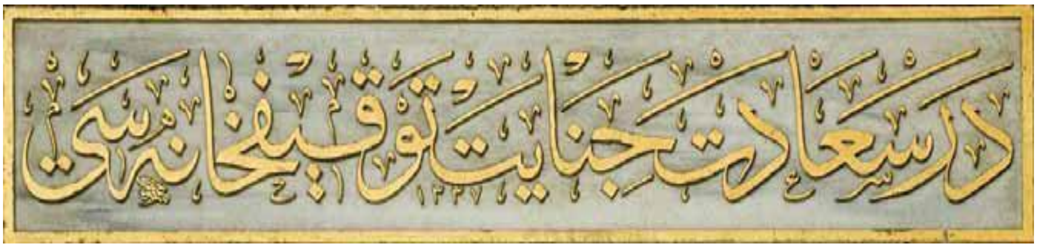
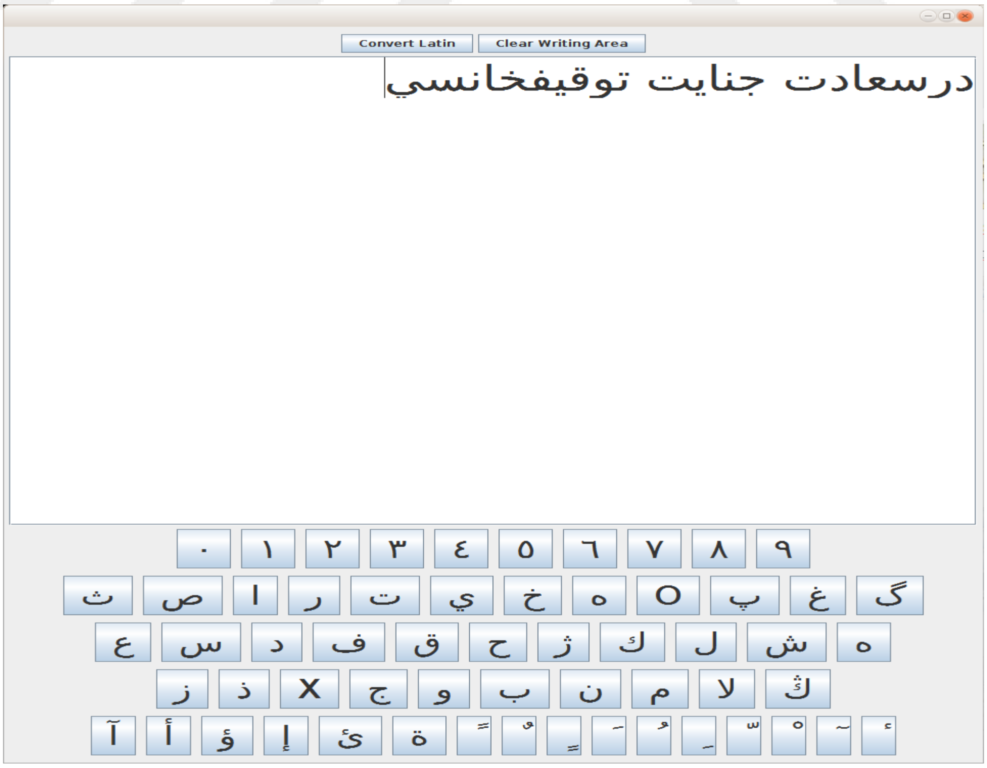


Figure 5.2 An Ottoman Inscription

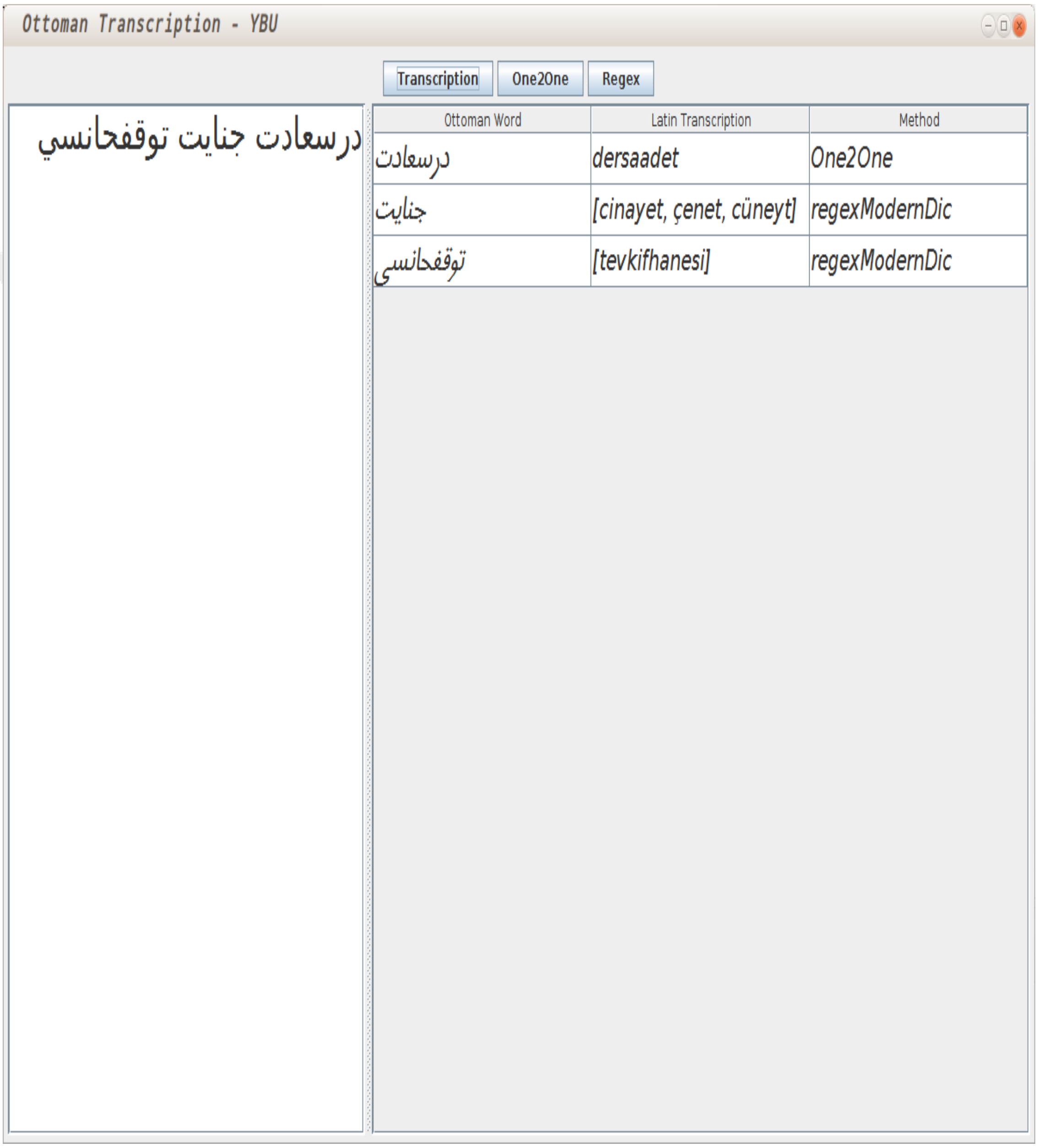
The transcription of this inscription is “Dersaadet Cinayet Tevkifhanesi”. Ottoman-Turkish Keyboard could be used in order to write this inscription. Only letters have written, the other orthographic signs haven't written for transaction of this inscription.



درسعادت جنايت توقيفخانسي

Figure 5.3 The writing of the inscription with Ottoman-Turkish Alphabet

Transcription of this inscription shows in figure 5.4. "درسعادت" and "توقفحانسي" words are found successfully. But cinayet, çenet and cüneyt are found for “جنايت” word. “cinayet” is correct transcription of this word. “çenet” and “cüneyt” is incorrect transcription of word.



درسعادت جنايت توقفحانسي

Ottoman Word	Latin Transcription	Method
درسعادت	dersaadet	One2One
جنايت	[cinayet, çenet, cüneyt]	regexModernDic
توقفحانسي	[tevkifhanesi]	regexModernDic

Figure 5.4 Transcription of the inscription

5.2. The Transcription of an Ottoman Tombstone Using the Ottoman-Turkish Keyboard

The following Ottoman tombstone have used with the purpose of transcription.

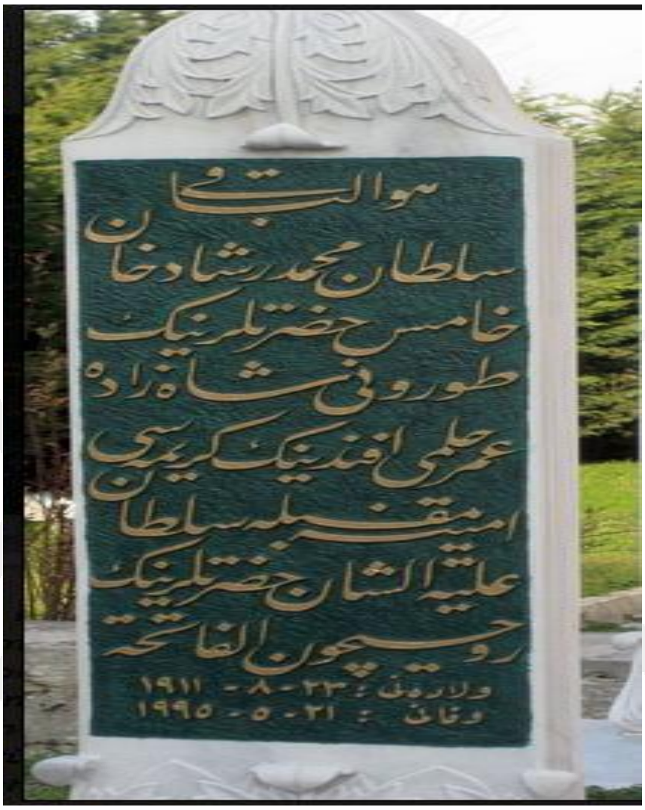
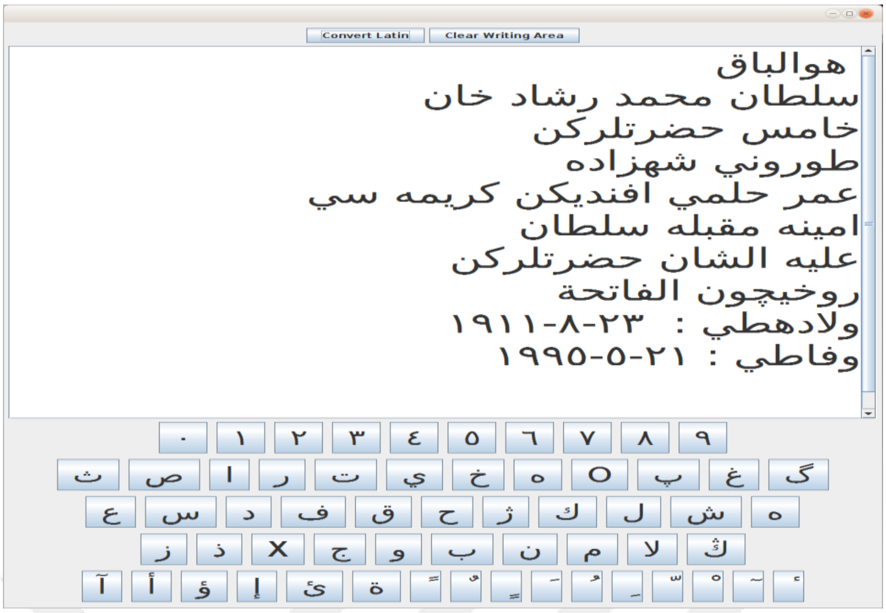


Figure 5.5 Ottoman tombstone

The transcription of this Ottoman tombstone is “Sultan Mehmed Reşad Han Hamis Hazretlerinin Torunu Şehzade Ömer Hilmi Efendinin Kerimesi Emine Makbule Sultan Aleyhişşan Hazretlerinin Ruhiçün Elfatiha Veledati: 23-8-1911 Vefat: 21-5-1995”. Ottoman-Turkish Keyboard could be used in order to write this Ottoman tombstone. The writing of the tombstone with used Ottoman-Turkish Keyboard is shown in the Figure 5.6.



هوالباق
سلطان محمد رشاد خان
خامس حضرتلركن
طوروني شهزاده
عمر حلمي افنديكن كريمه سي
امينه مقبله سلطان
عليه الشان حضرتلركن
روخيچون الفاتحة
ولادهطي : ١٩١١-٨-٢٣
وفاطي : ١٩٩٥-٥-٢١

Figure 5.6 The writing of the Ottoman tombstone with Ottoman-Turkish alphabet

The transcription of this Ottoman tombstone is shown in Table 5.1. According to this table, the following words' Latin transcription is found accurately.

"هوالباق", "سلطان", "محمد", "رشاد", "خان", "خامس", "شهزاده", "عمر"

"افنديكن", "سلطان", "عليه", "الشان", "روخيچون", "الفاتحة", " ولادهطي", "٢", "٣",

"٩","٩", "٥","٨", "١", "٩", "١", "١", "وفاطي", "٢", "١", "٥", "١"

The following words' Latin transcription could be found correctly, but incorrect words have found beside the correct word.

“حضرتلركن”, “مقبله”, “امينه”, “كريمهسي”, “حلمي”, “طوروني”, “حضرتلركن”

For instance, “*طوروني*" is correctly transcription of “torunu”. But "dairini", "davarını", "devrine", "devrini", "durunu", "duvarını", "tavrını", "teorine", "teorini", "törüne", "törünü", "turunu", "türevine", "türevini", "türüne", "türünü" are wrongly transcription of this word.

Table 5.1 Transcription of the tombstone.

Ottoman Word	Latin Transcription	Method
هوالباق	**Hüve'l-bâki**	One2One
سلطان	**sultan**	One2One
محمد	**Mehmed**	One2One
رشاد	**reşad**	One2One
خان	**han**	One2One
خامس	**[hamis]**	regexOldDic
حضرتلركن	[**hazretlerinin**, hidratlarının]	regexModernDic
طوروني	[dairini, davarını, devrine, devrini, durunu, duvarını, tavrını, teorine, teorini, **torunu**, törüne, törünü, turunu, türevine, türevini, türüne, türünü]	regexModernDic
شهزاده	**şehzade**	One2One
عمر	**Ömer**	One2One
حلمي	[halamı, haleme, halemi, halımı, halime, halimi, heleme, helme, hılımı, **hilmi**]	regexModernDic
افنديكن	**[efendinin]**	regexModernDic
كريمهسي	[giremese, göremese, **kerimesi**, koruması, kreması, kuruması, küremesi, kürümesi]	regexModernDic
امينه	[amine, amiyane, emene, **emine**, imine]	regexModernDic
مقبله	[makable, **makbule**, mukabele, mukbile]	regexModernDic
سلطان	**sultan**	One2One
عليه	**aleyhi**	One2One
الشان	**-şşan**	One2One
حضرتلركن	[**hazretlerinin**, hidratlarının]	regexModernDic
روخيچون	**[ruhiçün]**	regexOldDic
الفاتحة	**elfatiha**	One2One
ولادهطي	**[veladeti]**	regexModernDic
:	**:**	One2One
٢	**2**	One2One
٣	**3**	One2One

Table 5.1 (Continued) Transcription of the tombstone.

Ottoman Word	Latin Transcription	Method
-	**-**	One2One
٨	**8**	One2One
-	**-**	One2One
١	**1**	One2One
٩	**9**	One2One
١	**1**	One2One
١	**1**	One2One
وفاطي	**[vefatı]**	regexModernDic
:	**:**	One2One
٢	**2**	One2One
١	**1**	One2One
-	**-**	One2One
٥	**5**	One2One
-	**-**	One2One
١	**1**	One2One
٩	**9**	One2One
٩	**9**	One2One
٥	**5**	One2One

CHAPTER SIX

DATABASE and SOFTWARE

This chapter explains the process of design and implementation of software and database in this study.

6.1. Ottoman Optical Character Recognition and Transcription Software

This software is developed in Eclipse platform. Eclipse platform is one of the famous java integrated development environment. Also this software is implemented by java language.

The graphical main user interface shows in the appendix C. An Ottoman-Turkish images or documents can open in the image panel. Imagemagick library input values could be adjusted in this GUI and also some image processing operations like zooming or rotation image could be done by using this interface. In addition to, OCR operation could be done by OCR button and the OCR result show in the right text area.

Ottoman-Turkish Keyboard shows the appendix D. This keyboard already discuss in the chapter five. Transcription tool also shows the appendix E. This tool could be used to transcription from Ottoman-Turkish text to Latin text. If the table row in the transcription frame clicks, enrichment database tool opens. This tool could also used to enrich data in the database. The implemented classes show in the Figure 6.1.

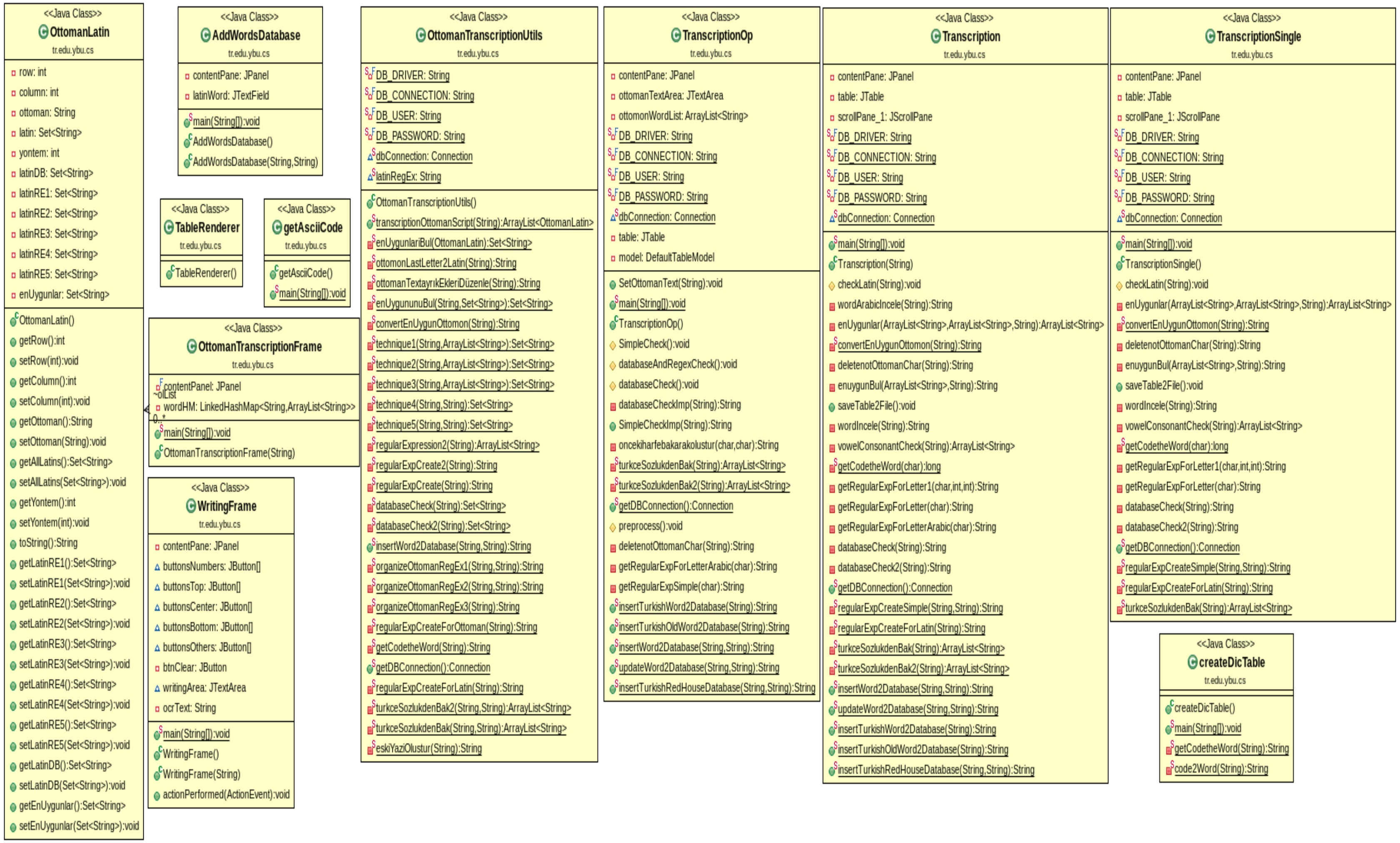
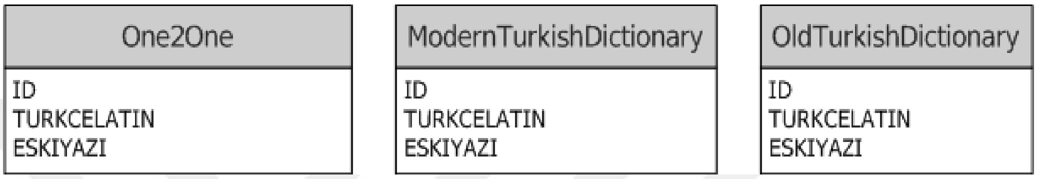

<<Java Class>>
OttomanLatin
tr.edu.ybu.cs
row: int
column: int
ottoman: String
latin: Set<String>
yontem: int
latinDB: Set<String>
latinRE1: Set<String>
latinRE2: Set<String>
latinRE3: Set<String>
latinRE4: Set<String>
latinRE5: Set<String>
enUygunlar: Set<String>
OttomanLatin()
getRow():int
setRow(int):void
getColumn():int
setColumn(int):void
getOttoman():String
setOttoman(String):void
getAllLatins():Set<String>
setAllLatins(Set<String>):void
getYontem():int
setYontem(int):void
toString():String
getLatinRE1():Set<String>
setLatinRE1(Set<String>):void
getLatinRE2():Set<String>
setLatinRE2(Set<String>):void
getLatinRE3():Set<String>
setLatinRE3(Set<String>):void
getLatinRE4():Set<String>
setLatinRE4(Set<String>):void
getLatinRE5():Set<String>
setLatinRE5(Set<String>):void
getLatinDB():Set<String>
setLatinDB(Set<String>):void
getEnUygunlar():Set<String>
setEnUygunlar(Set<String>):void
<<Java Class>>
AddWordsDatabase
tr.edu.ybu.cs
contentPane: JPanel
latinWord: JTextField
main(String[]):void
AddWordsDatabase()
AddWordsDatabase(String,String)
<<Java Class>>
TableRenderer
tr.edu.ybu.cs
TableRenderer()
<<Java Class>>
getAsciiCode
tr.edu.ybu.cs
getAsciiCode()
main(String[]):void
<<Java Class>>
OttomanTranscriptionFrame
tr.edu.ybu.cs
contentPanel: JPanel
~olList
wordHM: LinkedHashMap<String,ArrayList<String>>
0..*
main(String[]):void
OttomanTranscriptionFrame(String)
<<Java Class>>
WritingFrame
tr.edu.ybu.cs
contentPane: JPanel
buttonsNumbers: JButton[]
buttonsTop: JButton[]
buttonsCenter: JButton[]
buttonsBottom: JButton[]
buttonsOthers: JButton[]
btnClear: JButton
writingArea: JTextArea
ocrText: String
main(String[]):void
WritingFrame()
WritingFrame(String)
actionPerformed(ActionEvent):void
<<Java Class>>
OttomanTranscriptionUtils
tr.edu.ybu.cs
DB_DRIVER: String
DB_CONNECTION: String
DB_USER: String
DB_PASSWORD: String
dbConnection: Connection
latinRegEx: String
OttomanTranscriptionUtils()
transcriptionOttomanScript(String):ArrayList<OttomanLatin>
enUygunlariBul(OttomanLatin):Set<String>
ottomonLastLetter2Latin(String):String
ottomanTextayrikEkleriDüzenle(String):String
enUygununuBul(String,Set<String>):Set<String>
convertEnUygunOttomon(String):String
technique1(String,ArrayList<String>):Set<String>
technique2(String,ArrayList<String>):Set<String>
technique3(String,ArrayList<String>):Set<String>
technique4(String,String):Set<String>
technique5(String,String):Set<String>
regularExpression2(String):ArrayList<String>
regularExpCreate2(String):String
regularExpCreate(String):String
databaseCheck(String):Set<String>
databaseCheck2(String):Set<String>
insertWord2Database(String,String):String
organizeOttomanRegEx1(String,String):String
organizeOttomanRegEx2(String,String):String
organizeOttomanRegEx3(String):String
regularExpCreateForOttoman(String):String
getCodetheWord(String):String
getDBConnection():Connection
regularExpCreateForLatin(String):String
turkceSozlukdenBak2(String,String):ArrayList<String>
turkceSozlukdenBak(String,String):ArrayList<String>
eskiYaziOlustur(String):String
<<Java Class>>
TranscriptionOp
tr.edu.ybu.cs
contentPane: JPanel
ottomanTextArea: JTextArea
ottomonWordList: ArrayList<String>
DB_DRIVER: String
DB_CONNECTION: String
DB_USER: String
DB_PASSWORD: String
dbConnection: Connection
table: JTable
model: DefaultTableModel
SetOttomanText(String):void
main(String[]):void
TranscriptionOp()
SimpleCheck():void
databaseAndRegexCheck():void
databaseCheck():void
databaseCheckImp(String):String
SimpleCheckImp(String):String
oncekiharfebakarakolustur(char,char):String
turkceSozlukdenBak(String):ArrayList<String>
turkceSozlukdenBak2(String):ArrayList<String>
getDBConnection():Connection
preprocess():void
deletenotOttomanChar(String):String
getRegularExpForLetterArabic(char):String
getRegularExpSimple(char):String
insertTurkishWord2Database(String):String
insertTurkishOldWord2Database(String):String
insertWord2Database(String,String):String
updateWord2Database(String,String):String
insertTurkishRedHouseDatabase(String,String):String
<<Java Class>>
Transcription
tr.edu.ybu.cs
contentPane: JPanel
table: JTable
scrollPane_1: JScrollPane
DB_DRIVER: String
DB_CONNECTION: String
DB_USER: String
DB_PASSWORD: String
dbConnection: Connection
main(String[]):void
Transcription(String)
checkLatin(String):void
wordArabicIncele(String):String
enUygunlar(ArrayList<String>,ArrayList<String>,String):ArrayList<String>
convertEnUygunOttomon(String):String
deletenotOttomanChar(String):String
enuygunBul(ArrayList<String>,String):String
saveTable2File():void
wordIncele(String):String
vowelConsonantCheck(String):ArrayList<String>
getCodetheWord(char):long
getRegularExpForLetter1(char,int,int):String
getRegularExpForLetter(char):String
getRegularExpForLetterArabic(char):String
databaseCheck(String):String
databaseCheck2(String):String
getDBConnection():Connection
regularExpCreateSimple(String,String):String
regularExpCreateForLatin(String):String
turkceSozlukdenBak(String):ArrayList<String>
turkceSozlukdenBak2(String):ArrayList<String>
insertWord2Database(String,String):String
updateWord2Database(String,String):String
insertTurkishWord2Database(String):String
insertTurkishOldWord2Database(String):String
insertTurkishRedHouseDatabase(String,String):String
<<Java Class>>
TranscriptionSingle
tr.edu.ybu.cs
contentPane: JPanel
table: JTable
scrollPane_1: JScrollPane
DB_DRIVER: String
DB_CONNECTION: String
DB_USER: String
DB_PASSWORD: String
dbConnection: Connection
main(String[]):void
TranscriptionSingle()
checkLatin(String):void
enUygunlar(ArrayList<String>,ArrayList<String>,String):ArrayList<String>
convertEnUygunOttomon(String):String
deletenotOttomanChar(String):String
enuygunBul(ArrayList<String>,String):String
saveTable2File():void
wordIncele(String):String
vowelConsonantCheck(String):ArrayList<String>
getCodetheWord(char):long
getRegularExpForLetter1(char,int,int):String
getRegularExpForLetter(char):String
databaseCheck(String):String
databaseCheck2(String):String
getDBConnection():Connection
regularExpCreateSimple(String,String):String
regularExpCreateForLatin(String):String
turkceSozlukdenBak(String):ArrayList<String>
<<Java Class>>
createDicTable
tr.edu.ybu.cs
createDicTable()
main(String[]):void
getCodetheWord(String):String
code2Word(String):String

Figure 6.1 Implemented java classes

6.2. Ottoman Dictionary Database

PostgreSQL database is one of the most advanced open source databases. In this study, PostgreSQL database environment is used for the Ottoman Dictionary database. Ottoman dictionary database is key point for transcription of the Ottoman-Turkish texts. This database is a kind of dictionary database. Tables and columns of this database are shown in Figure 6.2.



One2One	ModernTurkishDictionary	OldTurkishDictionary
ID TURKCELATIN ESKIYAZI	ID TURKCELATIN ESKIYAZI	ID TURKCELATIN ESKIYAZI

Figure 6.2 Database tables

One2One table is used for One to One transcription. There is the Latin word equation for each Ottoman-Turkish word in this table. Currently, 65.420 rows exist in One2One table. Redhouse’s Lexicon [32] and Kanar’s dictionary [33] are benefitted for the vocabulary for this table. ModernTurkishDictionary table is used transcription of using regular expressions. The important column is TurkceLatin. There are 1.146.560 rows exist in this table for the present. OldTurkishDictionary table is also used transcription of using regular expressions. Currently, the number of rows is unsatisfactory in this table.

Enrichment of database tables is important for transcription. If all Latin words would include Ottoman-Turkish word in database, transcription and reversed transcription would be possible for all scripts.

CHAPTER SEVEN

CONCLUSION

In this thesis, optical character recognition and Latin transcription are integrated to convert modern Turkish Language from Ottoman Scripts. Experiments for optical character recognition show that recognition rates could change quality, style and printed or handwritten documents or images. The recognition rates for high quality and printed images can be with an accuracy of 100%. Optical character recognition is not possible for some handwriting or low quality documents or images. So these kind of documents or images need to write again with Ottoman-Turkish.

Also, novel approaches are presented for the transcription of Ottoman-Turkish texts. Transcription process is depending on dictionary database. The success rate of transcription can rise as high as an accuracy of 100%. But transcription is not done successfully for some words, if the word does not exist in the database.

As a future work, it is possible to append Persian character for Tesseract's trained data. In this way, the performance of optical character recognition will increase with new datasets. New version of Tesseract is worked for recognition for Persian characters. New transcription methods could be developed after the Arabic and Persian morphologies deeply analyze, It is also possible improve to enrich data in dictionary database for more successful transcription. Enrichment database may provide to prepare comprehensive lexicons. Also web and mobile applications could be developed for OCR and Latin transcription. Another future research is that Latin-based Turkish texts may transcript Ottoman-Turkish. In this manner, newspapers, books or speeches could be transcribed to Ottoman-Turkish. Thus Ottoman-Turkish can gain popularity as before.

REFERENCES

[1] Ozturk, A., Gunes, S. and Ozbay, Y. *“Multifont Ottoman Character Recognition”,* Proceedings of the 7th IEEE Int. Conf. on Electronics Circuits and Systems (ICECS), 2000, Jounieh, Lebenon, pp. 945-949, 2000.

[2] Kurt, Z., Turkmen, H.I., and Karslıgıl, M.E., *“Linear Discriminant Analysis in Ottoman Alphabet Character Recognition”,* Proceedings of the European Computing Conference, Lecture Notes in Electrical Engineering, 2009, Vol. 28, No. 7, pp. 601-607, 2009

[3] Onat, A., Yildiz, F., and Gündüz, M. *“Ottoman Script Recognition Using Hidden Markov Model”*, IEEE Transection on Engineering Computing Technology, 2006, 14: pp. 71-73, 2006.

[4] Altingövde, İ.S., Şaykol, E., Ulusoy, Ö., Güdükbay, U., Çetin, A.E. and Göçmen, M. *“Content-Based Retrieval (CBR) System for Ottoman Archives”,* Proceedings of 14th Conference on Signal Processing and Communications Application in Turkish, IEEE, Piscataway, NJ, 2006.

[5] Başar, E., Kılıç, N., Görgel, P. and Uçan, B. *“Ottoman Character Recognition with Artificial Neural Networks and Development of Automatic Intelligent Translation System into Turkish”,* Electrical, Electronics, Computer, Biomedical Engineers, 12th. National Conference, Eskişehir, Turkey, 2007

[6] Ataer, E. and Duygulu, P. *“Matching Ottoman Words: An image retrieval approach to historical document indexing”* in Proc. of ACM Int. Conf. on Image and Video Retrieval, ACM, 2007, New York, pp. 341–347, 2007

[7] Yalnız, İ. Z., *“Integrated Segmentation And Recognition Of Connected Ottoman Script”*, Thesis (M.Sc), Bilkent University, 2008

[8] Yalnız, İ.Z. & Altingövde, İ.S. & Güdükbay, U. & Ulusoy, Ö. "Ottoman Archives Explorer: A Retrieval System for Digital Ottoman Archives" ACM Journal on Computing and Cultural Heritage, Vol. 2, No. 3, Article No. 8, 20 pages, 2009

[9] Andrews, W. G., İnan, M., Kebeli, S. and Waters, S., *"Rethinking The Transcription Of Ottoman Texts. The Case For Reversible Transcription"* , Turkish Studies International Periodical for the Languages, Literature and History of Turkish or Turkic Volume 5/2 spring 2010

[10] Spuler, B., *"Persian Historiography & Geography"*, Pustaka Nasional Pte Ltd, Singapur, January 2003.

[11] Hagopian, V.H., *"Ottoman-Turkish conversation-grammar; a practical method of learning the Ottoman-Turkish language"*, Heidelberg, J. Groos, New York, 1907.

[12] Borovikov, E. *"A survey of modern optical character recognition techniques"* AMS 2004

[13] Satti, D. A., *"Offline Urdu Nastaliq OCR for Printed Text using Analytical Approach"*, Thesis (Ph.D.), Quaid-i-Azam University, 2013

[14] Turkish Grand National Assembly, Library and Documentation Center [online], https://kutuphane.tbmm.gov.tr, [accessed 08.01.2016]

[15] Turkey Ministry of Culture and Tourism [online], http://ekitap.kulturturizm.gov.tr, [accessed 08.01.2016]

[16] Farabi Digital Library [online], http://e-library.ircica.org/ [accessed 08.01.2016]

[17] Modawi, O.M.A. *"Optical Character Recognition Software Evaluation"*, Thesis (M.Sc.), Sudan University of Science & Technology, 2005

[18] Heliński, M., Kmieciak, M., and Parkola, T. *"Report on the comparison of Tesseract and ABBYY FineReader OCR engines"* IMPACT: Improving Access to Text, 2012

[19] The Tesseract open source OCR engine[online], https://github.com/tesseract-ocr, [accessed 08.06.2015]

[20] Smith, R., Antonova, D. and Lee, D., *"Adapting the Tesseract Open Source OCR Engine for Multilingual OCR"*, Proceedings of the International Workshop on Multilingual OCR, 2009, Barcelona, Spain July 25, 2009.

[21] Smith, R., "*Hybrid Page Layout Analysis via Tab-Stop Detection*", Proceedings of the 10th international conference on document analysis and recognition, 2009, pp 241-245, July 26-29,2009

[22] Smith, R., *"An overview of the Tesseract OCR Engine"*, Proc 9th Int. Conf. on Document Analysis and Recognition, Curitiba, Brazil, IEEE, Sep 2007.

[23] Smith R. "*Tesseract OCR Engine*" [online]; URL: http://tesseract-ocr.googlecode.com/files/TesseractOSCON.pdf, [accessed 08.06.2015]

[24] Mishra, N., Patvardhan, C., Lakshmi, C. V., and Singh, S. *"Shirorekha Chopping Integrated Tesseract OCR Engine for Enhanced Hindi Language Recognition"* International Journal of Computer Applications 39(6): pp 19-23, February 2012. Published by Foundation of Computer Science, New York, USA.2012

[25] Leptonica image processing and analysis library [Online], http://www.leptonica.com. [accessed 07.10.2015]

[26] Shapiro, L.G. and Stockman, G. C., *"Computer Vision"*, Prentice Hall, 2001.

[27] Duda, R. O. and Hart, P.E., *"Use of the Hough Transformation to Detect Lines and Curves in Pictures"*, Communication of ACM, - January, 1972. - pp. 11–15.

[28] Touj, S., Amara, N.E.B. and Amiri H., *"Generalized Hough Transform for Arabic Optical Character Recognition"*, Proc 7th. The International Conference on Document Analysis and Recognition, 2003

[29] Imagemagick image processing library[Online], http://www.imagemagick.org [accessed 07.09.2015]

[30] The Ottoman Archive Project [Online], http://otap.bilkent.edu.tr/, http://courses.washington.edu/otap/ [accessed 07.08.2015]

[31] Zemberek Natural Language Processing library [Online], https://github.com/ahmetaa/zemberek-nlp, [accessed 09.12.2015]

[32] Redhouse, J.W., *"Turkish and English Lexicon"*, New Edition Constantinople, 1890

[33] Kanar, M., *"Arap Harfli Alfabetik Osmanlı Türkçesi Sözlüğü"*, Say Yayınları, 2012.

APPENDIX A

OPTICAL CHARACTER RECOGNITION ON THE FIRST PAGE OF THE NUTUK

The following text is the result of optical character recognition with Tesseract V3 on the first page of the Nutuk.

سنهس مايسنك ثمتدكولى صامسوتح جيقدم ٠ وضعيث

نانلى دولتنك،إدأخل يرلنديني غروب ، حرب نموميده مغلوب اولمرء

عنانر اردوس هي طرفدم زدملنر ، شرائط آغير ، بر متابهلامه

امثلاةش . يويوك حربك اوزون سنهلرى ظرقندمم ملت بررغون

فقير بر مالده و ملت وءلكتى حرب هومديه سوق ايدر ، كندى حيارى

اندي -يلجه

هوشهرلي ، مملكتدن فرار افيلر٠ سلطت وخلافت موقعنى اشغال

ابدنوحيدالدين، متردى،

شخصنى ويانلز تختنى تأمين ايدوهجذض تخيل ايتدكى دفى ندبيرز

آراش}درنقدع ا داماد

فريد باشاك رياسقندمك مين ؛ذجز ، حينيقسز ، جبين ، بالكر بادشاهك

ارادته تابع

.و او٠كظه برار شضلرتر وقاع ايدهبهجك هى هانك بروضعينه راضى

...اردوك أ لندن اسلحه وجبخانهس آلفير وآلققدع

ائتلاف دولتلرى ، متار ٨ اتكامنه رعايته لزومكورميورلر ع برر زسي« ايا

، ابهنى

دوتح»لرى و عكرلرى اسثنبيلدع آطنه ولاته ، فرانسزلر بم اورفه ،

عرعس ك عينتاب ،

العهايزلر طرفندن اشغال ايدلمش ا آنطاليه وقونيهدء ، اينااليان قطعات

عكريهس بم هسزضن

و صامسوندم ايريز عكرلرى زونيور ، هدطرفدم ، اجنى ضابط و مأمورلرى

و خصوس

آدملرى فعاليتدم و ترايت ، ميدأكلام قبول اسندكمز تهربمدن درم،ن اول ،

هيا مايس

دم ائتلاف دولتلريخك موافقته سونان اردوس ازمير. اخراج إديلنور.

The correct result must be the following text. The green characters have been recognized correctly but the red characters have not been recognized after optical character recognition operation in this text. The total number of characters is 958. The number of correctly founded characters is 727 and incorrectly founded characters are 231. According to these results recognition rate is 75.88%.

١٣٣٥ سنه سي مايسنك ١٩ نجي كوني صامسونه جيقدم ٠ وضعيت

: و منظرهٔ عموميه

، عثمانلى دولتنك داخل بولنديغى غروپ، حرب عموميده مغلوب اولمش

عسمنلي اردوسي هر طرفده زدهلنمش ، شرائطي آغير ، بر متاركه نامه

امضالانمش . بويوك حربك اوزون سنهلرى ظرفنده , ملت بورغون

فقير بر حالده ملت و مملكتى حرب عمومي يه سوق ايدنلر ، كندى
حياتلرى انديشه سنه

دوشهرك ، مملكتدن فرار ايتمشلر ٠ سلطنت و خلافت موقعنى اشغال ايدن
وحيدالدين، متردى،

شخصنى و يالكز تختنى تأمين ايدوه بيله جكني تخيل ايتديكى دنى
تدبيرلر آراشديرمقده . داماد

فريد پاشانك رياستنده كي قابينه ؛ عاجز ، حيثيتسز ، جبين ، يالكز
پادشاهك ارادهسنه تابع

و اونكله برابر شخصلريني وقايه ايده بيله جك هر هانكي بروضعيته
. راضى

... اردونك ألندن اسلحه و جبخانهسي آلنمش و آلنمقده

ائتلاف دولتلرى ، متاركه احكامنه رعايته لزوم كورميورلر . برر وسيله
ايله ، ائتلاف

دونانمالرى وعسكرلرى استانبولده . آطنه ولايتي , فرانسزلر ; اورفه ،
، مرعش , عينتاب

انكليزلر طرفندن اشغال ايدلمش . آنطاليه وقونيهده ، ايتاليان قطعات عسكريهسي ; مرزيفون

و صامسونده انكليز عسكرلرى بولونيور . هرطرفده ، اجنبى ضابط و مأمورلرى و خصوصي

آدملرى فعاليتده .نهايت ، مبدأ كلام قبول ايتديكمز تاريخدن درت كون اول , ١٥ مايس

٣٣٥ ده ائتلاف دولتلرينك موافقتيله يونان اردوسي ازميره اخراج ايديليور.

APPENDIX B

OPTICAL CHARACTER RECOGNITION ON THE SECOND PAGE OF THE NUTUK

The following text is the result of optical character recognition with Tesseract V3 on the second page of the Nutuk.

برندن هيقه ، مملكتك ص طرفنده م عناصر خرسقيانيه غنى ، جلى ، خصوس أمل

ومقصدلرينك امين استحصاله ، دولتك بر ان اول ، جهكمسنه صرف مسار ابدكورلر ٠

بالاسخرم الدءايديلنمونيق معلومات وو»قايه تأيدايندىك اسبانيول روهطريقخانهسندء

قضض ايدن ط ماورو ميرا » عيئنى ، { ونيقه ، ٩ نا ولاينق داخلنده جتهلر تقليل
وادار.

اينكم متينغار و،رواأادالر بابدبرمقهمعغول . يولانصليب احمرى، رسى
مهاجرينةوميسيونديم

ماورى ميرا ا هيئنك تسهيل مساعيسنه خادم ٠ ط ماورى ميرا » هيئتى طرطدن ادار. »
اولان

روم كينبلرينك اينس قفكيلاتيرى ، كرس ياشتى متجاوزكنجلرهه داخل اولمق اوزر. ض
بردع اكل اولونيور ٠

ارهنى ٠لطريق (زاو٠ن) اقندىدم ، ط ماورى هيرا » هيئنيه مم نكر اولهرق باليشيور
٠

ارمنى حاضرلنددء تيمأ روم حاضرلنى كني ايلريليور م

طرنون ، سامسون وبوتون قرع نكز ساحللرنده تنل ايش واستاتهولدمك سكزم

(معط ق بونتوس جعينى » سهولته وموفقيتله باليشيور ٠ لا ونيقه ب ٢

ق وضعيتك دهشت ووخامنى قارشوسنده ، هى بردع ، هدمنطقهەء برطافي نواد طرفندن

مقابل خلاس بارمإى هوشونولمكه باشلانمر ايدى ٠ بودوشو» ايه آلان تشيثاهت ، برطافي

تشكللر دوغوردى ع مثلا ب ادرنه وحواليسندم ، تراكا هاشم ايلى ا عنوانه برجعنث واردى ق

شرقد٠ بم { وليقه ، ٣ تم ارضرومده والعنيزد٠ ل وثبقه ن ف تم مركز عموميس استاتهولدم

اولمقاوزرع ط ولايات شرقيه مدافعة حقوق عليه » جعينى تيكيد ايدلمشدى ع طرغوند٠ ومحافظة

حقوق . ،مندب بدجعيت موجود اولدينلإ درسعادتحمدم ، طرزون وحواليس عدمهمكزيع

جعيتى ، واردى ع يخجعيت عمكزينك كؤندردتي مرخص(م اوف قضاسه لازستان لواس

(داخلنده شعبهلر آجيلمعدى . { وبينه ئ ه تم و ل وبمهته ، ٦

ازمير(اشغال او»جغنه هار مايسك اون اوجندن برى فعلى امارءلر كرن ازميرد٠ ل٠عني

كنج وطنيرورلر، آيك لا مدكبس ، بوألع وضعيث حقنده مداولهء افكار اثير واسواق

ال.هكهلددكنه شهه فالمان و،ن اشغالنك الحاقه ئحهلقسنه ماني اولمق اسا–نده متفق فاكر

صا ع ال. ٠ عم ا

و ط رد الحاق. ٠رنسينى اورتحيه آنمشيردرم عيندكجهد٠ بومقصدك تشمرأميناكون ازميرد٠

يهودى ماشاطلغنه طو،لأويلن خلق طرفندن برمتثغ بالمشهد. ابدتسكون صباحا يران ك لربك ريختيمدم كورولمسبه نوقشهح امير إديلن درجهدء تأمين مقصد ابادر ٠ «ؤ بوجعيتلرلي مقصد قطكللرى وهدف سياسبؤى حقند٠ مختصرأ اعلى معلومات اثك موافق اولور مطالعهسندوم.

تراك باشا ايلى ا جعيكك رؤساسندن لعضيلربه دها استانيولدء اكنكروشمشايدممع »

The correct result must be the following text. The green characters have been recognized correctly but the red characters have not been recognized after optical character recognition operation in this text. The total number of characters is 1804 for this text. The number of correctly founded character is 1396 and incorrectly founded character is 409. According to these results recognition rate is 77.38% for this text.

بوندن بشقه ، مملكتك هر طرفنده , عناصر خرستيانيه خفى ، جلى ، خصوصي أمل ومقصدلرينك تأمين استحصالنه ، دولتك بر آن اول ، چوكمسنه صرف مساعي ايدييورلر ٠

بالاخره الده ايديلن موثوق معلومات ووثائق ايله تأيد ايتدي, كه استانبول روم پطريقخانه سنده

تشكل ايدن « ماوي ميرا » هيئتى ،[وثيقه : ١] ولايتلر داخلنده چته لر تشكيل و اداره

ايتمك , متينغلر وپروپاغاندالر ياپديرمقله مشغول . يونان صليب احمرى، رسمى ; مهاجرين قوميسيوني

« ماورى ميرا » هيئتنك تسهيل مساعيسنه خادم ٠ « ماورى ميرا » هيئتى طرفندن اداره اولنان

روم مكتبلرينك ايزجي تشكيلاتلرى ، يكرمي ياشنى متجاوز كنجلرده داخل اولمق اوزره هر

يرده اكمال اولونيور ٠

ارمنى پطريقي (زاوهن) افنديده ، « ماورى ميرا » هيئتيله هم فكر اولهرق چاليشيور ٠

ارمنى حاضرلغيده تماماً روم حاضرلغى كبي ايلريليور .

طربزون ، صامسون وبوتون قره دكز ساحللرنده تشكل ايتمش و استانبولده كي مركزه

مربوط « پونتوس جمعيتى » سهولتله و موفقيتله چاليشيور .] وثيقه : ٢]

وضعيتك دهشت و وخامتى قارشوسنده ، هر يرده ، هرمنطقهده برطاقم ذوات طرفندن

مقابل خلاص چارهلرى دوشونولمكه باشلانمش ايدى . بو دوشونجه ايله آلنان تشبثات

، برطاقم

تشكللر دوغوردى . مثلا : ادرنه وحواليسنده « تراكيا پاشاايلى » عنوانيله بر جمعيت .

واردى

شرقده ; [وثيقه : ٣] ارضرومده والعزيزده [وثيقه : ٤] مركز عموميسي استانبولده

اولمق اوزره « ولايات شرقيه مدافعهٔ حقوق مليه » جمعيتى تشكيل ايدلمشدى .

طربزون ده محافظهٔ

حقوق » نامنده بر جمعيت موجود اولديغي كبي درسعادتده ده ، طربزون وحواليسي عدم

مركزيت

جمعيتى ، واردى . بوجمعيت مركزينك كونددريكي مرخصلرله , اوف قضاسيله لازستان

لواسي

داخلنده شعبهلر آچيلمشدى . [وثيقه : ٥] و [وثيقه : ٦]

ازميرك اشغال اولنه جغنه دائر مايسك اون اوچندن برى فعلى اماره لر كورن ازميرده

بعض

١٥ نجي كيجهسي ، بو أليم وضعيت حقنده مداولههٔ افكار /كنج وطنپرورلر، آيك ١٤

ايلمثلر وامرواقع

حالنه كلديكنه شبهه قالميان يونان اشغالنك الحاقله نتيجهلنمسنه مانع اولمق اساسنده

متفق قالمشلر

و « رد الحاق » پرنسبپنی اورته یه آتمشلردر. عيني كيجهده بومقصدك تشميلني

تأمين ايچون ازميرده

يهودی ماشاطلغنه طوپلانهبيلن خلق طرفندن بر متينغ ياپيلمشسهده ابرتسي كون

صباحلين يونان

عسكرلر ينك ريختيمده كورولمسيله بوتشبث اميد ايديلن درجهده تأمين مقصد

ايدهممهشدر ٠

بوجمعيتلرك مقصد تشكللری و هدف سياسيلری حقنده مختصراً اعطای معلومات ايلمك

موافق اولور مطالعهسندهيم .

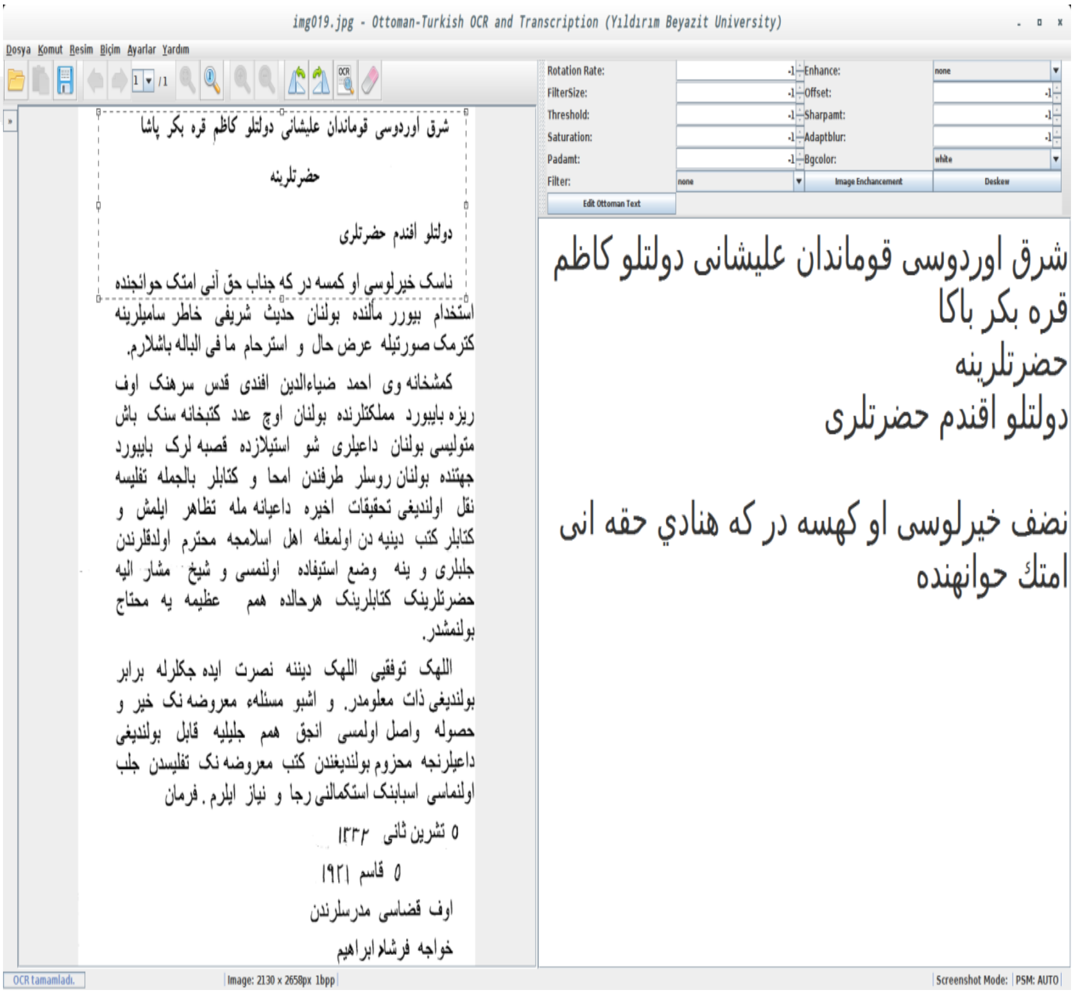
« تراكيا پاشاايلی » جمعيتنك رؤسا سندن بعضيلريله دها استانبولده ايكن كوروشمش

ايدم

APPENDIX C

MAIN GRAPHICAL USER INTERFACE

Figure C.1 shows the main graphical user interface (GUI). The user can open the Ottoman-Turkish documents or images. After opening a File, optical character recognition operation and some image processing operations could be done by this interface.


img019.jpg - Ottoman-Turkish OCR and Transcription (Yıldırım Beyazit University)
Dosya Komut Resim Biçim Ayarlar Yardım
1 /1
OCR
Edit Ottoman Text
Image Enchancement
Deskew

Rotation Rate: -1
FilterSize: -1
Threshold: -1
Saturation: -1
Padamt: -1
Filter: none
Enhance: none
Offset: -1
Sharpamt: -1
Adaptblur: -1
Bgcolor: white

شرق اوردوسی قوماندان علیشانی دولتلو کاظم قره بکر پاشا

حضرتلرینه

دولتلو افندم حضرتلری

ناسک خیرلوسی او کمسه در که جناب حق آنی امتک حوائجنده استخدام بیورر مآلنده بولنان حدیث شریفی خاطر ساميلرينه کترمک صورتیله عرض حال و استرحام ما فی الباله باشلارم.

کمشخانه وی احمد ضیاءالدین افندی قدس سرهنک اوف ریزه بایبورد مملکتلرنده بولنان اوچ عدد کتبخانه سنک باش متولیسی بولنان داعیلری شو استیلازده قصبه لرک بایبورد جهتنده بولنان روسلر طرفندن امحا و کتابلر بالجمله تفلیسه نقل اولندیغی تحقیقات اخیره داعیانه مله تظاهر ایلمش و کتابلر کتب دینیه دن اولمغله اهل اسلامجه محترم اولدقلرندن جلبلری و ینه وضع استیفاده اولنمسی و شیخ مشار الیه حضرتلرینک کتابلرینک هرحالده همم عظیمه یه محتاج بولنمشدر.

اللهک توفقیی اللهک دیننه نصرت ایده جکلرله برابر بولندیغی ذات معلومدر. و اشبو مسئلهء معروضه نک خیر و حصوله واصل اولمسی انجق همم جلیلیه قابل بولندیغی داعیلرنجه محزوم بولندیغندن کتب معروضه نک تفلیسدن جلب اولنماسی اسبابنک استکمالنی رجا و نیاز ایلرم . فرمان

٥ تشرين ثانی ١٣٣٢

٥ قاسم ١٩٢١

اوف قضاسی مدرسلرندن

خواجه فرشاد ابراهیم

شرق اوردوسی قوماندان علیشانی دولتلو کاظم
قره بکر باکا
حضرتلرینه
دولتلو اقندم حضرتلری

نضف خیرلوسی او کهسه در که هنادي حقه انی
امتك حوانهنده

OCR tamamladı.

Image: 2130 x 2658px 1bpp

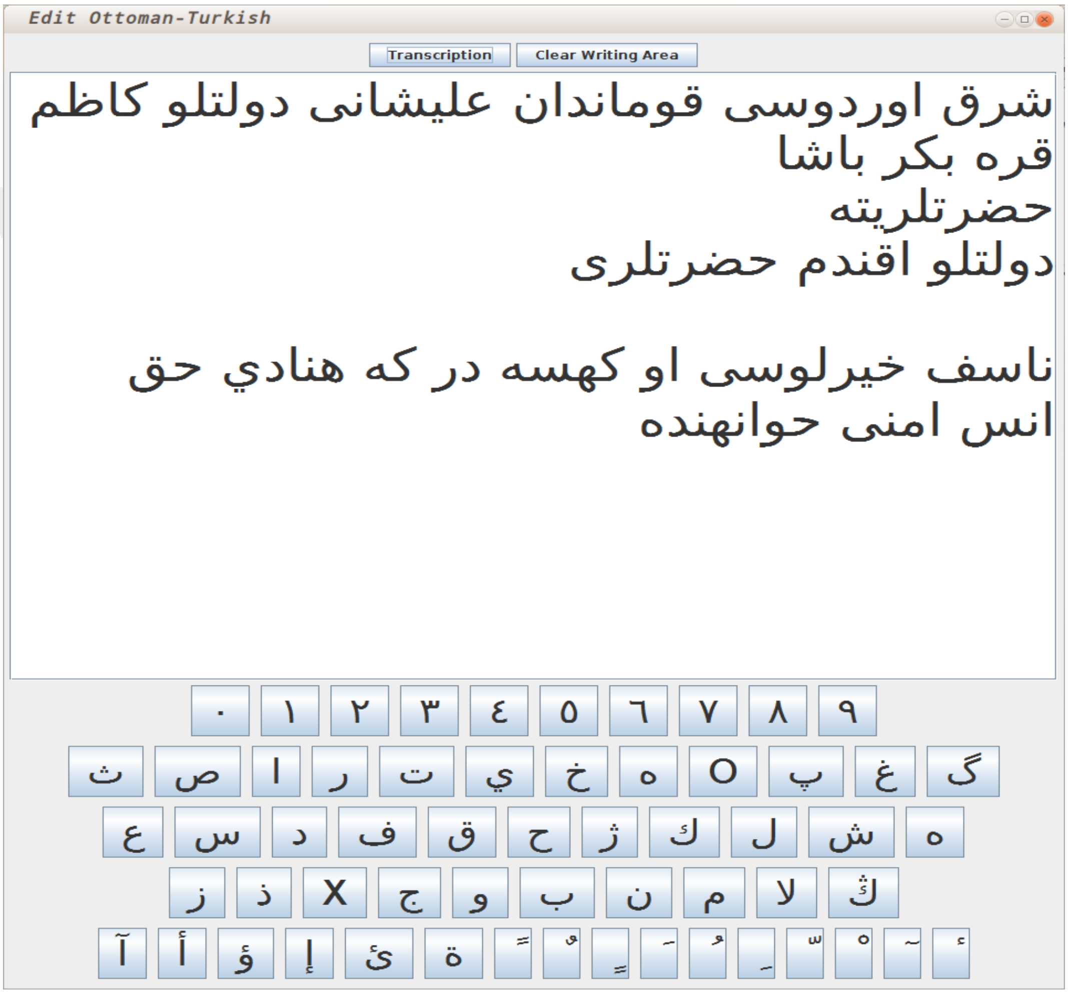
Screenshot Mode: | PSM: AUTO

Figure C.1 Main graphical user interface (GUI).

APPENDIX D

OTTOMAN-TURKİSH KEYBOARD TOOL

Figure D.1 shows Ottoman-Turkish keyboard tool. Ottoman-Turkish scripts could be written by this keyboard with Ottoman-Turkish alphabet.



شرق اوردوسى قوماندان عليشانى دولتلو كاظم
قره بكر باشا
حضرتلريته
دولتلو اقندم حضرتلرى

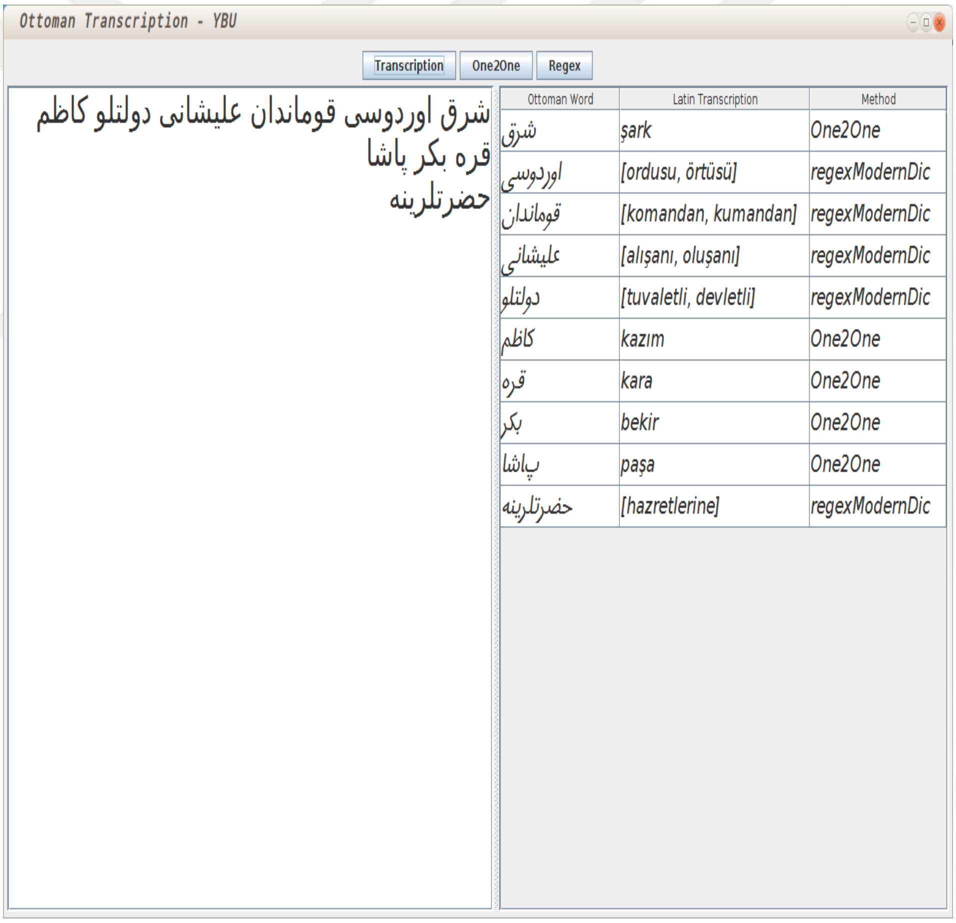
ناسف خيرلوسى او كهسه در كه هنادي حق
انس امنى حوانهنده

Figure D.1 Ottoman-Turkish keyboard

APPENDIX E

TRANSCRIPTION TOOL

Figure E.1 shows the transcription tool. Left panel shows Ottoman-Turkish text after editing with the keyboard. Right panel shows a table which shows the Latin words after transcription of the each Ottoman-Turkish words. Transcription button is combined one to one method and transcription using regular expressions. One2One button is performed for one to one method. Regex button performs transcription using regular expressions method.



شرق اوردوسی قوماندان علیشانی دولتلو کاظم
قره بکر پاشا
حضرتلرینه

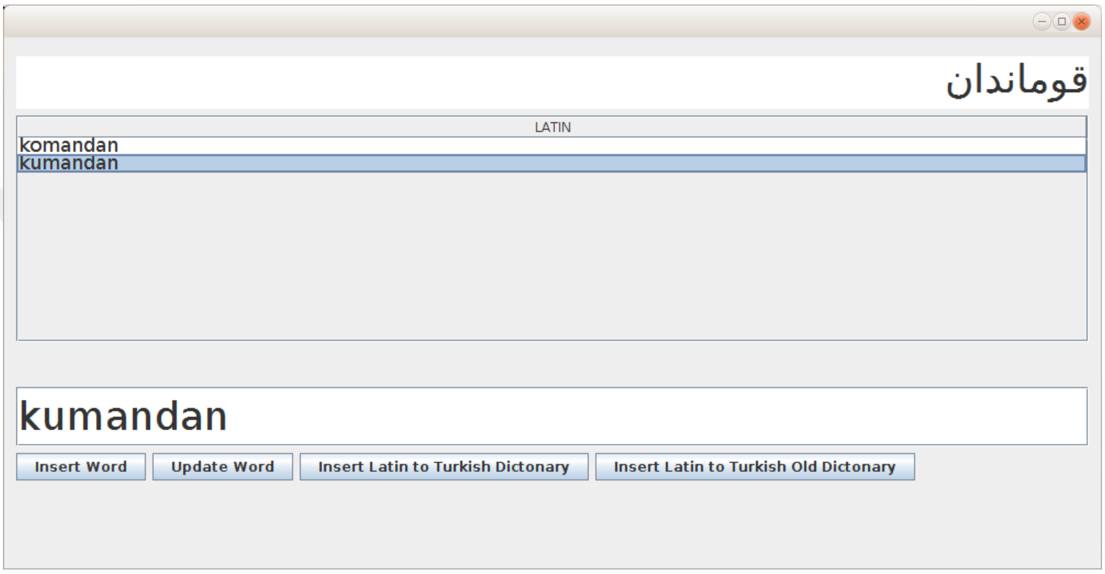
Ottoman Word	Latin Transcription	Method
شرق	şark	One2One
اوردوسی	[ordusu, örtüsü]	regexModernDic
قوماندان	[komandan, kumandan]	regexModernDic
علیشانی	[alışanı, oluşanı]	regexModernDic
دولتلو	[tuvaletli, devletli]	regexModernDic
کاظم	kazım	One2One
قره	kara	One2One
بکر	bekir	One2One
پاشا	paşa	One2One
حضرتلرینه	[hazretlerine]	regexModernDic

Figure E.1 Transcription tool

APPENDIX F

ENRICHMENT DATABASE TOOL

Figure F.1 shows the enrichment database tool. Dictionary database can be improved by using this tool.



قوماندان

LATIN
komandan
kumandan

kumandan

Insert Word
Update Word
Insert Latin to Turkish Dictonary
Insert Latin to Turkish Old Dictonary

Figure F.1 Enrichment database tool

RESUME

MUSTAFA DOĞRU

Address : İstanbul Yolu 13. Km. Hava
Müzesi Karşısı Şaşmaz 06630

E-mail : mustafadogru@adalet.gov.tr

PERSONEL INFORMATION

Marital Status : Married

Date of Birth : 15.06.1983

EDUCATIONAL BACKGROUND

2001 – 2006, B.S in Computer Engineering, Erciyes University, Kayseri/Turkey

WORK EXPERIENCE

2006 – Present, IT Specialist Ministry of Justice