DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

TEXT CONVERSION SYSTEM BETWEEN TURKIC DIALECTS

by
Emel ALKIM

September, 2013
İZMİR

DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

TEXT CONVERSION SYSTEM BETWEEN TURKIC DIALECTS

**A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Computer Engineering**

by
Emel ALKIM

September, 2013
İZMİR

Ph.D. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"TEXT CONVERSION SYSTEM BETWEEN TURKIC DIALECTS"** completed by **EMEL ALKIM** under supervision of **PROF.DR. YALÇIN ÇEBİ** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

....
Prof.Dr. Yalçın ÇEBİ

Supervisor

....
Yrd. Doç Dr. Gökhan DALKILIÇ

Thesis Committee Member

....
Prof. Dr. Gürer GÜLSEVİN

Thesis Committee Member

....
Yrd. Doç. Dr. Şen Çakır

Examining Committee Member

....
Bekir Tanaç DİNCER

Examining Committee Member

Prof.Dr. Ayşe OKUR
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to thank my advisor Professor Dr. Yalçın ÇEBİ for his support during my studies; my thesis tracking committee members Assistant Professor Dr. Gökhan DALKILIÇ and Professor Dr. Gürer GÜLSEVİN, and also my friends and colleagues in Dokuz Eylül University Natural Language Processing Research Group; Assistant Professor Dr. Özlem AKTAŞ, Research Assistant Çağdaş Can BİRANT, Assistant Professor Dr. Özden FİDAN and Instructor Dr. Özgün KOŞANER for their contribution to this study. I would also like to thank to Turkish Linguistic Association (Türk Dil Kurumu, TDK) for their support during the development of the resources for this study.

This research was supported under Contract Number: 2009_KB_FEN_11 by Dokuz Eylül University Scientific Research Projects (Bilimsel Araştırma Projeleri, BAP) Coordination Unit.

I have special thanks to my family and friends, Research Assistant Mete Uğur AKDOĞAN and Dr. İbrahim ARPALIYİĞİT for their support and patience during the development and writing of the thesis.

Emel ALKIM

TEXT CONVERSION SYSTEM BETWEEN TURKIC DIALECTS

ABSTRACT

Turkic communities come from a common culture; however the interaction with other communities over years caused diversion especially in written language. A system which can automatically translate documents written in different Turkic languages will be an important step towards eliminating the disunity of Turkic communities on written work of art over past ninety years and obtaining fusion of Turkic communities.

In this study, a rule-based and semi-supervised machine translation system (MT-Turk), which is designed for closely related Turkic languages and implemented on Turkish, Kirghiz and Kazan Tatar, is presented. MT-Turk is an extensible bidirectional translation infrastructure in which new Turkic dialects can be added by just adding the lexicon of roots/stems, suffixes, and the rules. Furthermore, it is open to extension by suggestion. In order to form a multilingual machine translation infrastructure, two subsets of rule-based approach, the interlingual machine translation approach and transfer-based approach were used in combination to achieve extensibility and interoperability.

The success of the translation process was evaluated using both BLEU and NIST metrics. The evaluated scores were between 5.04 and 15.12 for BLEU, between 3.12 and 4.64 for NIST in unsupervised translation and between 7.20 and 21.71 for BLEU, between 3.52 and 4.77 for NIST in semi-supervised translation for various language pairs and translation directions. Depending on these results, it was seen that the efficiency of the translation process is extremely dependent on the size of the lexicon and the rule base.

Keywords: Machine translation, natural language processing, rule-based machine translation, multi-word expressions, Turkic dialects, Turkish, Kirghiz, Kazan Tatar

TÜRK LEHÇELERİ ARASINDA ÇEVİRİ SİSTEMİ

ÖZ

Türk dilleri aynı kökenden gelmelerine rağmen yıllar içinde farklı topluluklarla olan etkileşimler nedeniyle farklılaşmışlardır. Farklı lehçelerde yazılmış metinlerin otomatik çevirisini yapan bir sistem, Türk topluluklarının iletişiminde ve kaynaşmalarında bir engel olan bu farklılaşmanın giderilmesinde ve kültür birliğinin geliştirilmesinde önemli bir adım olacaktır.

Bu çalışmada, akraba diller olan Türk dilleri için geliştirilip; Türkiye Türkçesi, Kırgız Türkçesi ve Tatar (Kazan) Türkçesi üzerinde uygulanan kural tabanlı ve yarı eğitmenli bir bilgisayarlı otomatik çeviri sistemi tanıtılmaktadır. MT-Turk, sadece sözlük, ek ve kurallar tanımlayarak yeni bir lehçe eklenmesi ile genişletilebilen iki yönlü bir çeviri altyapısıdır. Ayrıca, öneriler yardımıyla da genişletilmeye açıktır. Çok dilli bir bilgisayarlı çeviri altyapısı hazırlamak için kural tabanlı yaklaşımın iki alt alanı olan aktarım temelli ve interlingua temelli yaklaşımlar, genişletilebilirliği ve birlikte çalışabilirliği sağlamak amacıyla birlikte kullanılmıştır.

Çeviri işleminin başarısı BLEU ve NIST ölçekleri kullanarak değerlendirilmiştir. Ölçülen değerler farklı dil çiftleri ve çeviri yönleri için gözetimsiz çeviride BLEU 5,04 ve 15,12 arasında, NIST 3,12 ve 4,64 arasında, gözetimli çeviride ise BLEU 7,20 ve 21,71 arasında, 3,52 ve 4,77 arasında değişmektedir. Bu sonuçlara dayanarak, çeviri işleminin etkinliğinin sözlük ve kural tabanının boyutuna son derece bağlı olduğu gözlenmiştir.

Anahtar Kelimeler: Bilgisayarlı çeviri, doğal dil işleme, kural tabanı bilgisayarlı çeviri, sözcük öbekleri, Türk lehçeleri, Türkiye Türkçesi, Kırgız Türkçesi, Tatar (Kazan) Türkçesi

CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

Communication has become the most crucial topic in the era of globalization and the Internet. Moreover, there is a huge amount of information in the Internet in various languages which awaits exploring. Machine Translation (MT) is the method to achieve this connectivity and overcome the language barrier.

However, MT is an hard task in which collaboration of several fields are required (Hovy, 2001a; Şenkal, 2000). The most important problems of machine translation are the different structures of languages, different cultures and the ambiguities of natural language. Although Turkic dialects are generally similar in their structure and even consist of common stems, such as "at: *horse*" which has the same representation in Turkish, Azerbaijani, Bashkir, Kazakh, Kirghiz, Uzbek, Tatar, Turkmen and Uyghur (Ercilasun, 1992; Uğurlu, 2004), people cannot understand each other in different dialects. Turkic language family, which is consisted of 40 languages that are closely related to each other, is spread over a large geographical area ranging from Eastern Europe and the Mediterranean to northeastern Siberia and western China (SOROSORO, 2009); and is spoken by approximately 180 million people as mother language (SIL International, 2013). Hence, for enhancing the economic and trade relations between Turkic Republics there is a need for a common way of understanding each other and translating documents to other Turkic dialects easily.

Information technology (IT) applications are important instruments for the purpose of constructing this common way, therefore in 2005 "Turkic World Computer Assisted Linguistics Working Group" (TDK, 2005) was founded with the aim of conducting studies on Turkic Languages with IT technology, by constructing common dictionaries and conducting grammar studies with IT technology. The working group was founded with the initiative of "Turkic Republics Information Technologies Working Group" (TBD (Informatics Association of Turkey), 2000) within Undersecretariat of Foreign Trade and by the cooperation of Turkish Linguistic Association (TDK) and Turkish Informatics Association (TBD).

1.1 Aim of Thesis

The need of machine translation between Turkic dialects is crucial and much easily attainable than with other languages like English as they are closely related. However current studies either focus on a particular language pair or work only on one direction (from Turkic dialects to Turkish).

The aim of this study is to build an extensible infrastructure to translate from one Turkic dialect to another. The MT-Turk, infrastructure is extensible in two dimensions; the number of dialects supported and quality of translation. The number of dialects supported can be increased by adding new dialects to the infrastructure, in other words by supplying lexicon and rules with user friendly interfaces. Additionally, the quality and success of the translation can be extended by the suggestions made by users of the infrastructure. Currently, MT-Turk supports Turkish, Kirghiz and Kazan Tatar.

1.2 Thesis Organization

This thesis is divided into six chapters. In Chapter 2, machine translation is described with history, approaches and example studies including a subsection for machine translation between closely related languages.

The infrastructure of MT-Turk is described in detail including rule formats, the algorithms employing the components of MT-Turk and the database model in Chapter 3.

The case study of this thesis is a subset of Turkic dialects, Turkish, Kirghiz and Kazan Tatar. The grammatical characteristics of these dialects are described in Chapter 4.

In Chapter 5, the information about the case study and resources are given in addition to information about the evaluation using two metrics, BLEU and NIST. Finally, the conclusion is given in Chapter 6 including a brief summary and results of the thesis.

CHAPTER TWO
MACHINE TRANSLATION

Machine translation (MT) is an interdisciplinary study area that requires collaborated study of linguistics, computer science, artificial intelligence, translation theory, computational algorithms, cognitive science, study of human-computer interaction and occasionally anthropology (Hovy, 2001b; Şenkal, 2000).

2.1 History of Machine Translation

Machine translation is one of the first non-numerical applications on computers (Hutchins, 1986). Even before internet era, machine translation was a popular area of interest; as a matter of fact, the idea of machine translation dates back to 17th century when René Descartes proposed a universal language (Ulitkin, 2011) and continued with different proposals and patents. Nevertheless, a memorandum by Warren Weaver (Weaver, 1949) is considered as the initiation of machine translation studies.

In 1954, the first application of machine translation, public demonstration of a machine translation system is performed with the Georgetown-IBM experiment (J. Hutchins, 2004; IBM, 1954). Although the experiment consisted of only 250 words and performed translation of 49 carefully selected Russian sentences to English, it encouraged the studies on machine translation. The studies on machine translation were very popular for more than a decade and funded by the government with the public effect of the experiment and the cold war, especially in the United States and Soviet Union.

Unfortunately, first with the report prepared by Bar-Hillel for United States Government at 1960 (Bar-Hillel, 1960) and then the ALPAC report at 1966 (ALPAC, 1966), full-automatic high quality machine translation (FAHQMT) was said to be *not attainable in an open domain*. Consequently, the studies on machine translation slowed down till 80s when improvements in software and hardware technologies led to more success (Cieślak, 2011).

The first approach to machine translation or classical machine translation is rule-based machine translation (W John Hutchins, 1986). Corpus based approaches were introduced later, with the emergence of the internet in 90s and availability of the large amounts of online text that became accessible (Su & Chang, 1992). Lately, two approaches are combined in hybrid approach to achieve better translation systems (Chen & Chen, 1996; Thurmair, 2005; Xuan, Li, & Tang, 2012) .

2.2 Natural Language Processing

Machine Translation is a sub-field of Natural Language Processing (NLP) and is achieved using seven levels of NLP. NLP is stated in Liddy (2001) as “a theoretically motivated range of computational techniques for analysing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.” (Liddy, 2001)

Each level of NLP is responsible for analysis and extraction of linguistically meaningful units at different levels. Implementing and using all levels are not obligatory, however; more successful translation is achieved with deeper analysis.

- ***Phonology***

Phonology level is responsible for the interpretation of sounds within and across words. The level manages three types of rules: phonetic rules, phonemic rules and prosodic rules. Phonetic rules define the constraints on how sounds within words are combined whereas phonemic rules define the variations of pronunciation when words are spoken together. Lastly, prosodic rules are used to define the fluctuation in stress and intonation across a sentence (Liddy, 2001).

- ***Morphology***

Morphological level is responsible for the study of words. In this stage, words are analysed and their morphemes, the smallest meaningful units of a word, are extracted (Liddy, 2001).

- ***Lexical***

Lexical level is responsible for interpreting the meaning of individual words, by mapping the most probable part-of-speech or sense, especially for the words having only one possible sense. The lexical level may require and use a lexicon (Liddy, 2001).

- ***Syntactic***

Syntactic level is the study of how the words in the sentence are combined to uncover the grammatical structure (Liddy, 2001). The sequences of words are transformed into syntax trees using grammatical rules and constraints of the natural language.

- ***Semantic***

Semantic analysis is the study of meanings of sentences. In semantic analysis, the sentence structure, in other words syntax trees, must be assigned meaning by focusing on the interactions among word-level meanings in the sentence. The semantic disambiguation of words with multiple senses is also a part of this level (Liddy, 2001).

- ***Discourse***

Discourse integration is the study of connecting the sentences in a context as the meaning of a sentence may depend on other sentences in that context. Thus, the discourse level is responsible for the text as a whole rather than a sentence (Liddy, 2001).

- ***Pragmatic***

Pragmatics is the study of assigning the real meaning to the text depending on how language is used. The goal of this level is to explain how extra meaning is read into texts without actually being encoded in them and it requires world knowledge, understanding of intentions, plans and goals (Liddy, 2001).

2.3 Rule-Based (Classical) Machine Translation

Rule-based machine translation is achieved by the use of linguistic data and rules for translation (Douglas, Balkan, Meijer, Humphreys, & Sadler, 1993; W John Hutchins, 1986). It is also called Classical Machine Translation as it is the traditional and the first developed approach to machine translation.

In the remainder of the subsection, different approaches to rule-based machine translation is described briefly and brief information about multi-word expressions is given. Lastly, some applications developed using rule-based machine translation methodologies are listed.

2.3.1 Types of Rule-based Machine Translation

Rule-based machine translation is categorized in three types according to the depth of the process (both for analysis and generation) and whether a language-independent representation of meaning is attempted or not (Dorr, Hovy, & Levin, 2006):

- Direct Approach
- Transfer Approach
- Interlingua Approach

Vauquois triangle (Vauquois, 1968) which is illustrated in Figure 2.1 is a representation used for visualizing the approaches to rule-based machine translation. Each layer in the triangle constitutes to a layer in linguistic analysis layers.

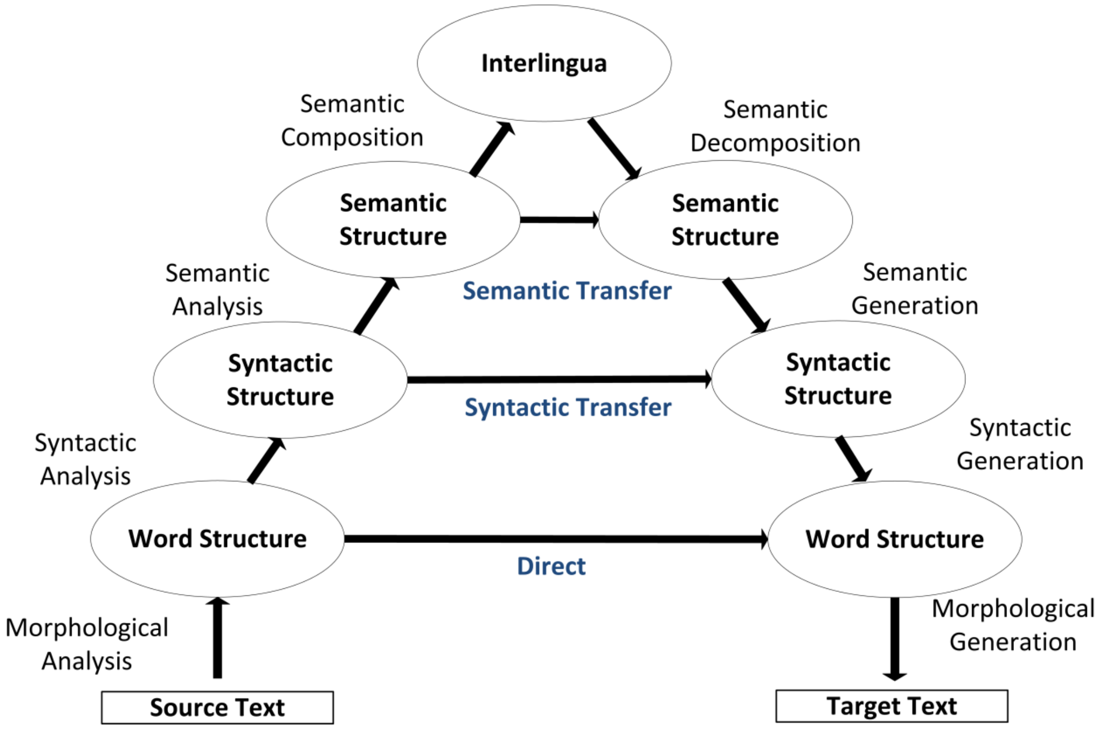


Figure 2.1 Vauquois triangle (Vauquois, 1968)

- ***Direct Approach***

In direct approach, each word is stored in the dictionary and the translation is achieved by a word-by-word replacement process which results in the need of huge dictionaries. In this approach, source texts are not analysed more than needed for generating texts in the target language (W J Hutchins, 1994; Şenkal, 2000).

- ***Transfer Approach***

In transfer approach, the input is analysed in the source language, then the transfer is achieved by a set of transfer rules and the output text is generated in the target language. These source-to-target transfer programs are developed with analysis and generation modules which are specific for each language pair (W J Hutchins, 1994).

- ***Interlingua Approach***

In interlingua approach, an interlingua, which is a language-neutral representation, is produced as the result of the analysis and used as the starting point for the generation. The translation is done in two steps; from the source language to the interlingua and from the interlingua into the target language, as the interlingua is language independent. In a multilingual configuration, programs for analysis are independent from programs for generation, and any analysis program can be used together with any generation program (W J Hutchins, 1994; Şenkal, 2000).

2.3.2 Multi-Word Expressions

Multi-Word Expressions (MWE) are defined as structures with more than one word, whose structure and meaning cannot be derived from their component words' independent meanings (Venkatapathy & Joshi, 2006). MWE can also be defined as idiosyncratic interpretations that cross word boundaries (Sag, Baldwin, Bond, Copestake, & Flickinger, 2002). MWE is a very complicated and problematic issue for natural language processing applications especially for morphologically rich languages like Turkish.

2.3.2.1 Fixed Expressions

Fixed expressions are fully lexicalized and are not subject to either morphosyntactic variation or internal. As a result, a simple words with spaces representation is sufficient (Sag et al., 2002).

2.3.2.2 Semi-Fixed Expressions

Semi-fixed expressions are subject to strict constraints on word order and composition, but can have some lexical variations like inflection. Some samples to semi-fixed expressions are non-decomposable idioms, compound nominals and proper names (Sag et al., 2002).

2.3.2.3 Syntactically Flexible Expressions

Syntactically-flexible expressions show a wide range of syntactic variability. Some examples to syntactically-flexible expressions are verb-particle constructions, decomposable idioms and light verbs (Sag et al., 2002).

2.3.2.4 Institutionalized Expressions

Institutionalized phrases are semantically and syntactically compositional, but they are statistically idiosyncratic. “Traffic light” is an example to institutionalized phrases, in which both “traffic” and “light” retain simplex senses and combine constructionally to produce a compositional form (Sag et al., 2002).

2.3.3 Rule-Based Machine Translation Applications

Many commercial and free translation systems were developed using rule-based translation approaches. Although it is the first method developed for machine translation and other methods are developed afterwards, there are still recent applications which are developed using rule-based machine translation methodologies.

Apertium is a free/open source platform for rule-based machine translation that is developed by Transducens research group at the University of Alicante in 2005. In Apertium, lexical processing is achieved by finite-state transducers, whereas hidden Markov models are used for part-of-speech tagging, and multi-stage finite-state chunking is used for structural transfer (Forcada et al., 2011). Apertium focuses mainly on Romance languages such as Spanish, Catalan, Portuguese, French, Occitan, Galician and English. It started as a platform for closely related languages and extended to include more divergent language pairs like English-Katalan.

GramTrans (GrammarSoft ApS & Kaldera Språkteknologi AS, 2006) is an internet-based machine translation application that provides rule-based constraint grammar parsing for mainly Scandinavian languages (like Danish, Norwegian, Swedish, Portuguese and Esperanto) and English (Wiechetek, 2008).

Matxin (Mayor, 2007) is another rule-based MT system and an open-source toolkit whose first implementation is used for translation from Spanish to Basque. Matxin is the first publicly available machine translation system for Basque and it is stated that although more study has been done on Basque, it is still a less-resourced language. The main focus of the study is the construction of a dependency analyser for Spanish and use of rich linguistic information to translate prepositions and syntactic functions (such as subject and object markers). Also the construction of an efficient module for verbal chunk transfer in addition to the design and implementation of modules for ordering words and phrases is achieved independently of the source language (Mayor et al., 2011).

Some commercial applications, such as (Systran, 2011) and (Apptek, 2012), have also started as a rule-based machine translation and then transformed into a hybrid system adding a corpus-based machine translation system usually after the rule-based system. Thus these applications are considered under Hybrid Machine Translation subsection.

2.4 Corpus-Based Machine Translation

In the beginning of 90s, the advancements in computer technologies and the availability of large amounts of online text have led to corpus-based techniques; which were proposed to “shift the burden of knowledge acquisition from human to computers by inducing linguistic knowledge from large corpora automatically” (Su & Chang, 1992).

The corpus-based approach is mainly studied in two sub-fields: example-based machine translation (EBMT) and statistical machine translation (SMT). The difference between EBMT and SMT is stated in (Hutchins, 2005) as, “input is decomposed into individual SL words and TL words is extracted by frequency data in SMT, whereas in EBMT input is decomposed into SL fragments and TL examples (in the form of corresponding fragments) are extracted from the database”. However it is also emphasized in (Hutchins, 2005) that the distinctions have become blurry after the studies on phrase-based and syntax-based SMT systems.

2.4.1 Example-Based Machine Translation

Example-based machine translation (EBMT) or 'translation by analogy' (as proposed by Nagao) is based on the extraction and combination of phrases (or other short parts of texts) (Nagao, 1984).

Hutchins stated that *translation by analogy* is *the most characteristic technique of EBMT* and it is *the one where the use of entire examples is most motivated* (Hutchins, 2005).

Nagao (1984) claimed that;

…man does not actually translate a sentence by doing deep linguistic analysis, rather, by properly decomposing an input sentence into certain fragmental phrases, then, by translating these fragmental phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase is done by the analogy translation principle with proper examples as its reference. (Nagao, 1984).

Nagao (1984) stated three tasks of EBMT: matching, alignment and recombination. These tasks correspond to the tasks in conventional machine translation as illustrated in Figure 2.2. The source-text analysis process in conventional machine translation is replaced by the matching of the input against the example set in EBMT. The transfer process is replaced by alignment and the generation process is replaced by recombination (Somers, 2001).

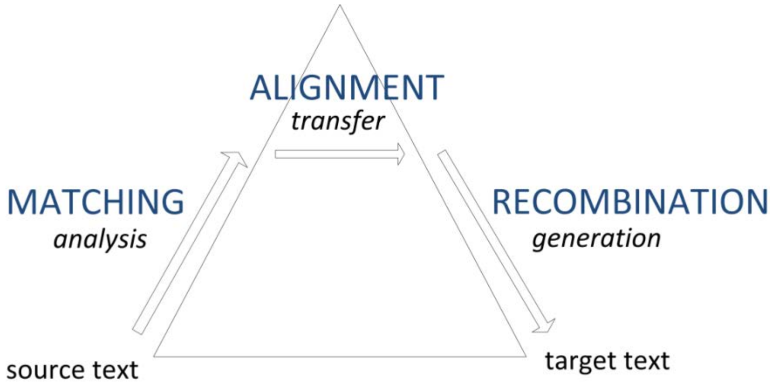


Figure 2.2 The "Vauquois pyramid" adapted for EBMT (Somers, 2001).

The “matching” task can be done in a relatively straightforward manner by using simple character-matching algorithms or with a more linguistically sophisticated matching (Somers, 2004).

The “alignment” task is the selection of the corresponding fragments in the target text, after the relevant example or examples have been selected (Somers, 2001). It can be done by comparing further similar examples and extracting the common elements. Somers (2004) stated that also the use of linguistic resources such as dictionaries can be very helpful.

The “recombination” task is the task of combining the fragments in the way that they fit together properly as the simple concatenation of the fragments may result in translation errors due to “boundary friction”, such as agreement and mutation which might not be covered by the chosen examples. Somers (2004) stated that this problem occurs especially when the target language is considerably more complex than the source language and RBMT can also be helpful at this task.

2.4.1.1 Stages of EBMT

In EBMT, there are four stages: example acquisition, example base management, example application and target sentence synthesis (Kit, Pan, & Webster, 2002).

- ***Example Acquisition***

Example acquisition is the process of acquiring examples from parallel bilingual corpus (i.e., existing translation). Text alignment of bilingual texts is a necessary step towards example acquisition at various levels. Manual alignment by experts can be a solution to produce quite reliable examples, but the price for precision and low speed leads to automatic text alignment technologies.

The studies on automatic text alignment can be categorized into two types; resource-poor and resource-rich approaches. The resource-poor approach focuses on sentence alignment with main focus on sentence length statistics, co-occurrence statistics and some limited lexical information whereas, the resource-rich approaches make use of all available and useful information, in particular, bilingual lexicon and glossary, to facilitate the alignment (Kit et al., 2002).

- ***Example Base Management***

Example base management is the process of storing and maintaining the examples. It is responsible for handling the storage, management (including addition, deletion and modification) and retrieval of examples at high speed, to support the translation process. As a result, an efficient example base management system must be capable of handling a massive volume of examples at an adequately high speed (Kit et al., 2002).

- ***Example Application***

The example application is the process of using the examples to facilitate translation, which also involves the decomposition or segmentation of an input sentence into a sequence of seen fragments (examples) in addition to converting the resulting fragments from the source language into the target language (Kit et al., 2002).

- ***Target Sentence Synthesis***

The target sentence synthesis is the process of composing a target sentence by combining the translated fragments with the aim of enhancing the readability of the target sentence after conversion and forming well-formed highly readable sentences (Kit et al., 2002).

2.4.1.2 Applications

Some example-based translation systems are; the system developed by (Stroppa, Groves, Way, & Sarasola, 2006) for Basque language, and the system developed by (Carnegie Mellon University, 1997) on increasing the efficiency of EBMT with generalizing the examples, which is applied and tested on English-French and English-Spanish languages.

2.4.2 Statistical Machine Translation

In statistical machine translation (SMT), techniques of statistical information extraction from large databases, which contain pairs of large corresponding texts that are translations of each other, are utilized. It was first proposed by IBM as a “purely statistical” approach which was inspired by the success in applications of statistical approaches to speech processing, lexicography and natural language processing (Brown & Cocke, 1988).

2.4.2.1 Noisy Channel Model of SMT

Noisy channel model of SMT, which is based on the noisy channel model introduced by (Shannon, 1948), is a commonly used way to describe SMT (Jurafsky & Martin, 2006; Lopez, 2008; Ramanathan, 2009). In noisy channel model, the sentence (sequence of words) *f* in the source language is considered to be a corrupted version of the sentence *e* in the target language, which is corrupted as a result of the noise in the communication channel. The goal of a SMT system is producing the equivalent sentence *e* (in the target language) of a given sentence *f* (in the source language) using statistical techniques (Ahmed & Hanneman, 2005). The noisy channel model of SMT for source language *f (French)* and target language *e (English)* is illustrated in Figure 2.3.

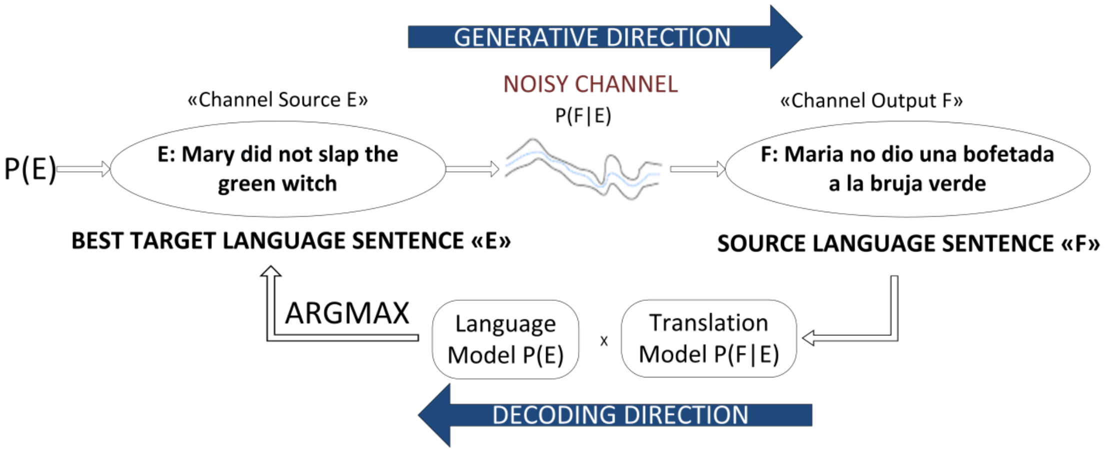


Figure 2.3 The noisy channel model of SMT (Jurafsky & Martin, 2006)

In the noisy channel model, it is required to think of ‘sources’ and ‘targets’ backwards to understand how SMT of a source sentence *f* in French to a target sentence *e* in English is achieved. A translation model, which is the model of the noisy channel, is built from target sentences through a channel to a source sentences. After the translation model is built, the best possible “source” English sentence is searched given a French sentence *f* to translate pretending that the French sentence *f* is the output of an English sentence *e* going through the noisy channel. The selection of the best possible sentence is done using the *translation model*, a *language model* that is built using the target language and a *search or decoding algorithm*. These are three main components of SMT (Jurafsky & Martin, 2006). A graphical illustration of how these components are connected is shown in Figure 2.4 and more detailed explanations of the components are given below for better understanding.

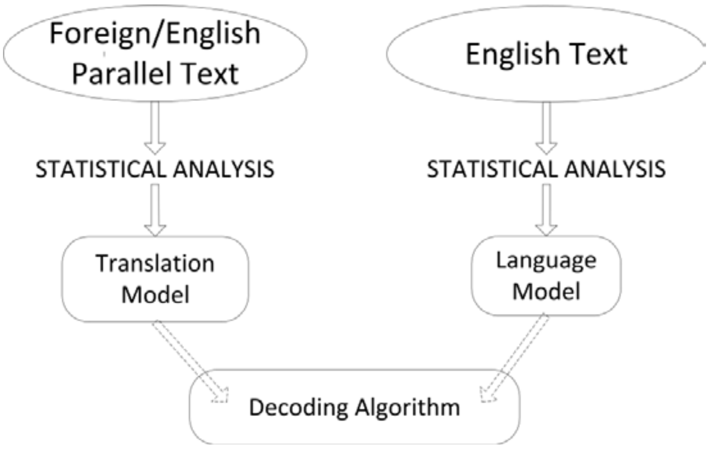


Figure 2.4 Components of statistical machine translation (Koehn, 2007)

- ***Language Model***

The language model, which is denoted by *P(e)* in Figure 2.3, is responsible for *fluency*, in other words for generating valid, fluent target sentences. Language modelling problem can be stated as “the problem of computing the probability of a single word given all of the words that precede it in a sentence” (Brown et al., 1990). Hence, given a word string s_1, s_2,…,s_n the language model can be written as equation 2.1 using the n-gram probabilities, without loss of generality (P. F. Brown et al., 1990; Way, 2010).

$$P(s_1, s_2, \ldots, s_n) = P(s_1)\, P(s_2|s_l) \ldots\, P(s_n|s_l\, s_2 \ldots s_{n-l}) \qquad (2.1)$$

- ***Search or Decoding Algorithm***

The search or decoding problem is to find the target sentence *e* which could have generated *f* with highest probability (Ahmed & Hanneman, 2005). The best *e* is found using equation 2.2 which uses Bayes theorem and the Fundamental Equation of Machine Translation (Brown, Pietra, Pietra, & Mercer, 1993).

$$e^{*} = \text{argmax}_{e}\, P(e)\, P(f|e) \tag{2.2}$$

- ***Translation Model***

Translation model, which is denoted with *P(f|e)* in Figure 2.3, is responsible for the *faithfulness* of the translation. It is the model of the generation process from an English sentence *e* through a noisy channel to a French sentence *f*. In the model, every pair of strings *(e, f)* is assigned a number *P(f|e)*, which is interpreted as the probability that the translator, when presented with *e*, will produce *f* as the result (P. E. Brown et al., 1993; Jurafsky & Martin, 2006).

There are three different groups of translation models for SMT: word-based models, phrase-based models and syntax-based models.

  - ***Word-Based Models***

Word-based models are the original models developed for statistical machine translation. In these models, translation process is tied to the translation of individual words. The first models developed for SMT are IBM Models, Model 1 to 5, which are discriminated by their use of different alignment models and parameters (Brown et al., 1993).

In IBM Model 1, all alignments have the same probability. In IBM Model 2, a zero-order alignment model is used whereas in IBM Model 3, an inverted zero-order alignment model with an additional fertility model that describes the number of words aligned to the word is used. In IBM Model 4, an inverted first-order alignment model and a fertility model is used. Finally, IBM Model 5 is a reformulation of IBM

Model 4 with a suitably refined alignment model in order to avoid deficiency, which is a result of the waste of probability mass on non-strings by IBM Models 3 and 4 (Och & Ney, 2000).

GIZA++ is an implementation of IBM Models which are still used for word alignment mostly as an initial training step of more complex models (Och & Ney, 2000).

- ***Phrase-Based Models***

In phrase-based models, the fundamental unit of translation is a phrase, any contiguous sequence of words, instead of a single word as entire phrases often need to be translated and moved as a unit (Jurafsky & Martin, 2006).

Each phrase in source language translates to exactly one nonempty phrase in destination language. The translation is done in three phases:

i. The source sentence is segmented into phrases.
ii. Each phrase is translated.
iii. The translated phrases are permuted into a final order, reordered if necessary.

A list of all source phrases and all of their translations are contained in a phrase table to use during this process. This phrase table is learned from the training data (Resnik & Park, 2006).

The construction of the phrase table can be done using different methods. A uniform evaluation framework was developed by Koehn, Och and Marcu (2003) and different methods were compared; moreover, it is stated in their study that their experiments showed that high levels of performance can be achieved with fairly simple means and also more sophisticated approaches that uses syntax do not lead to better performance (Koehn, Och, & Marcu, 2003).

- ***Syntax-Based Models***

Syntax-based statistical machine translation models can be based on some form of synchronous grammar to generate source and target sentences in addition to the correspondence between them simultaneously (Ahmed & Hanneman, 2005).

The first study on syntax-based SMT is the application of context-free transduction formalism, inversion transduction grammar, for modeling bilingual sentence pairs that allows some reordering of the constituents at each level (Wu, 1997).

Another important study on syntax-based models is a model that transforms a source-language parse tree into a target-language string using stochastic operations at each node, which capture linguistic differences such as word order and case marking (Yamada & Knight, 2001)

2.4.2.2 Applications

Google translate is the most famous free and online application of statistical machine translation (Google, 2012). Another online SMT is Bing which is developed by Microsoft (Microsoft, 2012). A free statistical machine translation system toolkit MOSES (Koehn et al., 2007) is also available on internet.

2.5 Hybrid Machine Translation

The aim of hybrid machine translation is combining the powers of two approaches above either by using statistics to adjust the outputs of a rule-based translation or using rules to guide corpus-based machine translation.

2.5.1 Applications

Systran is commercial machine translation software which supports 52 language pairs (excluding Turkish). Systran started as a RBMT and then added SMT to the system to achieve fluency and flexibility. They claim that the software can be trained on existing corpora and glossaries can be integrated (Systran, 2011).

Another commercial software, Apptek (Apptek, 2012) supports 30 language pairs. The hybrid machine translation approach is achieved by giving the statistical search process full access to the information available in Lexical Functional Grammar; lexical entries, grammatical rules, constituent structures and functional structures. This is accomplished by treating the pieces of information as feature functions in the Maximum Entropy (Sawaf, Gaskill, & Veronis, 2008).

Another study on hybrid machine translation is developed at University of Alicante for Spanish-English language pair. The system consists of a phrase-based statistical MT system whose phrase table was enriched with bilingual phrase pairs matching transfer rules and dictionary entries from the Apertium rule-based MT platform (Sánchez-Cartagena, Sánchez-Martínez, & Pérez-Ortiz, 2011).

2.6 Machine Translation of Closely Related Languages

Machine translation in grammatically similar languages is easier to achieve as they show similar structural and semantic properties (Altintas & Çiçekli, 2002; Homola & Kuboˇ, 2008). Therefore, just applying morphological analysis and direct translation of the resulting morphological structure can achieve good results. The addition of the syntax and semantic analysis is used for enhancing the performance of the translation system.

Hajič, Hric & Kubon (2000) analysed the examples of two machine translation applications, a transfer-based machine translation application that is developed between Czech and Russian and a word-for-word machine translation system between Czech and Slovak, and stated that machine translation can only be successful between very close languages as it is a very hard task (Hajič, Hric, & Kubon, 2000).

2.6.1 Machine Translation between Turkic Languages

Hamzaoğlu (1993) was one of the first researchers on Machine Translation for Turkish. In his study, a lexicon-based Turkish-Azerbaijani translator was built with no syntax analysis as Turkish and Azerbaijani languages are very similar.

Another study on Turkic languages is a translation system between Turkish and Crimean Tatar (Altıntaş, 2001). In this system, finite state machines were used for translating grammar structures, pre-defined structures and words from one language to the other. The outputs of the system are more than one possible sentence due to no disambiguation. The main steps followed by the system are:

i. Morphological analysis,
ii. Morphological disambiguation,
iii. Translation of pre-defined structures and idioms.
iv. Translation of structures with more than one word and the words which the translation of that word changes up to prior or following words,
v. Match the word in the target language's dictionary,
vi. Morphological generation in target language.

The system was tested on translating some Turkish sentences to Tatar, the results of translating one Turkish sentence to Tatar are given below.

Input : Turkish sentence
Akşam eve geleceğiz. (Tonight we will come home)

Output : Tatar sentence
aqSam evge kelecekmiz *(Tonight we will come home)* (2 different analysis leading to same translation)
aqSam evge istiqbalmIz (Tonight home our future)

It is stated in the study that syntactical analysis at the source language would enhance the performance of the system (Çiçekli, 2005).

A more recent study on Turkic languages is a hybrid translation model, which combines rule-based and statistical approaches using two level morphology, is developed by Tantuğ (2007). The hybrid translation model is based on a modified version of direct word-by-word translation model. Finite state methods with two level morphology are also employed in addition to multi-word processing and statistical methods for disambiguation (Tantuğ, 2007). A Turkmen to Turkish translation system (Tantuğ, Adali, & Oflazer, 2009) was designed and implemented for testing the hybrid translation model.

The steps followed by the hybrid translation model (Tantuğ, 2007) during translation are:

i. Tokenizer
ii. Source Language Morphological Analyzer
iii. Source Language Multi-word Processor
iv. Morphological Feature Transfer
v. Direct / Lexicalized Root Word Transfer
vi. Unified Statistical Language Model
vii. Sentence Level Lexical Form Rules
viii. Target Language Morphological Generator.
ix. Sentence Level Surface Form Rules

It is stated in the model that, both lexical and morphological ambiguities at the target language side (Turkish) are handled with the help of a unified statistical language model as Turkmen is a resource poor language. A set of rules working on the sentence level is also used to ease the drawback of direct transfer strategy. It is also stated that addition of new languages is possible for translation from any Turkic language to Turkish, but not in the opposite direction (from Turkish to other language) as they use the Turkish corpus for disambiguation. An Uyghur to Turkish translation system (Orhun, Adali, & Tantuğ, 2011) is also developed using this model (Tantuğ, 2007).

Another study is the DİLMAÇ Project (Fatih University, 2013), which is started by Turkmen and Turkish morphological analysers and a Turkmen-Turkish translator (Shylov, 2008). DİLMAÇ is also based on two-level morphological analysis. The translation is performed word by word; hence it does not support multi-word expressions. Twenty four languages are listed as available in DİLMAÇ although the lexicon size changes for different languages and also there are some problems with the rule files of some languages. At the date of the citation (04.10.2013), the richest lexicons are Uyghur, Japanese and Turkish with word counts respectively 37716, 19903 and 14906; however the word counts for languages Kirghiz and Kazan Tatar are really low with 39 and 23 words respectively. Also no translation can be achieved for Kirghiz as the rule file for Kirghiz is missing.

2.7 Machine Translation Evaluation

The evaluation of the machine translation results is achieved either by humans or machines. Human evaluation is very time consuming and maintaining objectivity is not always easy.

However machine evaluation is not also easy, as for evaluating a translation algorithm we need an algorithm which can define what a good translation is. Thus, the problem of machine translation evaluation is the same problem of translation. Nevertheless there are some features which can be evaluated automatically (Zwarts, 2010).

There are various machine evaluation techniques available like F-Measure (Turian, Shen, & Melamed, 1995), Sentence Level Evaluation (Kulesza & Shieber, 2004), Meteor (Lavie, Sagae, & Jayaraman, 2004) and String Accuracy Metrics (Marrafa & Ribeiro, 2001); however the most known and widely used machine translation evaluation metrics are BLEU (Papineni, Roukos, Ward, Zhu, & Heights, 2001) and an enhanced version of it, NIST (Doddington, 2002a). Although these techniques are mainly designed for evaluating statistical machine translation, they suffice enough at least as a start for evaluation as they measure the fluency of the translated text.

In this study, BLEU and NIST evaluation metrics are used for evaluating the results of MT-Turk translation. Hence, brief information about these evaluation metrics with formulas and parameters is given below.

2.7.1 BLEU(Bilingual Evaluation Understudy)

This metric is an IBM-developed metric and it defines the success of the translation based on an n-gram comparison of translated sentences with reference sentences.

It is the best known machine evaluation technique for machine translation and is accepted to being correlating well with human judgment although it does not measure the success of the algorithm instead how it scores against references. The unigram comparison tends to satisfy adequacy whereas the longer n-gram comparisons achieve the checking of the fluency of the produced sentence.

BLEU score is calculated using the equation 2.3 (Doddington, 2002b).

$$BLEU = exp\left\{\sum_{n=1}^{N} w_n \log p_n - max\left(\frac{L_{ref}^{*}}{L_{sys}} - 1 , 0\right)\right\} \tag{2.3}$$

BLEU is the calculated by first calculating the geometric mean of the test corpus' modified precision scores and then multiplying the result by an exponential brevity penalty factor.

Modified n-gram precisions for each n-gram (unigram, bigram, etc.) are calculated first. Then a weighted average of the logarithm of the modified precisions is calculated. The weight w_n equals to *1/N* where in the baseline *N* is 4.

Modified n-gram precision score, p_n is computed for any n using the equation 2.4: all candidate n-gram counts and their corresponding maximum reference counts are collected (Papineni et al., 2001). The candidate counts are clipped by their corresponding reference maximum value, summed, and divided by the total number of candidate n-grams. *Count_{clip}* truncates each word's count, if necessary, to not exceed the largest count observed in any single reference for that word using the equation 2.5.

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count(n-gram)} \tag{2.4}$$

$$Count_{clip} = min(Count;Max\_Ref\_Count) \tag{2.5}$$

The final part is brevity penalty factor; it penalizes candidates shorter than their reference translations using L^{*}_{ref} and L_{sys}.

L^{*}_{ref} = the number of words in the reference translation that is closest in length to the translation being scored.

L_{sys} = the number of words in the translation being scored.

Consequently, BLEU metric evaluates the likelihood of the candidate with regard to one or more reference texts and outputs a number between 0 and 1, 1 being the most similar.

2.7.2 NIST

This metric is developed by National Institute of Standards and Technology. It can be seen as an upgrade to BLEU metric as it uses the same n-gram technique and improves the BLEU metric by solving some problems of it. These improvements are:

- BLEU uses geometric mean of n-grams over N which makes the score equally sensitive to proportional differences in co-occurrence for all N. This can lead to the potential of counterproductive variance due to low co-occurrences for the larger values of N. NIST uses an arithmetic average of N-gram counts rather than a geometric average.

- BLEU treats all n-grams equally. Thus the n-grams with little information have the same value as the information rich n-grams. The richness of the information is in negative correlation with how often the n-gram occurs. Whereas NIST consider the fact that those N-grams that are most likely to (co-)occur would add less to the score than less likely N-grams.

- BLEU is not case insensitive whereas NIST is.

NIST score is calculated by the equation 2.6 (Doddington, 2002b).

$$NIST = \sum_{n=1}^{N} \left\{ \frac{\sum_{\substack{all\, w_1 \dots w_n \\ that\ co-occur}} Info(w_1 \dots w_n)}{\sum_{\substack{all\, w_1 \dots w_n \\ in\ sys\ output}} (1)} \right\} \cdot exp\left\{ \beta \log^{2} \left[min\left(\frac{L_{sys}}{\bar{L}_{ref}}, 1 \right) \right] \right\} \quad (2.6)$$

where

- β is chosen to make the brevity penalty factor = 0.5 when the # of words in the system output is $2/3^{rds}$ of the average # of words in the reference translation,
- N equals 5
- $\bar{L}_{ref}$ is the average number of words in a reference translation, averaged over all reference translations
- L_{sys} is the number of words in the translation being scored and
- *Info(w1 … wn)* is calculated by the equation 2.7.

$$Info(w_1 \dots w_n) = \log_2 \left(\frac{the\ \#\ of\ occurences\ of\ w_1 \dots w_{n-1}}{the\ \#\ of\ occurences\ of\ w_1 \dots w_n} \right) \qquad (2.7)$$

Notice that, in addition to the calculation of the co-occurrence score with information weighting, a change was also made to the brevity penalty. This change was made to minimize the impact on the score of small variations in the length of a translation.

It is stated by Doddington (2002b) that for human judgments of adequacy, the NIST score correlates better than the BLEU score. For fluency judgments, however, the NIST score correlates better than the BLEU score only in one of the corpora which are used on tests (Chinese corpus) (Doddington, 2002b).

CHAPTER THREE
MT-TURK INFRASTRUCTURE

The scope of this dissertation is the design and implementation of an infrastructure for a translation system between Turkic dialects. The translation system, MT-Turk, is a semi-supervised machine translation infrastructure for Turkic languages and was developed in a rule-based manner. The rule-based approach was selected for the reason that there are no parallel texts for Turkish and Kirghiz or Turkish and Kazan Tatar to train a corpus based machine translation infrastructure. Two subsets of rule-based approach, the interlingual machine translation approach and transfer-based approach, were used in combination to form the multilingual machine translation infrastructure developed in this study for the sake of achieving extensibility and interoperability. The combined rule based approach is more effective by uniting the advantages of both interlingual approach and transfer approach. The analysis of the input text is done to form a semi-interlingual representation and the transfer of this semi-interlingual form is performed using transfer rules.

The stems are translated using the interlingual machine translation approach, source language sentences are analyzed and each word or multi-word group is converted to a language-neutral representation of the concept they identify, common to more than one language (Drozdek, 1989) and no bilingual transfer dictionaries are required. Hence, the most crucial and problematic resource of the translation system, the lexicon, can be enhanced easily. On the contrary, the suffix transfers and word order changes are achieved using transfer based machine translation approach.

Extensibility of the system is achieved by the semi-interlingual representation as there is no need for language specific analyzers and generators. Disambiguation and forming a fully language independent canonical representation of meaning in the sentence (pure interlingua) is very difficult for Turkic dialects as a result of their under resourced property (lack of a large corpus) and not very necessary as Turkic dialects are closely related, word order is almost the same and semantics are similar. Hence, the language specific rules are used by the transfer phase to control and ensure the validity of the word group and suffix order in addition to proper suffix selection. The main architecture of the translation system is shown in Figure 3.1.

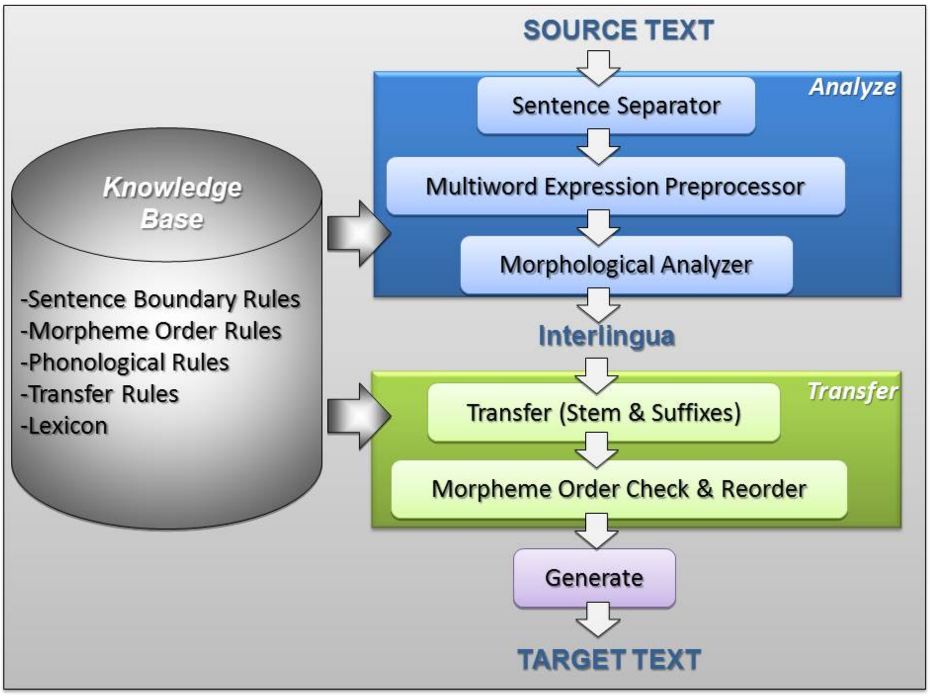


Figure 3.1 MT-Turk architecture

3.1 The Software Technologies Used for MT-Turk

The software which was developed in the scope of this dissertation, MT-Turk, is an ASP.NET web application and is implemented using Microsoft .NET Visual Studio 2010 (.NET Framework 4.0) environment with C# programming language.

The application is consisted of a Phonology library and MT-Turk web application. Phonology library has maintainability index 80, depth of inheritance 3 and 725 lines of code whereas MT-Turk has maintainability index 73, depth of inheritance 5 and 1812 lines of code.

The main data used by the application, lexicon and the suffixes, were stored in MS SQL Database Server whereas the rules were stored as text files or XML files. The structures of the lexicon, suffixes and rules are given in detail below.

3.2 Knowledge Base

There are three types of rule lists which are required by the system: sentence boundary rules, morpheme order rules and phonological rules. Besides, transfer rules for suffixes are also a requirement and are hold as a table in CONCEPTSET database.

3.2.1 Sentence Boundary Rules

The Sentence Boundary Rules, which are designed and implemented by (Aktaş & Çebi, 2010; Aktaş, 2006), are stored in an XML file and used by the Sentence Seperator component. A sample sentence boundary rule stating that a sentence boundary is matched when there is a punctuation mark between a lower letter and an upper letter is defined as:

```
<rule EOS="True">L.U</rule>
```

Note that, “L” is used for identifying lower letters, “U” for upper letters and “.” is used for punctuation marks which can either be “.”, “…”, “!” or “?”. The details on the format of the sentence boundary rules can be found in (Aktaş & Çebi, 2010; Aktaş, 2006).

3.2.2 Morpheme Order Rules

Morpheme order rules are designed and implemented by (Birant, Aktaş, & Çebi, 2010). The validity of the morpheme order is checked and achieved by using three rule files: “morpheme ordering rules”, “must rules” and “not rules”. All of the Morpheme Order Rules are stored in text format in separate files (Birant, 2008). “Morpheme ordering rules” file lists all the possible morpheme sequences that can result in a valid word. A sample morpheme ordering for verbs is defined with the rule:

Rule: E,TBEE\,DuEC,DuEOlz,Ytu,K,YS-y,DuEK

where;

- E : Verb
- TBEE : Derivation suffixes deriving verbs from verbs
- \ : Special symbol stating that the suffix group can be reiterated
- DuEC : Voice
- DuEOlz : Negation
- Ytu : Subordination
- K : Copula
- YS-y : Buffer sound
- DuEK : Personal endings

“Must rules” are used to define constraints that must be achieved, more specifically when there is a suffix that must be preceded by another. For example, when used with the verbs, copula must be used only after time suffixes. This constraint on the relation of copula and the time suffixes is defined with the rule:

Rule: E,K,DuEZ

where;

- E : Verb
- K : Copula
- DuEZ : Tense suffixes

“Not rules” specify the tag sequences that must be avoided, i.e. the suffixes that cannot occur in the same word. For example, case suffixes cannot be followed by number suffixes. This constraint on the relation of case and number suffixes is defined with the rule:

Rule: DuADur,DuASay

where;

- DuADur : Nominal Case
- DuASay : Nominal Number

3.2.3 Phonological Rules

The phonological rules were designed and implemented in the scope of this study, in collaboration with the Natural Language Processing Research Group (DEU CSE, 2004), with the aim of modelling the assimilation and harmony rules in Turkic dialects.

The rules are used for the analysis, alternation and the generation purposes. For the purposes of language independency, the required phonological information, the morphophonemics is supplied to the system as an XML file.

The XML file holds three parts: alphabet, substitutions and rules. The alphabet of the language is stored in the phonological XML file along with the type information (consonant, vowel).

Substitutions and rules are stored in the same format. Each consists of four properties: *id*, *name*, *valid* and *force_match* and two parts: *match* and *action.*

Parts:

- *Match* defines the pattern to be matched to apply the rule.
- *Action* defines what action to take if the rule is to be applied. Action part is especially used during the alternation process.

Properties:

- *Id* is a unique number identifying the rule.
- *Name* is used to hold a representative name for the rule.
- *Valid* is used to identify what the rule checks: validity or invalidity.
- *Force_match* is set to true for the kind of rules that the match part is required to be matched or the rule rejects the morpheme sequence. The rules with this property should be applied after other rules. A typical example to this kind of rules is vowel harmony rules.

The substitutions are used for character substitutions in suffix representations like A → a|e, whereas the rules are used for constraint checking. A sample rule for Final Devoicing is shown in Figure 3.2. A complete list of rules is given in Appendix A.

```
<rule index ="2" valid="True" force_match="False">
  <match>
    <stem>
      <pattern loc="last" type="char">
        <lex>p|ç|t|k</lex>
        <surf>b|c|d|g|ğ</surf>
      </pattern>
    </stem>
    <suffix>
      <pattern loc="first" type="char">
        <lex>vowel</lex>
        <surf>vowel</surf>
      </pattern>
    </suffix>
  </match>
  <action>
    <stem>
      <pattern loc="last" type="char">
        <pair index="0">
          <lex>p</lex>
          <surf>b</surf>
        </pair>
        <pair index="1">
          <lex>ç</lex>
          <surf>c</surf>
        </pair>
        <pair index="2">
          <lex>t</lex>
          <surf>d</surf>
        </pair>
        <pair index="3">
          <lex>k</lex>
          <surf>ğ</surf>
        </pair>
        <pair index="4">
          <lex>nk</lex>
          <surf>ng</surf>
        </pair>
      </pattern>
    </stem>
  </action>
</rule>
```

Figure 3.2 The rule for Turkish final devoicing

3.3 Translation Components

The components of MT-Turk translation, five software modules and the interlingua, are illustrated in Figure 3.1. Each module and the interlingua is described in detail below.

3.3.1 Sentence Separator

The sentence separator is the initial module of the analysis. In this module, a rule-based sentence boundary detection algorithm developed by (Aktaş & Çebi, 2010; Aktaş, 2006) is used. The sentence boundary definition rules and abbreviation lists that are used by the algorithm were defined in collaboration with the linguists and tested on large amounts of data. The average success rate of the algorithm was reported as 99.78% (Aktaş & Çebi, 2010; Aktaş, 2006).

The text to be translated is analyzed and separated into sentences by the sentence separator and each resulting sentence is sent to multi-word expression preprocessor for further analysis.

3.3.2 Multi-Word Expression Preprocessor

Each multi-word expression type needs special attention and different strategy. The first two types of MWEs, fixed and semi-fixed expressions, are the ones that will be matched from the lexicon. The existing Turkish lexicon holds multi-word expressions of these kinds as stems, because they represent a different meaning from the independent meanings of its component words' meanings. Hence, the morphological analyser developed by (Birant et al., 2010) is enhanced so that the multi-word expressions in the database is combined to form a single word with a word boundary symbol "#" in between.

The input text is analysed by the multi-word expression preprocessor prior to the morphological analysis. The possible multi-word list is gathered from the lexicon and the input text is searched for the existence of these multi-words. If they exist, input text is updated by combining these multi-words in a word with the boundary symbol “#” in between.

The multi-word groups of the third type, syntactically-flexible expressions, are handled by specific morphophonemic rules. These multi-word groups should be defined by linguists, matched and translated as a group structure by the system. The XML file which contains phonological information for the language should contain these rules for forming the multi-word groups of non-lexicalized collocations.

The rules are specified with a special tag <mwrule> (Multi-Word Rule). Each rule should specify the group name, the lexical form and the surface form of the match structure. The match structure can define structures to be matched in more than one adjacent word with different suffixes to be matched in each word. Some special abbreviations are used: W is for a word followed by the index (order) number of the word and # for identifying word boundary.

An example rule is shown in Figure 3.3. The rule specifies a multi-word group construction (“ir_mez”, as in “gelir gelmez*: as soon as he comes*”). The name of the group is “ir_mez”. The lexical form specifies that first word must have the suffix “YtuU1 - Ir” followed by the word boundary and the second word must have the suffix “YtuU2 - mAz”. The surface form of this group is formed by enclosing the two matched words with a group tag.

```
<Mwrule>
        <Group> ir_mez </Group>
        <Lex> W1<YtuU1>Ir</YtuU1>#W2<YtuU2>mAz</YtuU2></Lex>
        <Surf> <Group name= “ir_mez”>W1#W2 </Group> </Surf>
</Mwrule>
```

Figure 3.3 Sample multi-word rule

During the multi-word pre-process, the input string is analysed and all the multi-word rules are checked to see if the lexical structure of the rule is matched. If the rule is matched, it is applied by transforming the matched structure to form the surface structure. A sample interlingua that is formed by the application of the multi-word rule above is shown in Figure 3.4.

```
<Group name= “ir_mez”>
<Word>
        <ValueOfWord> gel ir</ ValueOfWord >
        <Root Index=”2545”>
        <Value> gel </Value>
        <Suffixes>
                <SuffixCombination Index=”0”>
                        <YtuU1>ir</YtuU1>
                </SuffixCombination>
        </Suffixes>
</Word>
<Word>
        <ValueOfWord> gelmez </ ValueOfWord >
        <Root Index=”2545”>
        <Value> gel </Value>
        <Suffixes>
                <SuffixCombination Index=”0”>
                        <YtuU2>mez</YtuU2>
                </SuffixCombination>
        </Suffixes>
</Word>
</Group>
```

Figure 3.4 Sample group interlingua

The multi-word groups of the forth type, institutionalized phrases, are proper names. Person names, which is a sub-group of the proper names, are stored at the database which is taken from Turkish Linguistic Association (TDK)’s Person Names Dictionary (TDK, 2009).

As it is impossible to obtain and store all proper names, the multi-word analyser also analyses the sentences up to the cases of the first letters of words. It marks:

- a group of words as a candidate for proper name if the first letters of adjacent words are capital,

- any single word in the middle or at the end of the sentence with a capital first letter as a candidate for proper name.

Note that the second marking process analyses only words in the middle or at the end of the sentence. Even though the first word of the sentence can also be a proper name, it cannot be marked by the case of the first character as it is always capital, it should be marked by the help of the dictionary or the users' input.

Apart from these multi-word expression groups, some of the suffixes may also form an extra word during conjugation. Turkish has only one example of these suffixes, which is the question suffix "mI", whereas Kirghiz has several. An example to these suffixes is one of the forms of continuous tense in Kirghiz, "–a cata(t)" (Çengel, 2005). The conjugation results in two words when a verb is conjugated with this suffix. For example, the word "gidiyor (*is going*)" in Turkish is formed with two words "bara catat" in Kirghiz.

These multi-word constructions are achieved with auxiliary verbs, some special words or negation words which are located in second part of the suffixes. However in Kirghiz, these auxiliary verbs may also be used individually. Some examples of auxiliary verbs are: cat-, cür-, tur- and otur-, special words are: bolso, kerek and bar and negation words are: elek, cok and emes. These suffixes are handled in a similar manner to multi-words. The suffix is represented as "-a#catat" and the multi-word preprocessor also searches for these suffixes, matches and combines them into a word.

3.3.3 Morphological Analyser

The morphological analysis of the source text is the last step of the analysis phase. The analysis is performed by a slightly improved version of the morphological analyser developed by Birant (2008).

In the original analyser the root list is used during the analysis, whereas for translation purposes stem list is more appropriate; moreover, the analyser should be able to work for different source languages. Thus, the original morphological analyser is enhanced with the ability to use the stem list and also with the ability to choose different language databases and rule lists on request.

Furthermore the text is analysed in a word by word manner in the original analyser, so multi-word expressions are not supported. Therefore, the original analyser is also enhanced to be able to analyse multi-word expressions by the multi-word expression pre-processor.

3.3.4 The Interlingua

In MT-Turk, a modified interlingua approach is used because of the output of the morphological analyser. The output of the original morphological analyser is a list of all possible root-suffix combinations in XML format and is language dependent; i.e. it contains the value of the root and the suffixes in the source language. This structure forms an XML output that is easy to read and interpret also by the human eye.

However, the interlingua must be language independent so as to be used during translation between any two languages in the system. The language independency is achieved by a small addition to the root information (*concept id*) in the output of the morphological analyser. The *concept id* is hold at in the *Index* attribute of the *Root* tag and the morphological analyser must be called with a specific parameter to return the output in this format. The use of the *concept id* achieves language independency in stems as it is common to all the languages. Suffix tags are also common between languages but some of the suffix tags may not exist in all languages which require a transfer mechanism for the suffix tags during translation.

Subsequently, although the roots are language independent, the tags are still in the source language and also the values of the roots and suffixes are still present. Therefore the output is semi-language specific. In other words, the interlingua has the language specific output of the analyser with the language independent *concept id* information of the roots added.

The main intention behind keeping the original design with a small addition is to maintain high readability of the XML output in addition to maintain interoperability between the morphological analyser and previously developed tools for Turkish. High readability of the XML output is very useful especially during language resources development process. A sample interlingua is given in Figure 3.5.

```
<Word>
        <ValueOfWord> evde </ ValueOfWord >
        <Root Index=”25313”>
        <Value> ev</Value>
        <Suffixes>
                <SuffixCombination Index=”0”>
                        <DuADurBul>de</DuADurBul>
                </SuffixCombination>
        </Suffixes>
</Word>
```

Figure 3.5 Interlingua sample

3.3.5 Transfer

As a result of a semi-language specific interlingua, the output of the analyser needs an additional transfer process before it can be generated in the destination language. The stem from the source language should be replaced by the corresponding stem in the destination language and the required transformations for the suffixes must be achieved. The transfer for stems is achieved with Algorithm 1.

Algorithm 1. Stem transfer algorithm

```
get the concept id of the stem
search for an entity with that concept id in the target language’s stem table.
if the entry is found then
    substitute the old representation with the found one
else
    search the concept cover relation for parents of the concept id and search for
    the parent concept id in the target stem table
substitute the old representation with the representation of the parent concept
```

For example; translating the word “yay- *summer*” from Azerbaijani to Turkish is done by following the steps below:

- Get the *concept id* of “yay” → 3
- Search stem with *concept id* 3 in target database (TR: Turkish) → *concept id* 3 is not found
- Search for the parent of the concept → parent of the concept 3 is concept 1
- Search *parent concept* in target database (TR) → “yaz”-noun

The databases and tables with entries that are used in this example are illustrated in Section 3.5 Database Model.

The transfer for suffixes is achieved in two levels, the correspondence of the suffixes and the reordering of the suffixes. The correspondence of the suffixes is achieved with Algorithm 2.

Algorithm 2. Suffix transfer algorithm

```
get the suffix combination to be translated
for each suffix combination s in the CONCEPTSET tag relation
        check and replace the combination s in the input suffix combination
for each suffix in the suffix combination
        check and replace the suffix with the corresponding suffixes in the
        CONCEPTSET tag relation
```

For example; translating the suffixes of the word “başta***gan***mın - *I had begun*” from Kirghiz to Turkish is done by following the steps below:

- Get the suffix combination of “başta***gan***mın” → ganmın : DuEZG2+DuEKGr2T1
- Check for suffix combinations to be replaced → no matching combination for DuEZG2+DuEKGr1
- Find correspondences for the existing suffixes→ replace DuEZG2 with DuEZGM (*past tense “mIş”*) + KDi (*copula* “DI”)
- The new form of the suffix combination is DuEZGM+KDi+DuEKGr2T1

The morpheme reordering is also achieved by the transfer component. Firstly, all the coalescence consonants are removed from the analyse output. Then, this form is checked through the morpheme ordering rules in the destination language. If it is not matched, the morpheme sequence is reordered in a combinational manner and all combinations are checked through the morpheme ordering rules.

Some suffixes have identical suffixes which are chosen or eliminated by the order rules. Person suffixes can be given as an example to these suffixes which has four different groups in Turkish. DuEKGr1T1 and DuEKGr2T1 are both first singular person suffixes, but used with different tense suffixes. The suffixes of this type must be defined as identical suffixes and also the tag part which discriminates the groups should be defined. The “Gr1” or “Gr2” parts, relatively, are the parts which discriminates the two identical suffixes.

If the morpheme order is not valid and if the morpheme sequence contains a suffix with identical suffixes, the suffix is replaced with the identical suffixes and the order check is renewed. The morpheme sequences with a valid order are then sent to the generator.

In this phase, parallel processing is also used for speeding up the translation. The processes of morpheme reordering and checking the validity of the new order are executed in parallel.

3.3.6 Generation

The generation is achieved by the phonology library which takes the stems and morpheme sequences and combines them in a word matching the language specific phonological constraints. The algorithm of the generation is given in Algorithm 3.

Algorithm 3. Generation algorithm

```
find all possible representations of the morphemes
for each possible morpheme sequence
    analyze the morpheme sequence through the phonology constraints (As defined in Phonological Rules and Phonology Library subsection above)
    add coalescence consonants where necessary
combine different analyze results that result in the same representation in the destination language into one
```

One word can have more than one possible translation; therefore, translation of a sentence can result in multiple sentences. In MT-Turk, all possible translations are listed instead of choosing one with disambiguation techniques as some ambiguities cannot be resolved at sentence level. Also most of the disambiguation studies require a corpus, which reveals a problem with languages with low resources like Kirghiz. These ambiguous sentence formations are hard to read and evaluate, thus the MT-Turk online translator is designed to put just ambiguous words in a dropdown box and display the unambiguous ones as normal text. Furthermore, if the translation is unsuccessful at the analysis level, the original word is displayed with a dark blue background whereas if the translation is unsuccessful at the generation level the original word is displayed with a red background outlining the problem.

A suggestion system is also integrated in the system to assist in resolving ambiguities. The system collects suggestions from the user and stores with context information, in other words the sentence it was used in. Therefore during a new translation; when there is an ambiguity in a word that was suggested before, the suggestions are shown to the user at the top of the ambiguous words drop down list in an descending order of total suggestion counts.

3.4 Software Modules

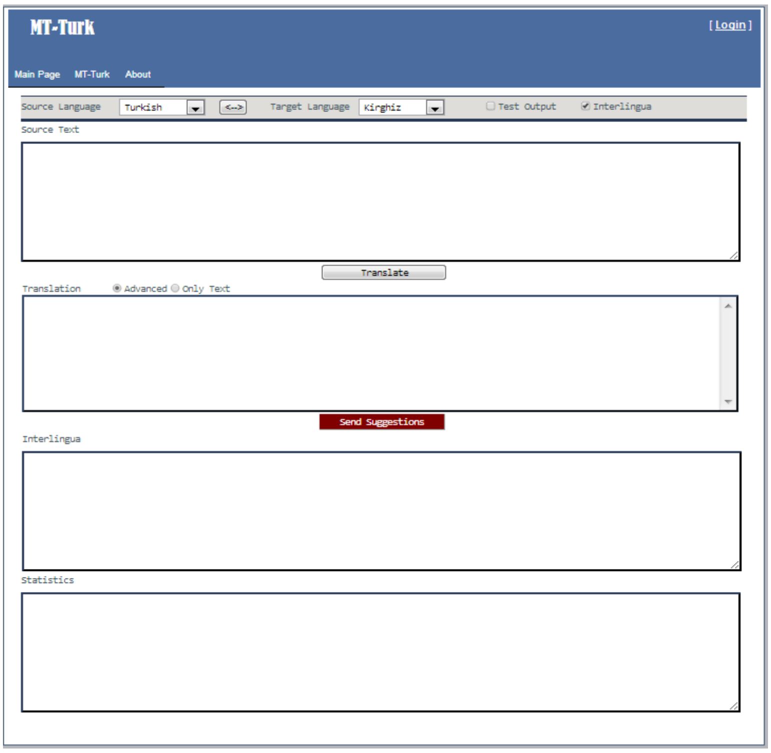
MT-Turk consists of five main software modules: translator, lexicon, suffix manager, multiword manager and rule processors; and it works for two types of users: anonymous user and administrative user. Each user has different access levels.

In the reminder of this subsection the software modules will be described with screenshots in groups of the user types.

3.4.1 Anonymous User

Anonymous user is the default user. If the user does not login, it is considered as an anonymous user and (s)he can only use translator module, however (s)he cannot do any suggestions.

The screenshot of the translation software module is given in Figure 3.6. The anonymous user can select the source language and the target language or swap the languages using the button in the middle of the two language dropdown boxes. The “Interlingua” checkbox can be used to see the interlingua after translation. The “Test output” checkbox can be used to indicate that output must be given in SGML (Standard Generalized Markup Language) format which is a standard markup language SGML, defined by ISO 8879:1986 (ISUG (International SGML/XML Users’ Group), 2010), for supplying evaluation input to *mteval-v13a* (NIST (National Institute Of Standards And Technology), 2010) evaluation program.



Source Language Turkish

Target Language Kirghiz

☐ Test output

☑ Interlingua

Translation ● Advanced ○ Only Text

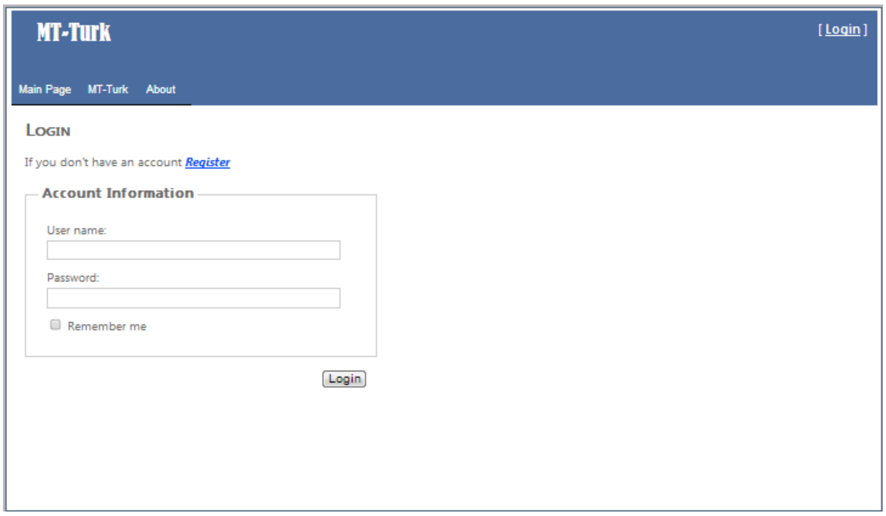
Figure 3.6 Anonymous user translation module screenshot

3.4.2 Administrative User

The administrative user is the main user of the system and has access to all modules with full administrative rights.

3.4.2.1 User Login

The administrative user must log in to be able to access the modules. The screenshot of the login screen is given in Figure 3.7.



If you don't have an account Register

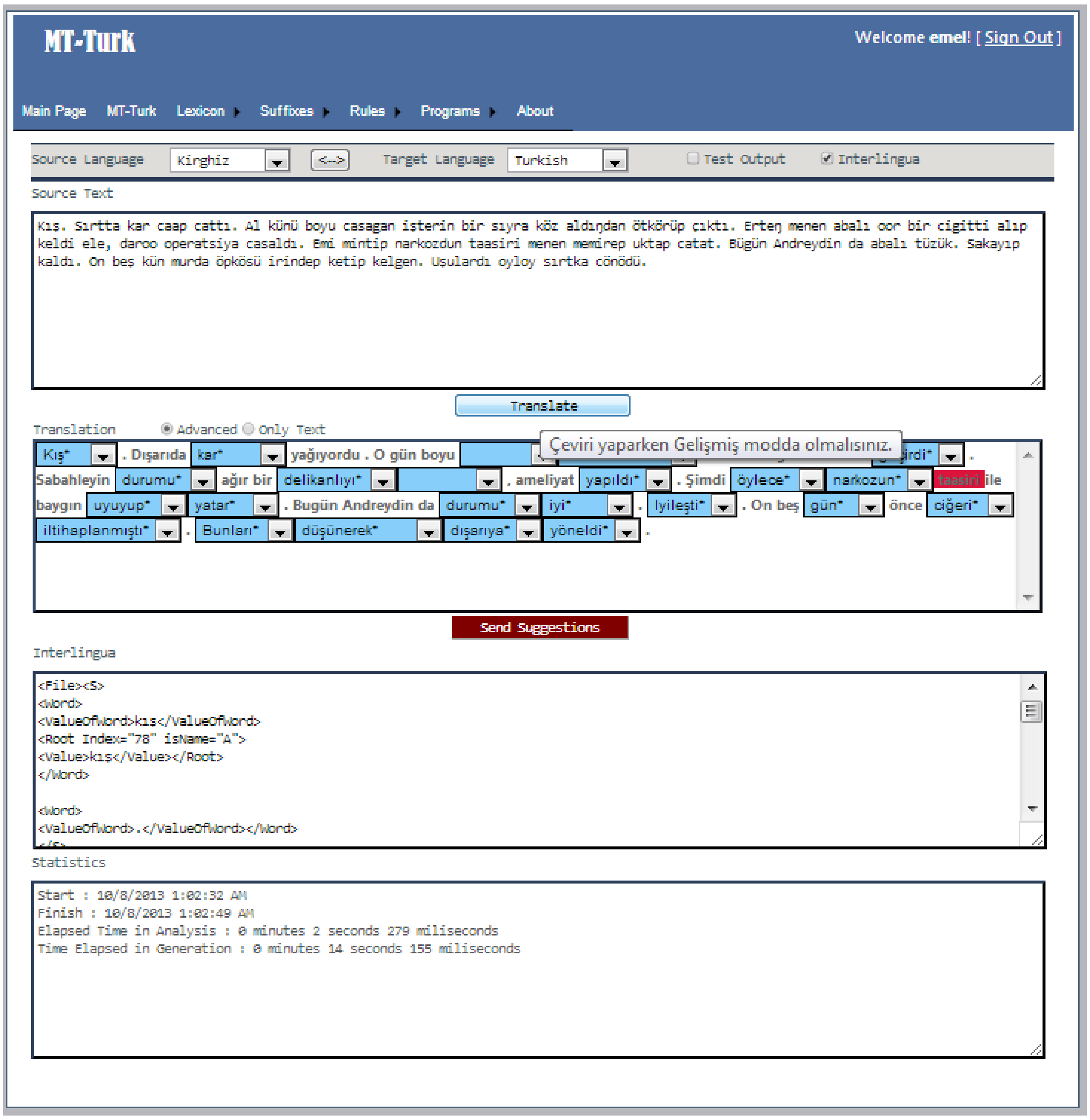
Figure 3.7 Login screenshot

3.4.2.2 Translator

The translator module is the same as the one anonymous user has access to with additional suggestion sending capability. The administrative user can select the correct translation outputs from ambiguous output dropdown boxes and send suggestions.

The system collects suggestions from the user, stores them in the system and shows suggestions, marked by a *, when there is a disambiguation of the word that is used in that suggestion. A sample translation result is given in Figure 3.8.

By means of listing the ambiguities that could not be resolved, the user is given the ability to select and edit translation output. The ambiguous words in the translation output are shown in a dropdown box and the unambiguous ones are shown as normal text to achieve a simple and efficient view of the output.



Source Language: Kirghiz
Target Language: Turkish
□ Test Output ☑ Interlingua

Source Text

Kış. Sırtta kar caap cattı. Al künü boyu casagan işterin bir sıyra köz aldıŋdan ötkörüp çıktı. Erteŋ menen abalı oor bir cigitti alıp keldi ele, daroo operatsiya casaldı. Emi mintip narkozdun taasiri menen memirep uktap catat. Bügün Andreydin da abalı tüzük. Sakayıp kaldı. On beş kün murda öpkösü irindep ketip kelgen. Uşulardı oyloy sırtka cönödü.

Translation ● Advanced ○ Only Text

Kış* . Dışarıda kar* yağıyordu . O gün boyu … Çeviri yaparken Gelişmiş modda olmalısınız. … irdi* .
Sabahleyin durumu* ağır bir delikanlıyı* , ameliyat yapıldı* . Şimdi öylece* narkozun* taasiri ile
baygın uyuyup* yatar* . Bugün Andreydin da durumu* iyi* . İyileşti* . On beş gün* önce ciğeri*
iltihaplanmıştı* . Bunları* düşünerek* dışarıya* yöneldi* .

Interlingua

```
<File><S>
<Word>
<ValueOfWord>kış</ValueOfWord>
<Root Index="78" isName="A">
<Value>kış</Value></Root>
</Word>

<Word>
<ValueOfWord>.</ValueOfWord></Word>
```

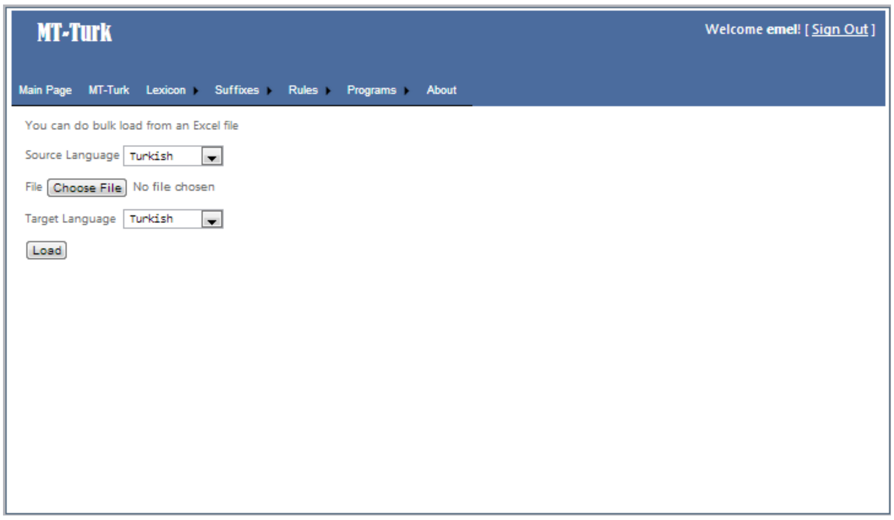
Statistics

```
Start : 10/8/2013 1:02:32 AM
Finish : 10/8/2013 1:02:49 AM
Elapsed Time in Analysis : 0 minutes 2 seconds 279 miliseconds
Time Elapsed in Generation : 0 minutes 14 seconds 155 miliseconds
```

Figure 3.8 Translator module output screenshot

3.4.2.3 Lexicon

Lexicon manager is consisted of two sub-modules: bulk lexicon uploader and lexicon manager. Bulk lexicon uploader gets bilingual lexicon data as an Excel file with two columns in addition to source and target languages through dropdown boxes and stores the data in the lexicon. Bulk lexicon uploader is required for large data entry as manually digitizing a lexicon requires too much time and effort. Still it should be noted that, when there are two languages in the system bulk loading will be less efficient than manual entry as it connects the new stems with only one language. The bulk lexicon loader screenshot is given in Figure 3.9.



You can do bulk load from an Excel file

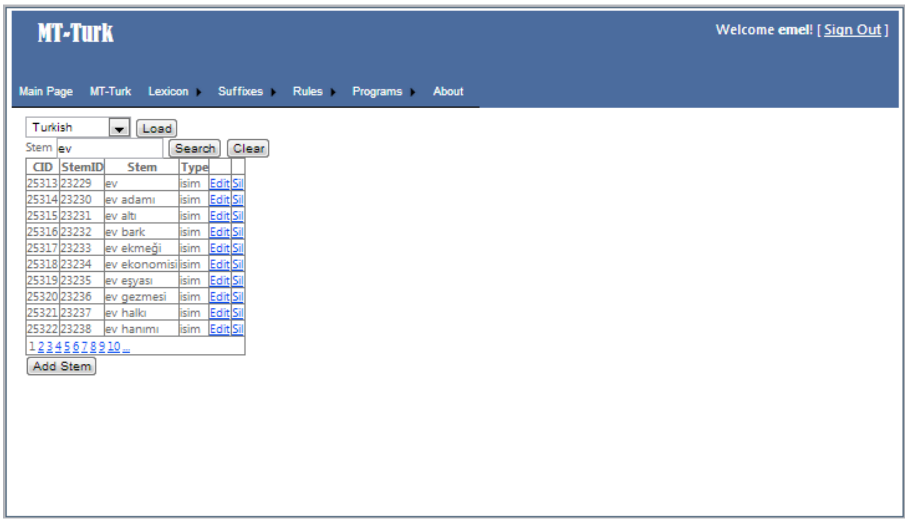
Source Language: Turkish

File: No file chosen

Target Language: Turkish

Figure 3.9 Bulk lexicon uploader

In the lexicon manager, on the other hand, the current stem list of the lexicon is displayed. A stem starting with a string can be searched by the user in addition to editing existing stems and adding a new stem from scratch. The screenshot of lexicon manager for a sample search is given in Figure 3.10.



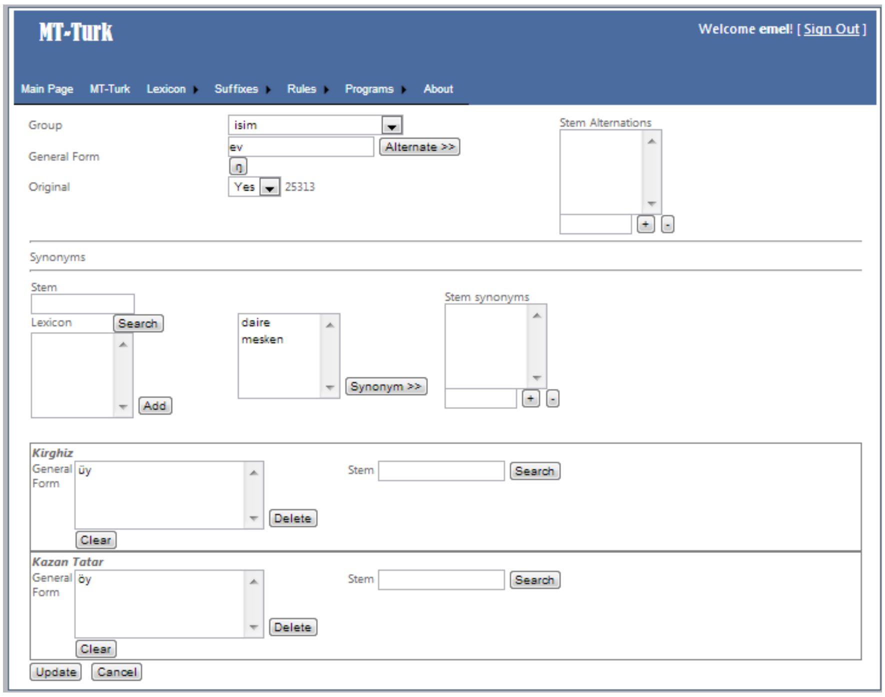
Turkish

Stem: ev

CID	StemID	Stem	Type		
25313	23229	ev	isim	Edit	Sil
25314	23230	ev adamı	isim	Edit	Sil
25315	23231	ev altı	isim	Edit	Sil
25316	23232	ev bark	isim	Edit	Sil
25317	23233	ev ekmeği	isim	Edit	Sil
25318	23234	ev ekonomisi	isim	Edit	Sil
25319	23235	ev eşyası	isim	Edit	Sil
25320	23236	ev gezmesi	isim	Edit	Sil
25321	23237	ev halkı	isim	Edit	Sil
25322	23238	ev hanımı	isim	Edit	Sil

Figure 3.10 Lexicon manager screenshot

The stem information should include corresponding stems in all the languages defined in the system; however, it is not compulsory. The screenshot of the lexicon stem editor for a sample entry is given in Figure 3.11.



Group: isim

General Form: ev

Original: Yes 25313

daire
mesken

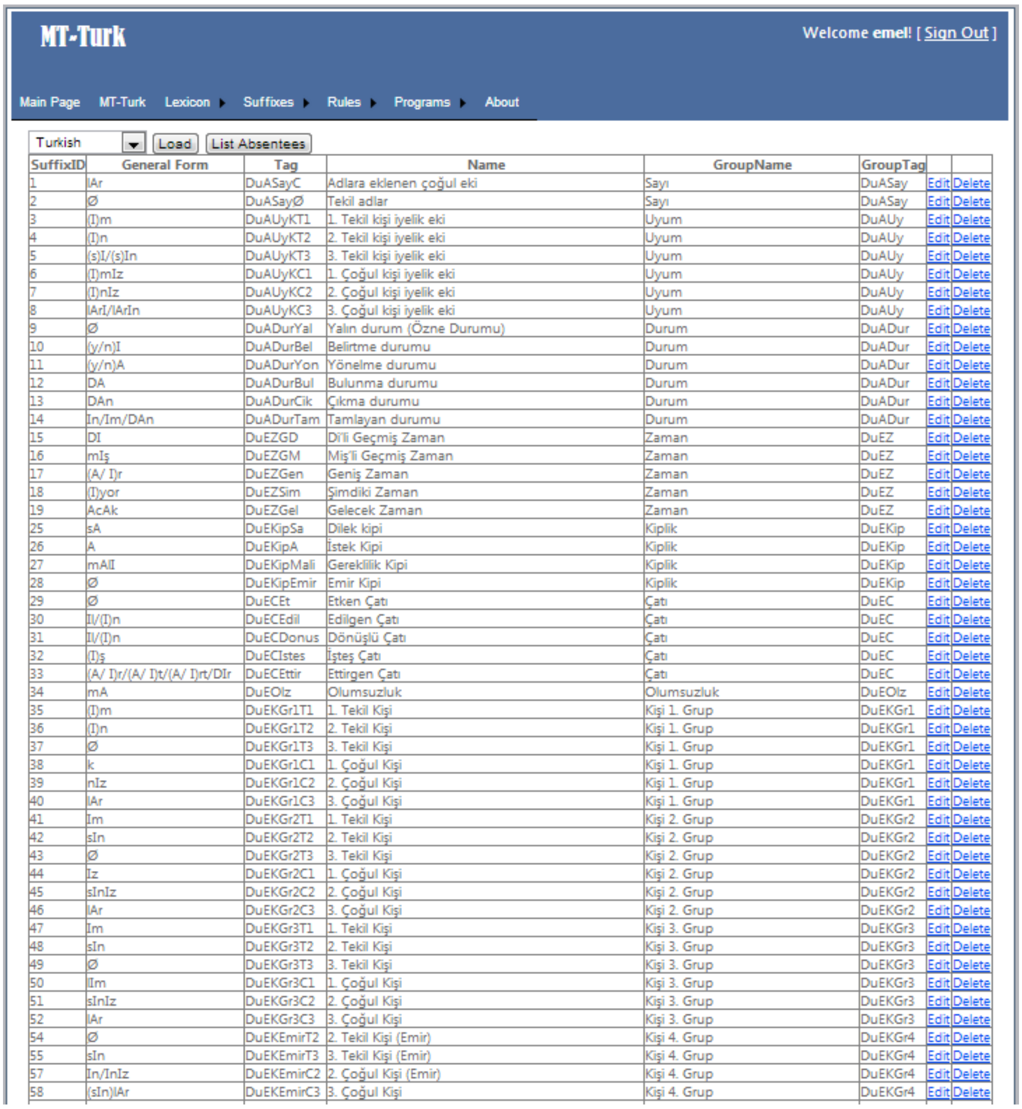
Kirghiz General Form: üy

Kazan Tatar General Form: öy

Figure 3.11 Stem editor screenshot

3.4.2.4 Suffix Manager

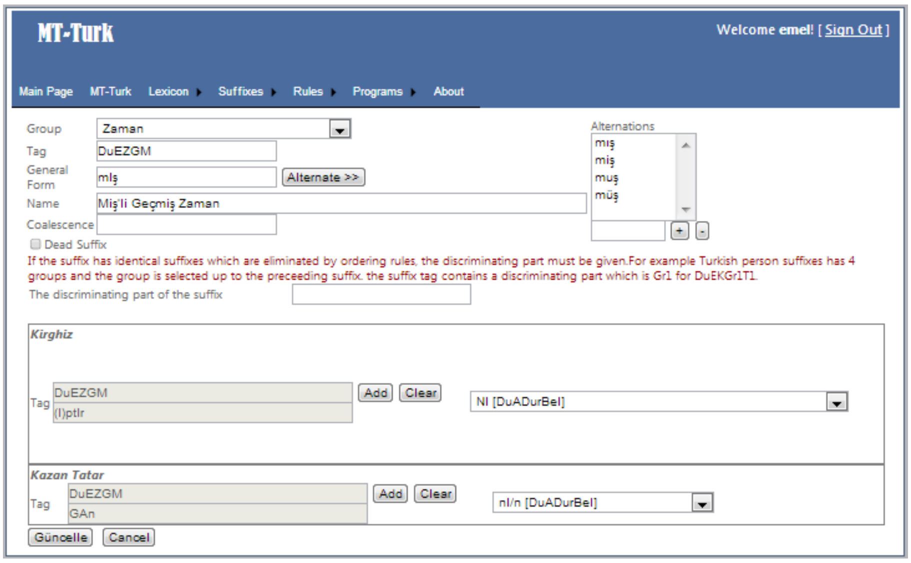
The current suffix list is displayed by the suffix manager. Existing suffixes can be edited by the user and also new suffixes ca be added to the system using suffix manager. The screenshot of the suffix manager is given in Figure 3.12.



SuffixID	General Form	Tag	Name	GroupName	GroupTag		
1	lAr	DuASayC	Adlara eklenen çoğul eki	Sayı	DuASay	Edit	Delete
2	Ø	DuASayØ	Tekil adlar	Sayı	DuASay	Edit	Delete
3	(I)m	DuAUyKT1	1. Tekil kişi iyelik eki	Uyum	DuAUy	Edit	Delete
4	(I)n	DuAUyKT2	2. Tekil kişi iyelik eki	Uyum	DuAUy	Edit	Delete
5	(s)I/(s)In	DuAUyKT3	3. Tekil kişi iyelik eki	Uyum	DuAUy	Edit	Delete
6	(I)mIz	DuAUyKC1	1. Çoğul kişi iyelik eki	Uyum	DuAUy	Edit	Delete
7	(I)nIz	DuAUyKC2	2. Çoğul kişi iyelik eki	Uyum	DuAUy	Edit	Delete
8	lArI/lArIn	DuAUyKC3	3. Çoğul kişi iyelik eki	Uyum	DuAUy	Edit	Delete
9	Ø	DuADurYal	Yalın durum (Özne Durumu)	Durum	DuADur	Edit	Delete
10	(y/n)I	DuADurBel	Belirtme durumu	Durum	DuADur	Edit	Delete
11	(y/n)A	DuADurYon	Yönelme durumu	Durum	DuADur	Edit	Delete
12	DA	DuADurBul	Bulunma durumu	Durum	DuADur	Edit	Delete
13	DAn	DuADurCik	Çıkma durumu	Durum	DuADur	Edit	Delete
14	In/Im/DAn	DuADurTam	Tamlayan durumu	Durum	DuADur	Edit	Delete
15	DI	DuEZGD	Di'li Geçmiş Zaman	Zaman	DuEZ	Edit	Delete
16	mIş	DuEZGM	Miş'li Geçmiş Zaman	Zaman	DuEZ	Edit	Delete
17	(A/ I)r	DuEZGen	Geniş Zaman	Zaman	DuEZ	Edit	Delete
18	(I)yor	DuEZSim	Şimdiki Zaman	Zaman	DuEZ	Edit	Delete
19	AcAk	DuEZGel	Gelecek Zaman	Zaman	DuEZ	Edit	Delete
25	sA	DuEKipSa	Dilek kipi	Kiplik	DuEKip	Edit	Delete
26	A	DuEKipA	İstek Kipi	Kiplik	DuEKip	Edit	Delete
27	mAlI	DuEKipMali	Gereklilik Kipi	Kiplik	DuEKip	Edit	Delete
28	Ø	DuEKipEmir	Emir Kipi	Kiplik	DuEKip	Edit	Delete
29	Ø	DuECEt	Etken Çatı	Çatı	DuEC	Edit	Delete
30	Il/(I)n	DuECEdil	Edilgen Çatı	Çatı	DuEC	Edit	Delete
31	Il/(I)n	DuECDonus	Dönüşlü Çatı	Çatı	DuEC	Edit	Delete
32	(I)ş	DuECIstes	İşteş Çatı	Çatı	DuEC	Edit	Delete
33	(A/ I)r/(A/ I)t/(A/ I)rt/DIr	DuECEttir	Ettirgen Çatı	Çatı	DuEC	Edit	Delete
34	mA	DuEOlz	Olumsuzluk	Olumsuzluk	DuEOlz	Edit	Delete
35	(I)m	DuEKGr1T1	1. Tekil Kişi	Kişi 1. Grup	DuEKGr1	Edit	Delete
36	(I)n	DuEKGr1T2	2. Tekil Kişi	Kişi 1. Grup	DuEKGr1	Edit	Delete
37	Ø	DuEKGr1T3	3. Tekil Kişi	Kişi 1. Grup	DuEKGr1	Edit	Delete
38	k	DuEKGr1C1	1. Çoğul Kişi	Kişi 1. Grup	DuEKGr1	Edit	Delete
39	nIz	DuEKGr1C2	2. Çoğul Kişi	Kişi 1. Grup	DuEKGr1	Edit	Delete
40	lAr	DuEKGr1C3	3. Çoğul Kişi	Kişi 1. Grup	DuEKGr1	Edit	Delete
41	Im	DuEKGr2T1	1. Tekil Kişi	Kişi 2. Grup	DuEKGr2	Edit	Delete
42	sIn	DuEKGr2T2	2. Tekil Kişi	Kişi 2. Grup	DuEKGr2	Edit	Delete
43	Ø	DuEKGr2T3	3. Tekil Kişi	Kişi 2. Grup	DuEKGr2	Edit	Delete
44	Iz	DuEKGr2C1	1. Çoğul Kişi	Kişi 2. Grup	DuEKGr2	Edit	Delete
45	sInIz	DuEKGr2C2	2. Çoğul Kişi	Kişi 2. Grup	DuEKGr2	Edit	Delete
46	lAr	DuEKGr2C3	3. Çoğul Kişi	Kişi 2. Grup	DuEKGr2	Edit	Delete
47	Im	DuEKGr3T1	1. Tekil Kişi	Kişi 3. Grup	DuEKGr3	Edit	Delete
48	sIn	DuEKGr3T2	2. Tekil Kişi	Kişi 3. Grup	DuEKGr3	Edit	Delete
49	Ø	DuEKGr3T3	3. Tekil Kişi	Kişi 3. Grup	DuEKGr3	Edit	Delete
50	lIm	DuEKGr3C1	1. Çoğul Kişi	Kişi 3. Grup	DuEKGr3	Edit	Delete
51	sInIz	DuEKGr3C2	2. Çoğul Kişi	Kişi 3. Grup	DuEKGr3	Edit	Delete
52	lAr	DuEKGr3C3	3. Çoğul Kişi	Kişi 3. Grup	DuEKGr3	Edit	Delete
54	Ø	DuEKEmirT2	2. Tekil Kişi (Emir)	Kişi 4. Grup	DuEKGr4	Edit	Delete
55	sIn	DuEKEmirT3	3. Tekil Kişi (Emir)	Kişi 4. Grup	DuEKGr4	Edit	Delete
57	In/InIz	DuEKEmirC2	2. Çoğul Kişi (Emir)	Kişi 4. Grup	DuEKGr4	Edit	Delete
58	(sIn)lAr	DuEKEmirC3	3. Çoğul Kişi	Kişi 4. Grup	DuEKGr4	Edit	Delete

Figure 3.12 Suffix manager screenshot

The suffix information should include corresponding suffix or suffixes in all the languages defined in the system; however, it is not compulsory. A screenshot of a sample editing information for the suffix "Past tense MIŞ" is given in Figure 3.13.



Group Zaman

Tag DuEZGM

General Form mIş

Name Miş'li Geçmiş Zaman

Coalescence

Alternations
mış
miş
muş
müş

□ Dead Suffix

If the suffix has identical suffixes which are eliminated by ordering rules, the discriminating part must be given.For example Turkish person suffixes has 4 groups and the group is selected up to the preceeding suffix. the suffix tag contains a discriminating part which is Gr1 for DuEKGr1T1.

The discriminating part of the suffix

Kirghiz

Tag DuEZGM
(I)ptIr

NI [DuADurBel]

Kazan Tatar

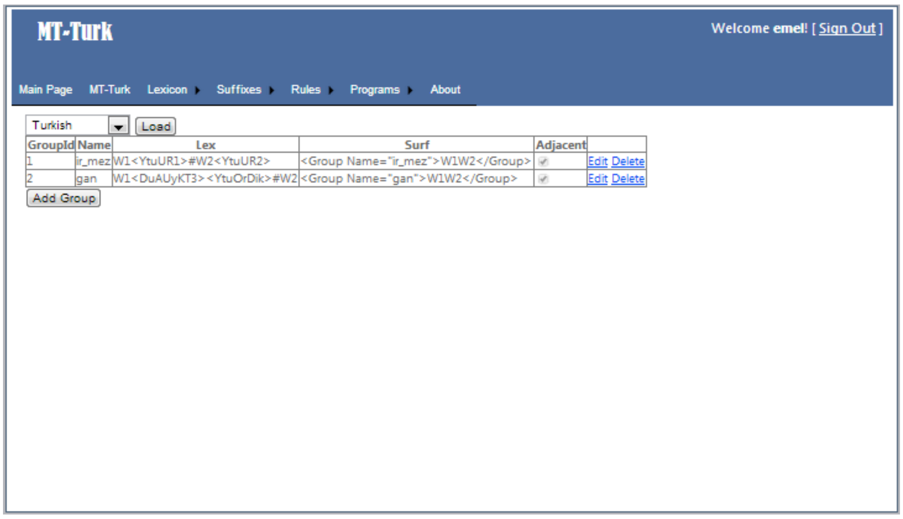
Tag DuEZGM
GAn

nI/n [DuADurBel]

Figure 3.13 Suffix editing screenshot for suffix "past tense MIŞ"

3.4.2.5 Multi-Word Manager

Multi-word expressions of the third form Non-Lexicalized Collocations is managed using multi-word manager interface. The screenshot of the Turkish multi-word manager is shown in Figure 3.14.



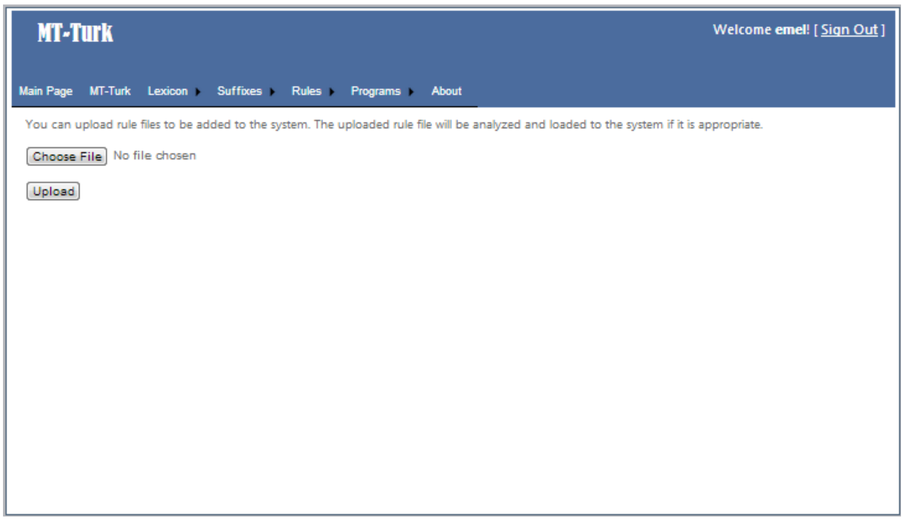
Turkish

GroupId	Name	Lex	Surf	Adjacent	
1	ir_mez	W1<YtuUR1>#W2<YtuUR2>	<Group Name="ir_mez">W1W2</Group>	✓	Edit Delete
2	gan	W1<DuAUyKT3><YtuOrDik>#W2	<Group Name="gan">W1W2</Group>	✓	Edit Delete

Figure 3.14 Multi-word manager screenshot

3.4.2.6 Rule Uploaders

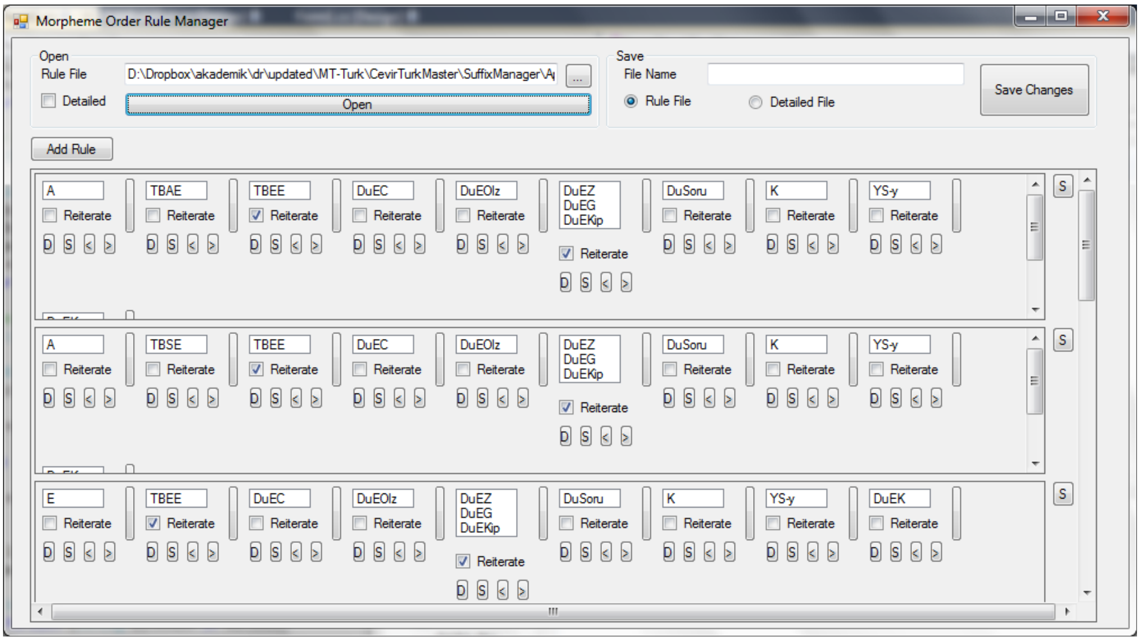
One of the operations that are not available online is morpheme order rule editing as the system requires the rule file to be uploaded as a file using the interface shown in Figure 3.15. Instead of an online editing of rules, morpheme order rule editing windows application can be downloaded using the programs tab and the constructed rule file can be uploaded to the system for analysis and approval. The windows application, which is shown in Figure 3.16, enables flexibility and speed on editing the rules.



You can upload rule files to be added to the system. The uploaded rule file will be analyzed and loaded to the system if it is appropriate.

No file chosen

Figure 3.15 Morpheme rule upload interface



Rule File: D:\Dropbox\akademik\dr\updated\MT-Turk\CevirTurkMaster\SuffixManager\A

□ Detailed

● Rule File ○ Detailed File

Figure 3.16 Morpheme rule editing windows application

3.5 Database Model

As the main aim of this study is to achieve an extensible system which has no pivot language, each language in the system has its own database. Each database consists of tables holding roots, stems, suffixes and their alternations. The alternations for each root, stem and suffix are generated by the phonology library and the used by the analyser.

Currently, there are three languages defined in the system, thus three databases for languages; Turkish, Kirghiz and Kazan Tatar are created. Since the database for Turkish was designed prior to system development (Aktaş & Çebi, 2010; Birant et al., 2010), it is taken as is and databases of other languages are designed depending on this model.

Although separate databases are used, a connection between the databases has to be provided to achieve transfer from one language to another, and it must be realized in two levels: stems and suffixes. This is achieved by a separate database named CONCEPTSET.

The tables in the language databases are:

- *Kok :* holds the list of roots.
- *KokTip :* holds part of speech information of the roots.
- *KokSanal :* holds alternation information of the roots.
- *Govde :* holds the list of stems.
- *GovdeTip :* holds part of speech information of the stems.
- *GovdeSanal :* holds alternation information of the stems.
- *Tip :* holds part of speech list.
- *Ek :* holds the list of suffixes in lexical form.
- *Ekler :* holds all possible alternations of suffixes (surface forms).
- *EkGrup :* holds the group information for the suffixes (for use in morphological analysis)

The entity relationship (ER) diagram of language databases is illustrated in Figure 3.17.

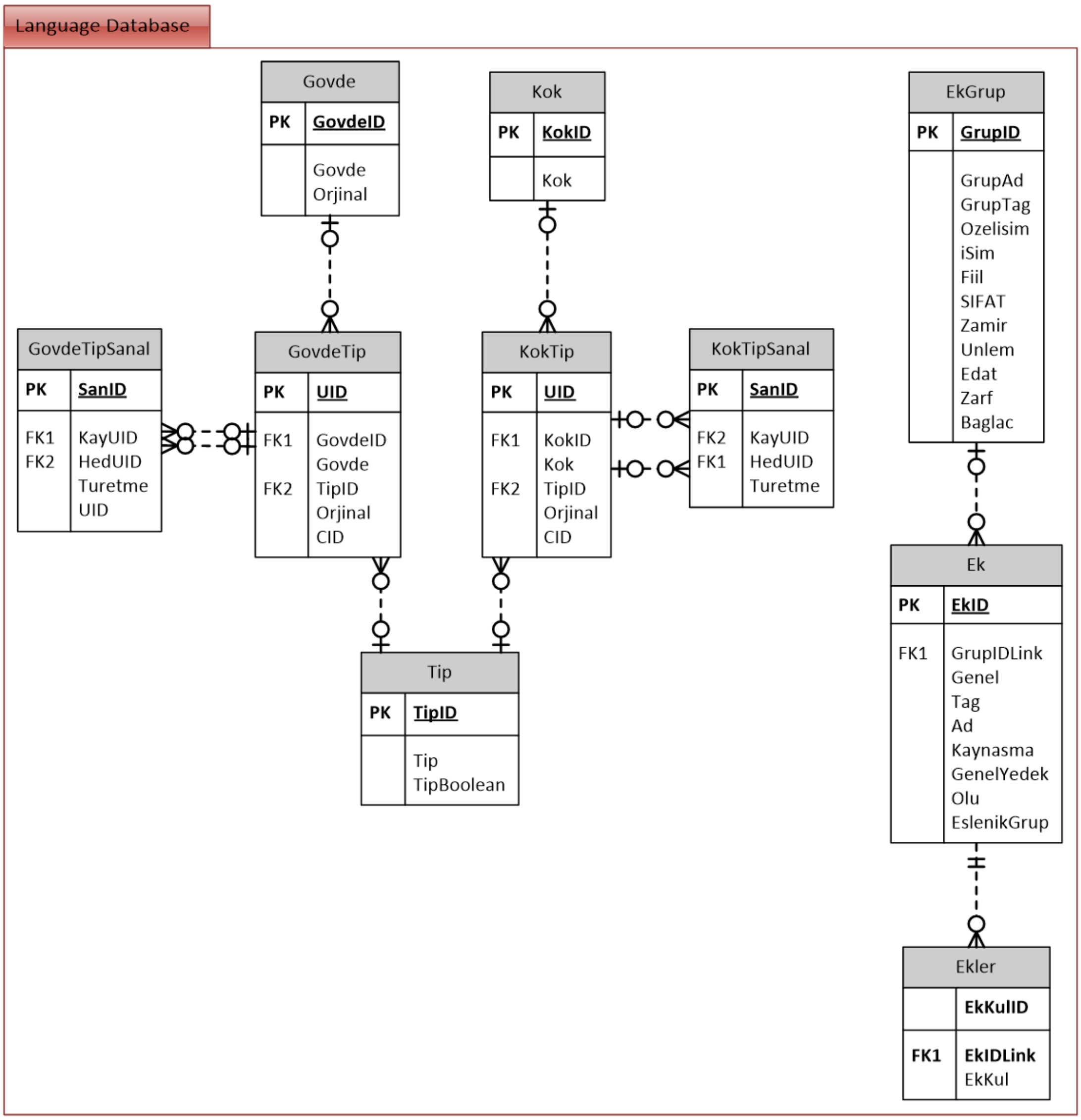


Figure 3.17 ER diagram of language databases

The tables in the CONCEPTSET database are:

- Languages
- Concepts
- ConceptCoverRel
- TagSubstituteRel
- User
- Suggestion

The ER Diagram of CONCEPTSET database is illustrated in Figure 3.18.

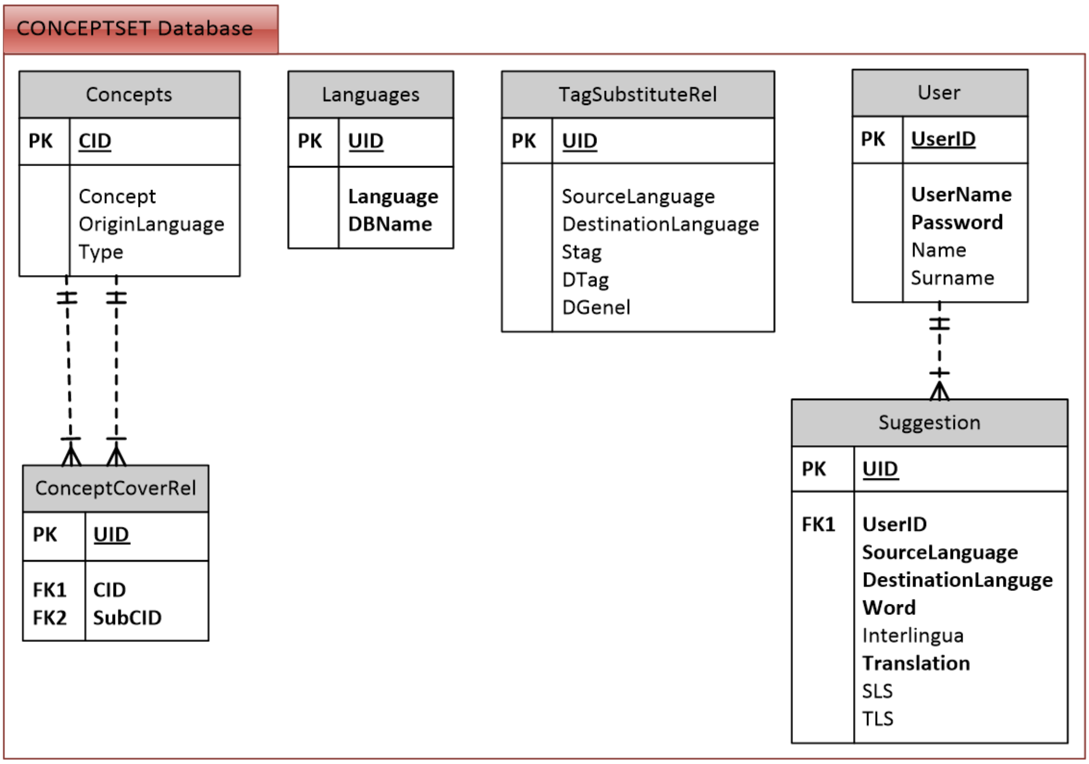


Figure 3.18 ER diagram of CONCEPTSET database

3.5.1 The Table "Languages"

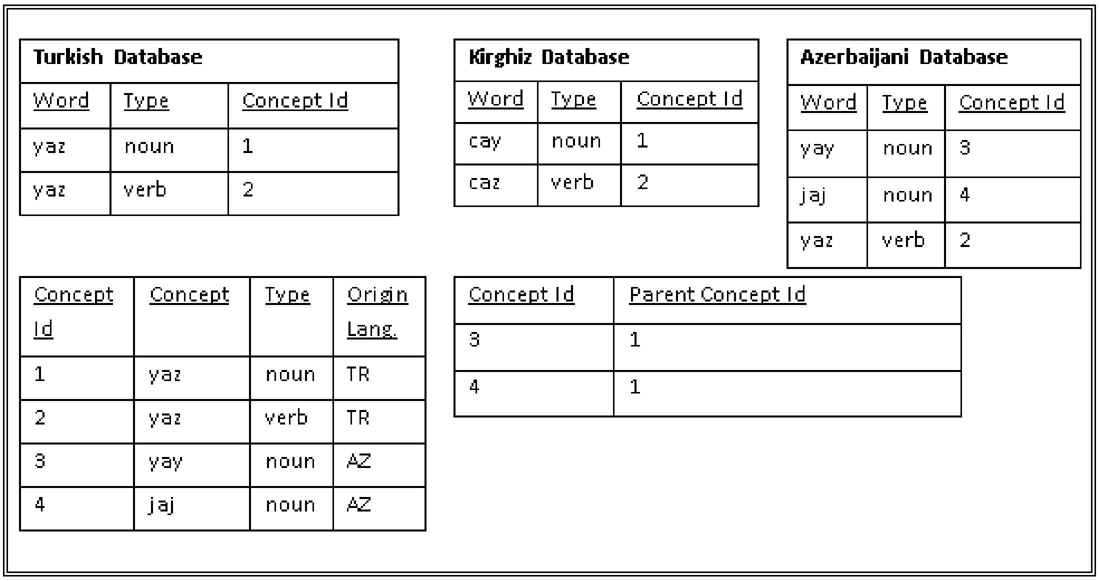
"Languages" table holds the list of available languages in the system with database names. These entries are used for gathering the list of available languages and also the database connection strings are constructed automatically with these values. Therefore, the database connection to the new database can be achieved by a tuple entry to this table with no extra settings or processing required. The current table is shown in Table 3.1.

Table 3.1 Languages database table

Language	DBName
Türkiye Türkçesi	tdkgov_yeni
Kırgız Türkçesi	tdkgov_yeni_kr
Tatar (Kazan) Türkçesi	tdkgov_yeni_tk

3.5.2 The Table “Concepts”

The “Concepts” table holds the list of concepts including the information on which language the concept is introduced by. Thus, the connection between the languages is achieved by the “Concepts” table. A sample view of the table is shown in Figure 3.19.



Turkish Database

Word	Type	Concept Id
yaz	noun	1
yaz	verb	2

Kirghiz Database

Word	Type	Concept Id
cay	noun	1
caz	verb	2

Azerbaijani Database

Word	Type	Concept Id
yay	noun	3
jaj	noun	4
yaz	verb	2

Concept Id	Concept	Type	Origin Lang.
1	yaz	noun	TR
2	yaz	verb	TR
3	yay	noun	AZ
4	jaj	noun	AZ

Concept Id	Parent Concept Id
3	1
4	1

Figure 3.19 Sample representation of the relation between CONCEPTSET and languages

3.5.3 The Table “ConceptCoverRel”

The “ConceptCoverRel” table holds the covering relations over concepts. For the example of word “yaz” given in Table 3.2, the containments of “Concepts” and “ConceptCoverRel” tables are listed in Figure 3.19.

Table 3.2 Different representation of Turkish word "yaz"

Turkish	**Kirghiz**	**Azerbaijani**
yaz (summer)	ay	yay / jaj
yaz (write)	caz-	yaz-

3.5.4 The Table “TagSubstituteRel”

Each suffix tag in a language has a corresponding tag or tag sequence in each language that is defined in the system. The “TabSubstituteRel” table holds a mapping of this correspondence with transfer rules. The representations of two sample suffix correspondence rules are illustrated in Table 3.3.

Table 3.3 Tag correspondence relation

SourceLanguage	**DestinationLanguage**	**STag**	**DTag**
KR	TR	DuEZG2	DuEZGM+KDi
KR	TR	DuEZG4	DuEZGen+KDi

The first rule represents that DuEZG2 - *Type Two Past Tense* in Kirghiz is represented by combination of two suffixes (DuEZGM - *past tense “mIş”* +KDi – *copula* “DI”) in Turkish.

3.5.5 The Table “User”

The user table holds the list of system user with a username, name, surname and a password as illustrated in Table 3.4.

Table 3.4 User table

Username	**Name**	**Surname**	**Password**
emel	Emel	ALKIM	**********
yalcin	Yalçın	ÇEBİ	**********

3.5.6 The Table “Suggestion”

The suggestions supplied by the users are stored in the system with the context information. The information consists of the id of the user, the word, the result of the analysis (the interlingua form), the selected translation, the sentence in which the suggested word occurs, and the suggested translated form of the sentence.

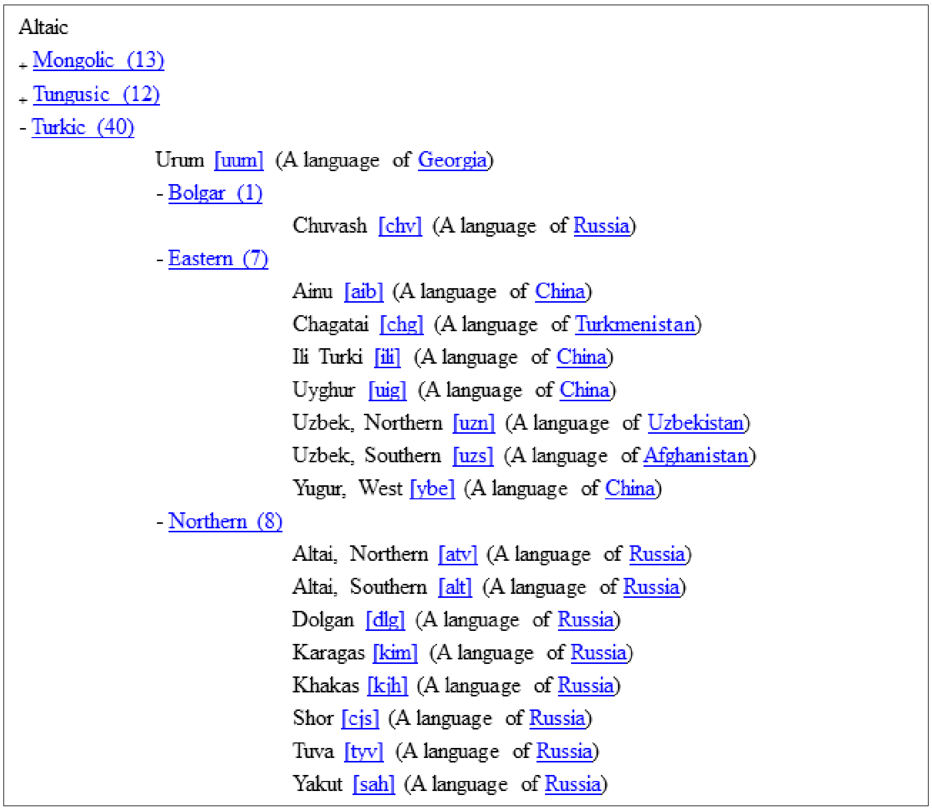
During the translation; when there is a disambiguation in a word that was suggested before, the suggestions are shown to the user. The data stored for one suggestion is given in Table 3.5.

Table 3.5 Sample suggestion information

User	Word	Intermediate Form	Translation	SLS (Source Language Sentence)	TLS (Target Language Sentence)
emel	gördü	<root>gör</root> <suffixes> <DuEG>DI</DuEG> </suffixes>	boldu	Kadın yolcunun sözünü makul gördü	Katın colooçunun sözünü makul boldu

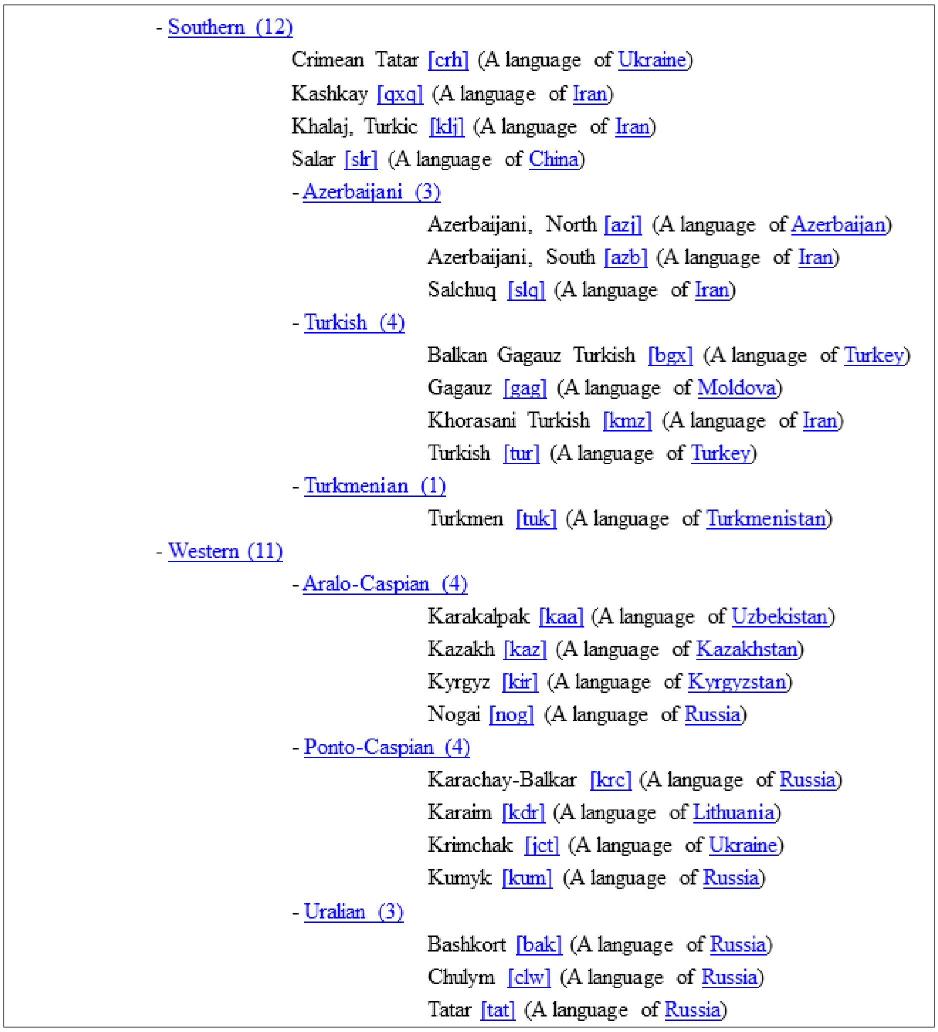
CHAPTER FOUR
GRAMMATICAL CHARACTERISTICS OF TURKISH, KIRGHIZ AND KAZAN TATAR

Turkic language family belongs to the Ural-Altaic group (Bozkurt, 2002) and consists of 40 languages (M. P. Lewis, 2009). The languages of the Turkic language family, which are listed in Figure 4.1 and Figure 4.2, are closely related to each other. Thus, they are similar in their structural and semantic properties.



Altaic
+ Mongolic (13)
+ Tungusic (12)
- Turkic (40)
    Urum [uum] (A language of Georgia)
    - Bolgar (1)
        Chuvash [chv] (A language of Russia)
    - Eastern (7)
        Ainu [aib] (A language of China)
        Chagatai [chg] (A language of Turkmenistan)
        Ili Turki [ili] (A language of China)
        Uyghur [uig] (A language of China)
        Uzbek, Northern [uzn] (A language of Uzbekistan)
        Uzbek, Southern [uzs] (A language of Afghanistan)
        Yugur, West [ybe] (A language of China)
    - Northern (8)
        Altai, Northern [atv] (A language of Russia)
        Altai, Southern [alt] (A language of Russia)
        Dolgan [dlg] (A language of Russia)
        Karagas [kim] (A language of Russia)
        Khakas [kjh] (A language of Russia)
        Shor [cjs] (A language of Russia)
        Tuva [tyv] (A language of Russia)
        Yakut [sah] (A language of Russia)

Figure 4.1 Turkic language family (a) (SIL International, 2013)



- Southern (12)

Crimean Tatar [crh] (A language of Ukraine)
Kashkay [qxq] (A language of Iran)
Khalaj, Turkic [klj] (A language of Iran)
Salar [slr] (A language of China)

- Azerbaijani (3)

Azerbaijani, North [azj] (A language of Azerbaijan)
Azerbaijani, South [azb] (A language of Iran)
Salchuq [slq] (A language of Iran)

- Turkish (4)

Balkan Gagauz Turkish [bgx] (A language of Turkey)
Gagauz [gag] (A language of Moldova)
Khorasani Turkish [kmz] (A language of Iran)
Turkish [tur] (A language of Turkey)

- Turkmenian (1)

Turkmen [tuk] (A language of Turkmenistan)

- Western (11)

- Aralo-Caspian (4)

Karakalpak [kaa] (A language of Uzbekistan)
Kazakh [kaz] (A language of Kazakhstan)
Kyrgyz [kir] (A language of Kyrgyzstan)
Nogai [nog] (A language of Russia)

- Ponto-Caspian (4)

Karachay-Balkar [krc] (A language of Russia)
Karaim [kdr] (A language of Lithuania)
Krimchak [jct] (A language of Ukraine)
Kumyk [kum] (A language of Russia)

- Uralian (3)

Bashkort [bak] (A language of Russia)
Chulym [clw] (A language of Russia)
Tatar [tat] (A language of Russia)

Figure 4.2 Turkic language family (b) (SIL International, 2013)

Turkic Languages are spread over a large geographical area in eastern Europe and Central and North Asia; ranging from the Balkans to the Great Wall of China and from central Iran (Persia) to the Arctic Ocean (“Turkic languages,” 2012). Figure 4.1 and Figure 4.2 also show the countries in which the languages are used. Table 4.1 shows the usage statistics of the most widely used Turkic languages.

Table 4.1 Turkic languages' usage statistics (M. P. Lewis, 2009)

Language	Lang. Code	Usage Area (Widely)	Usage Popularity (Appr.)
Turkish	TRK	Turkey	72 million
Azerbaijani	AZB	Iran	24,3 million
Azerbaijani	AZE	Azerbaijan	7 million
Turkmen	TCK	Turkmenistan	6,4 million
Kazakh	KAZ	Kazakhstan	8 million
Kirghiz	KDO	Kyrgyzstan	2,6 million
Uyghur	UIG	China	7,6 million
Uzbek	UZB	Uzbekistan	18,5 million
Uzbek	UZS	Afghanistan	1.4 million
Chuvash	CJU	Russia	2 million
Bashkir	BXK	Russia	1 million

Turkic languages come from the same origin and have changed in time as a result of spreading in a large geographical area and being influenced by other languages. Hence, nearly all Turkic languages share the same phonology, morphology and syntax structure except Chuvash, Khalaj, Yakut and Dolgan which show different characteristics.

The most significant property of Turkic languages is that they are agglutinative languages in which the words are formed by adding affixes to a root. Therefore, a single word can represent a whole sentence and the morpho-syntactical information is very important for analyzing and translating the text. Such an example where a Turkish word with twelve suffixes forms an English sentence with thirteen words is:

Çekoslavakyalılaştıramadıklarımızdansınız.
Eng: You are one of those that we could not turn into a Checkoslavakian.

4.1 Turkish

Turkish, which is from the southern branch of Turkic languages, is the most widely used Turkic language and the thirteenth most spoken language in the world (Jonsay, 2013). Turkish is also considered to be the most developed Turkic language (Bozkurt, 2002).

4.1.1 Alphabet

The current Turkish alphabet is a slightly modified Latin alphabet which does not contain *Qq*, *Ww, Xx* and modified versions of seven existing letters, *Çç, Ğğ, ı, İ, Şş, Öö, Üü,* are added. It was formed with the aim of representing Turkish pronunciation as accurate as possible and accepted during the alphabet reform in 1928. The alphabet, which is consisted of 29 letters: 21 consonants and 8 vowels, is given in Table 4.2.

Table 4.2 Turkish alphabet

Letter	**Type**
Aa	Vowel
Bb	Consonant
Cc	Consonant
Çç	Consonant
Dd	Consonant
Ee	Vowel
Ff	Consonant
Gg	Consonant
Ğğ	Consonant
Hh	Consonant
Iı	Vowel
İi	Vowel
Jj	Consonant
Kk	Consonant
Ll	Consonant
Mm	Consonant
Nn	Consonant
Oo	Vowel
Öö	Vowel
Pp	Consonant
Rr	Consonant
Ss	Consonant
Şş	Consonant
Tt	Consonant
Uu	Vowel
Üü	Vowel
Vv	Consonant
Yy	Consonant
Zz	Consonant

In Turkish, morphophonemic rules are defined according to the classes of vowels and consonants. Hence, the groups of vowels and consonants, which are used in the morphophonemic rules, are given below.

4.1.1.1 Vowels

The vowels in Turkish are grouped according to roundedness of the lips, the frontness of the tongue and amount of space left between tongue and palate (Göksel & Kerslake, 2005; G. L. Lewis, 1967). The vowels of Turkish are represented in groups in Table 4.3.

Table 4.3 Turkish vowels (G. L. Lewis, 1967)

	Unrounded		**Rounded**	
	Open	**Close**	**Open**	**Close**
Back	a	ı	o	u
Front	e	i	ö	ü

Some characters are used as alternations of one another and form allomorphs of a suffix. An uppercase letter is the common way to show such character groups; either vowel or consonant. The choice of which character will be used is achieved by the harmony rules. The allomorph vowels of Turkish are: *A* for *a* and *e*; *I* for *ı, i, u and ü*.

4.1.1.2 Consonants

The consonants in Turkish are grouped in terms of whether they are voiced or voiceless, their point of articulation and their manner of articulation; however, the former is the most significant in phonological and morphological processes (Göksel & Kerslake, 2005).

The voiceless consonants are *p, ç, t, k, s, ş, f, h* and voiced consonants are *b, c, d, g, ğ, j, l, m, n, r, v, y, z*. Four of the voiceless consonants have voiced equivalents: *p-b, ç-c, t-d, k-ğ* or *g*; but just *ç-c* and *t-d* occur in suffixes. Hence, the allomorph consonants for expressing these alternations are: *C* for *c* and *ç*, *D* for *t* and *d*.

4.1.2 Characteristics

The characteristics of Turkish is analysed in two groups according to the relevance to the study: morphophonemic characteristics, morphological and multi-word characteristics. The syntactical characteristics of languages are not analysed because they show the same characteristics. All of the three languages are Subject -

Object - Verb ordered languages. However, the order of the words can be rearranged for "distinguishing new information from background information and making a certain constituent prominent in the discourse" (Göksel & Kerslake, 2005). The subject or object can be moved in the sentence, to the beginning or before the verb, to emphasize. Hence, MT-Turk does not do syntactical analysis; instead, the grouping of words or the suffixes that create a new word when added to the stem are studied.

4.1.2.1 Morphophonemic Characteristics

Turkish is a highly phonological language, so that morphology of it cannot be studied without considering phonology. Each suffix has allomorphs which are selected by the morphophonemic rules. Vowel and consonant harmonies in addition to other phoneme alterations are the main morphophonemic rules.

- ***Vowel Harmony***

In Turkish, each vowel after the first vowel of the word harmonizes by the preceding vowel with only exception of some borrowed words. Moreover, when a suffix is attached to a stem it harmonizes with the vowel at the last syllable of the stem whether the stem is borrowed or not. Vowel harmony is achieved by two assimilations: palatal assimilation and labial assimilation.

  - ***Palatal Assimilation***

Palatal assimilation or front/back harmony is the harmony of front and back vowels. Front vowels (e, i, ö, ü) are followed by front vowels whereas back vowels (a, ı, o, u) are followed by back vowels.

For example, plural suffix in Turkish has two allomorphs *lar* and *ler*. It is used as *ler* when it comes to a stem with a front vowel at the last syllable whereas it is used as *lar* after a stem with a back vowel at the last syllable.

kitaplar: *books* (kitap+lar)
ödüller: *prizes* (ödül+ler)

However there are some exceptions to this rule such as some borrowed words, *insan (human), dünya (world),* and some suffixes, -*gil* (home, family), -*leyin* (time of day), -*mtrak* (adjectival suffix), -*ki* (locative), -*ken* (while), -*yor* (Present Continuous Tense).

- ***Labial Assimilation***

Labial assimilation or rounded/unrounded harmony is the harmony of rounded and unrounded vowels. Unrounded vowels are followed by unrounded vowels whereas rounded vowels are followed by either unrounded, open vowels or rounded close vowels. The valid letter sequences for two subsequent syllables of a word are given in Table 4.4.

Table 4.4 Labial assimilation in Turkish

Vowel in the First Syllable	**Vowel in the Next Syllable**
Unrounded vowels a, e, ı, i	Unrounded vowels a, e, ı, i
Rounded vowels o, ö, u, ü	Unrounded, open vowels a, e
	Rounded, close vowels u, ü

- ***Raising***

When one of the suffixes, –()yor, -(y)AcAk, -(y)An, or /y/ is added to a stem ending with an open vowel, the last vowel is assimilated to the close allomorph.

ara + yor > arıyor (*calling*)
söyle + yor > söylüyor (*telling*)

- ***Rounding***

Unrounded vowel becomes rounded when it is between two rounded vowels.

sorma + yor > sormuyor (*is not asking*)
olma+yor > olmuyor (*is not happening*)

- ***Syncope***

If a suffix starting with a vowel is added to a stem that has a close vowel in the second syllable, the close vowel is deleted.

ağız >ağzı (*his mouth*)
burun >burnu (*his nose*)

The rule does not apply to reiterations like:
omuz omuza (shoulder to shoulder)

Moreover; the close vowel in the second syllable is deleted when the stems like devir- (*knock over*), çevir- (*turn*), sıyır- (*scrape*), which have v/y +close vowel +r in their second syllable, are conjugated with –i, -im, -inti, -ik derivation suffixes or voice suffix "il".

devir+im > devrim (*revoluation*)
sıyır+ıl> sıyrıl- (*wriggle*)

- ***Consonant Harmony***

The consonant groups voiced and voiceless is the basis for the consonant harmony rules in Turkish. Mainly the rules can be summarized as "the voiceless consonants must be followed by voiceless consonants whereas voiced consonants can be followed by either voiced consonants or vowels". A more detailed representation of the rules, final voicing and devoicing are given with examples below.

  - ***Final Voicing***

When there is a voiceless consonant at the end of a stem and a suffix starting with a vowel is added to the stem, the voiceless consonant changes into the voiced form. Five voiceless consonants which change with final voicing rule are given with samples in Table 4.5.

Table 4.5 Final voicing in Turkish

Change	**Sample Lexical Form**	**Sample Surface Form**
p → b	kitap+I (book + Accusative case)	kita**b**ı
ç → c	ağaç+In (tree + Genitive case)	ağa**c**ı
t → d	kağıt+A (paper + Dative case)	kağı**d**a
k → g if there is consonant *n* before *k*	renk + I (color + Accusative case)	ren**g**i
k→ğ otherwise	ayak + I (foot + Accusative case)	aya**ğ**ı
g → ğ	analog + A (analog + Dative case)	analo**ğ**a

  - ***Devoicing***

The suffixes starting with voiced consonants are assimilated when they are added after a stem ending with a voiceless consonant. Three voiced consonants which are affected by the devoicing rule are given in Table 4.6 with samples.

Table 4.6 Devoicing in Turkish

Change	**Sample Lexical Form**	**Sample Surface Form**
c → ç	kitap+CI (book + Noun to noun derivation)	kitap**ç**ı
d → t	git+DI (go + Past tense with Dİ)	git**t**i
g → k	renk + I (color + Accusative case)	ren**g**i

- ***Apocope***

The last consonant is deleted when –cIk ve –rAk suffixes are added to a stem ending with consonant /k/.

küçük > küçü(k)cük (*tiny*)

The last consonant is also deleted when –(A)l suffix is added to stems like ufak (*little*), alçak (*low*), yüksek (*high*), küçük (*small*)

ufak > ufal- (*shrink*)
yüksel > yüksel- (*rise*)

4.1.2.2 Morphological and Multi-Word Characteristics

The gathered list of suffixes consists of 188 suffixes. Turkish does not consist of any infixes and prefixes except some old borrowed prefixes which are considered in the lexicon as a stem altogether (na-negation suffix e.g. namüsait: *not available*).

The rules on valid suffix combinations are gathered by two members of the Natural Language Processing Research Group in DEU, Özgün Koşaner and Özden Fidan (Fidan & Koşaner, 2007). The rules for noun and verb inflection are given below.

The general rule for noun inflection is:

NounStem>DuASay>DuAUy>DuADur

where;

NounStem	: Noun
DuASay	: Nominal Number
DuAUy	: Nominal Possessive
DuADur	: Nominal Case

The general rule for verb inflection is:

*VerbStem>DuEC>DuEOlz>DuEZ/DuEG/DuEKip>Koşaç>DuEK-**DIr***

or

*VerbStem>DuEC>DuEOlz>Ytu>Koşaç>DuEK-**DIr***

where;

VerbStem	: Verb
DuEC	: Voice
DuEOlz	: Negation
DuEZ	: Nominal Case
DuEG	: Aspect
DuEKip	: Mode
K	: Copula
DuEK	: Personal endings
Ytu	: Subordination

Multi-word characteristics of Turkish can be grouped in two categories: multi-word groups and suffixes that create new words.

- ***Multi-Word Groups***

The multi-word expressions (MWE) in Turkish can be grouped under four types (Oflazer, Çetinoğlu and Say 2004), which are:

  - ***Lexicalized Collocations (Fixed Expressions):*** MWEs are formed with duplication of same word or in a predefined structure. (e.g.: hiç olmazsa: *at least*; ipe sapa gelmez: *nonsensical*)
  - ***Semi-Lexicalized Collocations (Semi-Fixed Expressions):*** MWEs are already stored in the database (e.g.: kafayı ye-: *go nuts*).
  - ***Non-Lexicalized Collocations (Syntactically Flexible Expressions):*** MWEs are formed with use of some suffixes (e.g.: koş-a koş-a: *by running*; uyu-r uyu-maz: *as soon as he sleeps*)
  - ***Multi-Word Named-Entities (Institutionalized Expressions):*** MWEs are proper names (e.g.: Dokuz Eylül Üniversitesi: *Dokuz Eylul University*)

- ***Suffixes that Create New Words***

In Turkish, there is only one suffix that creates a new word when added to a stem: *#mI* question suffix.

kaçacak mı? (will the run away?)
bitti mi? (did it finish?)

4.2 Kirghiz

Kirghiz is from the Aralo-Caspian part of the western branch of Turkic languages. It is the sixth most spoken Turkic language and it is generally used in Kirgizstan.

4.2.1 Alphabet

At the end of 30s, the Turkic Republics in the Soviet Union were each given a different Cyrillic based alphabet. Kirghiz alphabet is one of those alphabets and still in use. The alphabet contains 34 letters, 22 consonants and 8 vowels, and two special characters for an apostrophe and a softening mark. The Cyrillic Kirghiz alphabet is given in Table 4.8 with transliterations in Latin and the types of the letters.

4.2.1.1 Vowels

The vowels in Kirghiz are grouped according to the place of outfall of the sound, status of the lips and openness of the mouth (Çengel, 2005). The vowels of Kirghiz are represented in groups in Table 4.7. The vowel allomorphs for Kirghiz are *A* for *a, e, o* and *ö*; *I* for *ı, i, u and ü*.

Table 4.7 Kirgiz vowels (Çengel, 2005)

	Back		**Front**	
	Unrounded	**Rounded**	**Unrounded**	**Rounded**
Open	a	o	e	ö
Close	ı	u	i	ü

4.2.1.2 Consonants

As in Turkish, the consonants in Kırgız are grouped in terms of whether they are voiced or voiceless, their point of articulation and their manner of articulation. The voiced/voiceless property is the main factor when defining phonemic rules for consonants. The voiceless consonants are *ç, f, x (h), k, p, s, ş, t, ts, tş* whereas voiced consonants are *b, c, d, g, j, l, m, n, ŋ, r, v, y, z*. More information about consonants in Kirghiz can be find in (Çengel, 2005).

In Kirghiz, suffixes can have many allomorphs. The allomorph consonants of Kirghiz are: *B* for *b* and *p*, *G* for *g* and *k*, *L* for *l*, *d* and *t*, *N* for *n*, *d* and *t*.

Table 4.8 Kirghiz alphabet (Çengel, 2005)

Cyrillic	Transliteration in Latin	Type
А а	A a	Vowel
Б б	B b	Consonant
В в	V v	Consonant
Г г	G g	Consonant
Д д	D d	Consonant
Е е	E e / ye	Consonant + Vowel
Ё ё	Yo yo	Consonant + Vowel
Ж ж	J j / C c	Consonant
З з	Z z	Consonant
И и	İ I	Vowel
Й й	Y y	Consonant
К к	K k	Consonant
Л л	L l	Consonant
М м	M m	Consonant
Н н	N n	Consonant
Ң ң	Ŋ ŋ (Ñ ñ in (Öner, 1998))	Consonant
О о	O o	Vowel
Ө ө	Ö ö	Vowel
П п	P p	Consonant
Р р	R r	Consonant
С с	S s	Consonant
Т т	T t	Consonant
У у	U u	Vowel
Ү ү	Ü ü	Vowel
Ф ф	F f	Consonant
Х х	X x (H h in (Öner, 1998))	Consonant
Ц ц	Ts ts	Consonant
Ч ч	Ç ç	Consonant
Ш ш	Ş ş	Consonant
Щ щ	Şc şc	Consonant
Ъ ъ	Apostrophe	
Ы ы	I ı	Vowel
Ь ь	Softening mark	
Э э	E e	Vowel
Ю ю	Yu yu	Consonant + Vowel
Я я	Ya ya	Consonant + Vowel

4.2.2 Characteristics

4.2.2.1 Morphophonemic Characteristics

Kirghiz has vowel and consonant harmony just as Turkish and they are very similar to the versions in Turkish.

- ***Vowel Harmony***

In Kirghiz, vowel harmony is similar to Turkish. It is used both inside the word, except borrowed words, and during suffix attachment; and is achieved by two assimilations: palatal assimilation and labial assimilation.

- ***Palatal Assimilation***

Palatal assimilation is very strong in many Turkic languages and Kirghiz is one of these languages (Öner, 1998). In Kirghiz, just the same as in Turkish, front vowels (*e, i, ö, ü*) are followed by front vowels whereas back vowels (*a, ı, o, u*) are followed by back vowels.

For example, dative case suffix in Kirghiz has eight allomorphs *ka, ke, ko, kö, ga, ge, go* and *gö.* The selection of which allomorph will be used is achieved according to vowel and consonant harmony. If the stem it is added ends with a syllable with a front vowel *ke, kö, ge* or *gö* is used.

kızga: *to the girl* (kız+ga)
küçkö: *to the power* (küç+kö)

However there are some exceptions to this rule such as some borrowed words, *araket (movement), gazeta (newspaper);* some derivation suffixes, *-ek* (derivation suffix "little"), *-ke* (derivation suffix "sweet"), *-tay* (derivation suffix "my sweet"); and some borrowed affixes, *be-/bey-/na-* (negation prefix from Persian), *-iy* (belonging possessive from Arabic) and *-zar* (place name maker from Persian) (Çengel, 2005).

Labial Assimilation

Labial assimilation or rounded/unrounded harmony is the harmony of rounded and unrounded vowels. Unrounded vowels are followed by unrounded vowels whereas rounded vowels are followed by either unrounded, open vowels or rounded close vowels. The valid letter sequences for two subsequent syllables of a word are given in Table 4.9 (Çengel, 2005).

For example, noun derivation suffix "DAş" suffix in Kirghiz has twelve allomorphs *daş, deş, doş, döş, taş, teş, toş, töş, laş, leş, loş* and *löş*. The selection of which allomorph will be used is achieved according to vowel and consonant harmony.

klasstaş: *class mate* (klass+taş)
coldoş: *companion* (col+doş)

Table 4.9 Labial assimilation in Kirghiz

Vowel in the First Syllable	**Vowel in the Next Syllable**
Unrounded vowels a, e, ı, i	Unrounded vowels a, e, ı, i
Rounded, open vowels o, ö	a
	Rounded u, ü, o, ö
Rounded, close vowels u, ü	a
	Rounded (except "o") u, ü, ö

Syncope

If a suffix starting with a vowel is added to a stem that has a close vowel in the second syllable, the close vowel is deleted.

iyin >iyni (*his shoulder*)
ayıl >aylı (*his wife*)

Consonant Harmony

The consonant groups voiced and voiceless is the basis for the consonant harmony rules in Kirghiz. If the stem ends with a voiced consonant, the suffix must start with a voiced consonant whereas if the stem ends with a voiceless consonant the suffix must start with a voiceless consonant. The consonant harmony rules are given with examples below.

- ***Final Voicing***

Similar to Turkish, when there is a voiceless consonant at the end of a stem and a suffix starting with a vowel is added to the stem, the voiceless consonant changes into the voiced form. Voiceless consonants which change with final voicing rule are given with samples in Figure 4.10.

Table 4.10 Final voicing in Kirghiz

Change	**Sample Lexical Form**	**Sample Surface Form**
p → b	kap+alabat (*getting a pot*)	ka**b**alat
k → g	ak + nan (*white bread*)	a**g**nan

- ***b* > *p* *Assimilation***

The suffixes starting with consonant *p* are assimilated when they are added after a stem ending with a voiceless consonant.

bat+per > bat**b**er (appreciate)

4.2.2.2 Morphological and Multi-Word Characteristics

The morphological characteristics of Kirghiz are similar to Turkish. The rules of noun and verb inflections are the same with minor order differences.

Multi-word characteristics of Kirghiz is also grouped and analysed in two categories: multi-word groups and suffixes that create new words. The multi-word groups in Kirghiz are similar to Turkish; however Kirghiz has a much more complicated structure of suffixes that create new words.

- ***Suffixes That Create New Words***

In Kirghiz, the use of auxiliary verbs is quite common to reinforce the meaning. Also there are some suffixes which form additional words when added to a stem.

There are seventeen auxiliary verbs which are: *al, bar, başta, ber, cat, ciber, cür, çık, kal, kel, ket, koy, otur, sal, taşta, tur* and *tüş*. The auxiliary words are added after a word with a gerundial suffix (generally *A, (I)p or y*) and each auxiliary word defines the verb to express a different manner. Some examples are;

kör-ö al-a-t (*he can see*)
urmattay başta-dı (he started to appreciate)

The auxiliary words *cat, cür, tur, otur* are also used in some suffixes like in present tense and future tense.

oku-p cat-a-t (he is reading)
oku-ganı tur-a-t (he will read)

There are also some additional words that are caused by suffixes. Some examples are *ele, boldu, kerek* and *eken*.

ket-e elek (he is going to go)
cürüşüm kerek (*i need to go*)

Moreover, there is no limit to the combinations of these, i.e. one auxiliary verb can be connected with another and also be inflected by a suffix with an additional word.

caz-ba-y tur-gan ele-m (*i would not write*)
tos-up al-ı-ş-ım kerek (*i need to meet them*)

4.3 Kazan Tatar

Kazan Tatar is from the Uralian part of the western branch of Turkic languages. It is the Turkic language with the oldest history with written literature and spoken in a large geographical area (Öner, 1998).

4.3.1 Alphabet

Although several alphabets, Arabic, Latin and Cyrillic, are used by Kazan Tatars; the current alphabet is a Cyrillic alphabet which is defined by Russia in 2002 (Şahin, 2003). The alphabet contains 37 letters, 28 consonants and 9 vowels, and two special characters for an apostrophe and a softening mark. The Cyrillic Kazan Tatar alphabet is given in Table 4.12 with transliterations in Latin and the types of the letters.

4.3.1.1 Vowels

The vowels in Kazan Tatar are grouped according to the place of outfall of the sound, status of the lips and openness of the mouth just like in Turkish and Kirghiz. As it can be seen from the alphabet, there are three letters with the representation *e*. Two of them, Ee and Ээ, are *close e* and is represented as *é* in this study. The vowels of Kazan Tatar are represented in groups in Table 4.11. The allomorph vowels for Kazan Tatar are: *A* for *a* and *e*; *I* for *ı* and *é*.

Table 4.11 Kazan Tatar vowels (Öner, 2007a)

	Back		**Front**	
	Unrounded	**Rounded**	**Unrounded**	**Rounded**
Open	a	o	e	ö
Half Open			é	
Close	ı	u	i	ü

Table 4.12 Kazan Tatar alphabet (Öner, 2007a)

Cyrillic	Transliteration in Latin	Type
А а	A a	Vowel
Б б	B b	Consonant
В в	V v	Consonant
Г г	G g	Consonant
Д д	D d	Consonant
Е е	E e / ye	Consonant + Vowel
Ё ё	Yo yo	Consonant + Vowel
Ж ж	J j	Consonant
З з	Z z	Consonant
И и	İ I	Vowel
Й й	Y y	Consonant
К к	K k	Consonant
Л л	L l	Consonant
М м	M m	Consonant
Н н	N n	Consonant
О о	O o	Vowel
П п	P p	Consonant
Р р	R r	Consonant
С с	S s	Consonant
Т т	T t	Consonant
У у	U u	Vowel
Ф ф	F f	Consonant
Х х	X x	Consonant
Ц ц	Ts ts	Consonant
Ч ч	Ç ç	Consonant
Ш ш	Ş ş	Consonant
Щ щ	Şc şc	Consonant
Ъ ъ	Apostrophe	
Ы ы	I ı	Vowel
Ь ь	Softening mark	
Э э	E e	Vowel
Ю ю	Yu yu	Consonant + Vowel
Я я	Ya ya	Consonant + Vowel
Ә ә	E e	Vowel
Ө ө	Ö ö	Vowel
Ү ү	Ü ü	Vowel
Җ җ	C c	Consonant
Ң ң	Ñ ñ	Consonant
Һ һ	H h	Consonant

4.3.1.1 Consonants

As in Turkish and Kirghiz, the consonants in Kazan Tatar are grouped in terms of whether they are voiced or voiceless, their point of articulation and their manner of articulation; and the first one is taken into consideration for rules.

The voiceless consonants are *p, f, t, s, ç, ş, k, h* whereas voiced consonants are *b, m, v, d, n, c, j, l, r, y, g, z*. The allomorph consonants for expressing the alternations are: *K* for *k* and *g*, *D* for *t* and *d*, *L* for *l* and *n*, *N* for *n*, *t* and *d*.

4.3.2 Characteristics

4.3.2.1 Morphophonemic Characteristics

Kazan Tatar has vowel and consonant harmony just as Turkish and Kirghiz, besides they are very similar.

- ***Vowel Harmony***

In Kazan Tatar, similar to Turkish and Kirghiz, vowel harmony is used both inside the word and during suffix attachment; and is achieved by two assimilations: palatal assimilation and labial assimilation.

- ***Palatal Assimilation***

Palatal assimilation is used in Kazan Tatar as in the other two Turkic languages; Kirghiz and Turkish. In Kazan Tatar, front vowels (e, i, ö, ü) are followed by front vowels whereas back vowels (a, ı, o, u) are followed by back vowels.

For example, the noun derivation suffix *lIk* has two allomorphs *lık* and *lék*. The selection of which allomorph will be used is achieved according to vowel and consonant harmony. If the stem it is added ends with a syllable with a front vowel *lék*, otherwise *lık* is used.

utınlık: *woodshed* (utın+lık)
çüplék: *rubbish dump* (çüp+ lék)

However there are some exceptions to this rule, such as newly borrowed Russian words, *agentlık (being an agent), printsipsızlık (being unprincipled)* (Öner, 1998).

- ***Syncope***

In Kazan Tatar, as in Turkish and Kirghiz, if a suffix starting with a vowel is added to a stem that has a close vowel in the second syllable, the close vowel is deleted.

avız >avzı (*his mouth*)
borın >bornı (*his nose*)

- ***Raising***

When a suffix starting with a vowel is added to a stem with a semi vowel *"y"* at the end, the last vowel assimilated to the close allomorph.

anla+y+a > anlıy (*he is understanding*)
kür+me+y+e> kürmiy (*he doesn't see*)

- ***Consonant Harmony***

The consonant groups voiced and voiceless is the basis for the consonant harmony rules. If the stem ends with a voiced consonant, the suffix must start with a voiced consonant whereas if the stem ends with a voiceless consonant the suffix must start with a voiceless consonant. A more detailed representation of the rules, final voicing and devoicing are given with examples below.

  - ***Final Voicing***

Similar to Turkish and Kirghiz, when there is a voiceless consonant at the end of a stem and a suffix starting with a vowel is added to the stem, the voiceless consonant changes into the voiced form. Voiceless consonants which change with final voicing rule are given with samples in Figure 4.13.

Table 4.13 Final voicing in Kazan Tatar

Change	Sample Lexical Form	Sample Surface Form
p → b	tap+ış (*to meet*)	ta**b**ış
k → g	kük + er (*to green*)	kü**g**er

- ***l > n Assimilation***

When a suffix starting with allomorph *L* is added to a stem with a consonant *m* or *n* at the end, the last vowel assimilated to consonant *n*.

ülen+ler > ülenner (*grasses*)
uram+dan> uramnan (*from the street*)

4.3.2.2 Morphological and Multi-word Characteristics

The morphological characteristics of Kazan Tatar are similar to Turkish and Kirghiz. The rules of noun and verb inflections are the same with minor order differences.

The multi-word groups in Kazan Tatar are similar to Turkish and Kirghiz. Furthermore, Kazan Tatar also makes use of auxiliary verbs and special words in suffix constructions as Kirghiz.

4.4 Differences and Problems

The differences in the structures of the languages make translation harder. Thus, for the success of the study the differences between the languages are studied. Most significant differences which affect the translation performance are suffix conflicts and suffix binding change.

- ***Suffix Conflicts***

One of the problems between Turkic languages is that a suffix in one language can be translated in the other language using a suffix group instead of one corresponding suffix.

Such an example is one of the past tenses in Kirghiz “*GAn*”. The Turkish translation of this suffix is achieved by using two suffixes as “*mIş-DI*” using Past Tense “mIş” and aspect “DI”.

- ***Suffix Binding Change***

Although word order is the same in all Turkic languages, suffixes can require binding to different members of a phrase in different languages. For the example of the phrase “casagan işter**in**: *the jobs he did*” which is formed with a participle, the possessive suffix *in* is added to the noun in Kirghiz whereas in Turkish it should be added to the participle as “yaptığ**ı** işleri”.

CHAPTER FIVE
CASE STUDY: MACHINE TRANSLATION BETWEEN TURKISH, KIRGHIZ AND KAZAN TATAR

MT-Turk infrastructure is tested on three Turkic languages for being able to evaluate cross-lingual translation. In this chapter, initially, the motivation behind the selection of these languages is described. All three languages are modelled in MT-Turk. The crucial information to model and represent a language, specifically in MT-Turk, is grouped in three sections: lexicon, grammar and suffixes. In the remainder of this chapter, the language resources for each language are described including the sources and the quantities. Furthermore, the test data that is used for evaluation is also decribed in this chapter.

5.1 Language Selection

MT-Turk is built upon resources that are retrieved and used in other products of the Natural Language Processing Research Group in Dokuz Eylül University Computer Engineering Department (DEU CSE, 2004). As a result, the first language in the system is Turkish.

The second language which is defined in the system is Kirghiz. The second language selection is made considering the closeness of the languages. Kirghiz and Turkish are not either too close to each other, like Azerbaijani and Turkish which need nearly no translation at all, or too distant, like Cuvash or Yakut which are too distinct from other Turkic dialects. Moreover, the existence of resources was an important factor for the selection of the second language. Two linguistic students, Cıldız Alimova and Darkan Akunbay, were asked for their help during the resource retrieval and translation evaluation phases in addition to help from Prof.Dr. Gürer Gülsevin and linguistic books (Çengel, 2005; Gedikli, 1993).

The third language which is added to the system is Kazan Tatar. The selection of the third language is done by firstly considering resources available, because of the problems that were met during the resource retrieval of the second language, then the closeness of the language. The retrieval of the resources was achieved by the help of Prof. Dr. Mustafa Öner from Ege University and the translation of the test data during the evaluation phase was achieved by Güzel Sabitova.

5.2 Language Resources

Three language resources are needed by MT-Turk for defining a new language to the system: the lexicon, the suffix list and the rules that is used to combine words and morphemes together, the grammar.

5.2.1 Lexicon

Lexicon is the most important resource because if a stem is not stored in the lexicon it cannot be translated. Unfortunately finding reliable and large digital lexicons was a real obstacle.

“A root is the portion of a word that is not further analyzable into meaningful elements, being morphologically simple, and carries the principle portion of meaning of the words in which it functions” (SIL International, 2004a) whereas “A stem consists minimally of a root, but may be analyzable into a root plus derivational morphemes” (SIL International, 2004b). Stems are used for the purposes of translation in this study as the derivated forms can be represented by a different representation in the target language instead of using a corresponding derivation suffix.

- ***Turkish***

The stem list for Turkish is retrieved from TDK (TDK, 2011a) during the development of other projects of the Natural Language Processing Research Group in DEU(DEU CSE, 2004). Two sample stems for Turkish are karyola (*bed*, *bedstead*) and yatak (*bed*, *matress*).

- ***Kirghiz***

The Kirghiz lexicon were initially filled with data from the Comparative Dictionary of Turkic Dialects (Ercilasun, 1992), which was published by The Ministry of Arts and Culture and computerized by Hoca Ahmet Yesevi University (TDK, 2011b). The lexicon was retrieved from Yalçın Özkan at Hoca Ahmet Yesevi University. Unfortunately the dictionary contains only 7398 Turkish words, although the Turkish dataset holds 16430 roots and 70968 stems. The dictionary holds eight different Turkic languages in addition to Turkish and Russian. Each Turkish word can have at most four different corresponding words for each language.

Unfortunately the low coverage of this dictionary decreased the success of the translation severely and also all of the dictionary couldn't be loaded automatically as a result of scanning errors and mapping problems. This caused the search for an additional dictionary source to be crucial. A new dictionary of 7195 entries has been found (Gülensoy & Sağınbayeva, 2004), but unfortunately the dictionary is in hardcopy and the scanning of the dictionary didn't give any good results. The storage of all the roots manually or correction of the scanner (OCR) errors require huge amount of time. Thus the roots which were used in the selected bilingual test data, a subset of the new lexicon data which did not exist in the previous lexicon, have been entered to the system manually.

Two sample stems for Kirghiz are kerebet (*bed, bedstead*) and töşök (*bed, matress*) and they are the correspondences of the sample words given for Turkish: karyola and yatak.

- ***Kazan Tatar***

The lexicon of Kazan Tatar was retrieved from Tatar-Turkish Dictionary (GƏCİL, 2012). The retrieved Kazan Tatar-Turkish lexicon is in digital format and consists of 6986 bilingual entries.

The bilingual data in the retrieved lexicon is transferred to the lexicon by the help of a small program developed for storing bilingual data to the system. However it should be noted that, the design of the lexicon enables cross mapping of the concepts to achieve better mapping of meanings; and the usage of just a bilingual set to store entries in the lexicon lacks the connection to the third language, hence decreasing the completeness of the lexicon.

A sample stem for Kazan Tatar is karawat (*bed*) and it is the correspondence of the sample words given for Turkish and Kazan Tatar: karyola and yatak, kerebet and töşök respectively.

5.2.2 Grammar

The grammar resource required for MT-Turk is grouped in three rules: morphophonemic rules, morpheme order rules and special reordering rules.

- ***Turkish***

The grammar rules of Turkish is retrieved from collaborative studies with the Department of Linguistics (DEU) in the Natural Language Processing Research Group in DEU (DEU CSE, 2004).

- ***Kirghiz***

The second crucial information, the grammar, for Kirghiz is retrieved from Çengel (2005). Çengel (2005) gives detailed information about Kirghiz grammar and language characteristics in addition to sample Kirghiz texts and their translations in Turkish.

- ***Kazan Tatar***

The grammatical information about Kazan Tatar is retrieved from two books: Öner (2007b) and Öner (1998). The phonological rule file for Kazan Tatar is constructed, but it can be enhanced through testing for better results.

5.2.3 Suffix List

The suffix list and the allomorphs of the suffixes are required for each language with the corresponding representations in other languages.

- ***Turkish***

The suffix list of Turkish is gathered together from several resources by two members of the Natural Language Processing Research Group in DEU (DEU CSE, 2004), Özgün Koşaner and Özden Fidan.

- ***Kirghiz***

The suffixes of Kirghiz are retrieved from Çengel (2005) and Öner (1998). Detailed information about Kirghiz suffixes is given in Çengel (2005) whereas in Öner (1998) the suffixes are listed in comparison with Kazan Tatar and Kazakh. Additionally Gedikli (1993) is used as a reference book as it gives information about Turkic dialects with a comparative grammar. The comparison of the suffixes and their usage with examples enables us to get information about the suffix usages in different Turkic dialects.

- ***Kazan Tatar***

The suffix list for Kazan Tatar and the correspondences in Turkish and Kirghiz are gathered together from two books: Öner (2007b) and Öner (1998).

5.3 Test Data for Evaluation

The evaluation is carried out using bilingual texts with two reference translations. The first reference translation is taken from published bilingual texts on Kirghiz and Kazan Tatar. The Ministry of Culture and Tourism has published a collection of literary works on Turkic Languages. The collection contains two volumes on Kirghiz Kültür Bakanlığı (2005a) and Kültür Bakanlığı (2005b); and three volumes on Kazan Tatar Kültür Bakanlığı (2001a), Kültür Bakanlığı (2001b) and Kültür Bakanlığı

(2001c) which are also available online as separate documents (Kültür Bakanlığı, n.d.).

The second reference translation is the translation of the same original texts by a native speaker and gathered by the help of Prof.Dr. Gürer Gülsevin. The original texts are in Kirghiz and Kazan Tatar whereas the translations are in Turkish. More detailed references are given for language pairs below and the texts are given in Appendix B.

5.3.1 Kirghiz to Turkish

Preliminary tests were done on a 35 sentence (258 words) bilingual (Kirghiz to Turkish) text (“Doctor” from a journal) which is translated by a native speaker and gathered by the help of Prof.Dr. Gürer Gülsevin. The original texts are in Kirghiz and the translations are in Turkish.

Five Kirghiz tales from three different authors are used as a secondary basis for translation evaluation from Kirghiz to Turkish. As a second reference translation, the tales were also translated by a native speaker and a linguist Cildiz Alimova. The names of the tales, the authors are listed in Table 5.1 with sentence and word counts.

Table 5.1 Kirghiz - Turkish test data with sentence statistics

Tale Name	Author	Translation 1	Translation 2	Sentence Count	Word Count
Ayıldın Baldarı Village Children	Sagındık Ömürbayev	(Kültür Bakanlığı, n.d.)	Cildiz Alimova	92	1607
Iyık Sezim Sacred Emotion	Aman Saspayev	(Kültür Bakanlığı, n.d.)	Cildiz Alimova	58	1063
İşenböö Not Believing	Şatman Sadıbakasov	(Kültür Bakanlığı, n.d.)	Cildiz Alimova	28	217
Mebel Furniture	Şatman Sadıbakasov	(Kültür Bakanlığı, n.d.)	Cildiz Alimova	44	304
At Cakşı Körgön Bala The boy horse loves	Şatman Sadıbakasov	(Kültür Bakanlığı, n.d.)	Cildiz Alimova	41	301
Total				263	3492

5.3.2 Turkish to Kirghiz

The five Kirghiz tales used at the evaluation of Kirghiz to Turkish translation are also used as the basis for translation from Turkish to Kirghiz as there was no Turkish resource with Kirghiz translation. The translations which are done by a native speaker and a linguist Cildiz Alimova are taken as the source and the translation output is evaluated with the original Kirghiz text. The names of the tales and the authors are listed in Table 5.1 with sentence and word counts.

5.3.3 Kazan Tatar to Turkish

Two Kazan Tatar tales from two different authors are used as the basis for translation from Kazan Tatar to Turkish. Unfortunately, only one of the tales was translated by a native speaker, Güzel Sabitova, as a second reference translation. The names of the tales, the authors are listed in Table 5.2 with sentence and word counts.

Table 5.2 Kazan Tatar - Turkish test data with sentence statistics

Tale Name	**Author**	**Translation 1**	**Translation 2**	**Sentence Count**	**Word Count**
Kiyim Stone Dress	Mehbüpcemal Akçurina	(Kültür Bakanlığı, n.d.)	Güzel Sabitova	127	2388
Ölüf,Yaki Güzel Kız Hediçe (Part I) *Aleph or Beautiful Girl Hatice*	Zahir Bigiyev	(Kültür Bakanlığı, n.d.)	-	63	910
Total				190	3298

5.3.4 Turkish to Kazan Tatar

The two Kazan Tatar tales used at the evaluation of Kazan Tatar to Turkish translation are also used as the basis for translation from Turkish to Kazan Tatar as there was no Turkish resource with Kazan Tatar translation. The translations at the book are taken as the source and the translation output is evaluated with the original Kazan Tatar text. The names of the tales, the authors are listed in Table 5.2 with sentence and word counts.

5.3.5 Kirghiz to Kazan Tatar

There are no available Kirghiz texts with Kazan Tatar translation available or an opportunity to get one; however, there are some Turkish texts in Ercilasun (1992) with translations in Azerbaijani, Bashkir, Kazakh, Kirghiz, Uzbek, Tatar, Turkmen, Uyghur and Russian. The Kirghiz translation is taken as a source and Kazan Tatar is taken as a reference translation. The text is given in Appendix B.

5.3.6 Kazan Tatar to Kirghiz

Just as Kirghiz to Kazan Tatar translation there is no bilingual data available for Kazan Tatar to Kirghiz translation. Hence the texts from Ercilasun (1992) are also used as a basis for Kazan Tatar to Kirghiz translation using Kazan Tatar translation of Turkish text as the source and Kirghiz translation as the reference.

5.4 Evaluation Results

The evaluation is achieved by an evaluation tool which is developed by NIST (National Institute of Standards and Technology) and used in NIST Open Machine Translation (OpenMT) Evaluations which are done since 2001 (NIST (National Institute Of Standards And Technology), 2010).

5.4.1 Kirghiz and Turkish

Although the translation from Kirghiz to Turkish and the translation from Turkish to Kirghiz are evaluated on larger documents, a small sample is given in Table 5.3 for better understanding of the process and evaluation.

Table 5.3 Kirghiz and Turkish translations

İŞENBÖÖ İttin köŋülü cay, büün mışık ekӧö dostoşot. Mından kiyin ekӧö birin-biri körgöndö murundarı tırışpayt, eç kim da alardı bir-birine “kas” dep aytışpayt. Bassa-tursa ele “av-av” deçü it şimşip cürüp tapkan bir kesim mayın aldı da mışıktı izdep cönödü. Bayatan kaparsız küngö kaktangan mışık ittin şıbırtın alda kaydan sezdibi, cerge andan-mından bir tiyip zımıradı.	İNANMAMA Köpeğin gönlü rahat , bugün kedi ikiye barışar. Bundan sonra ikiye birbiri gördüğünde burunları buruşmur , kimse da / de bunları birbiri "düşman" diye bahsetmeyecek . bassa#tursa sanki "avav" derdi köpek , koklayarak buuşmur bir parça yağıyı oyu da / de kediyi aramaya yöneldi . Sürekliden gamsız güneşlenen kedi , köpeğin hışırtıyı oda nereden sezdibi, yıkara oraya andan buradayı bir diyip zıpladı.
İNANMAMA Köpeğin içi rahat, bugün kediyle arkadaş olacak. Bundan böyle birbirin gördüklerinde burun bükmeyecek, kimse de onların “düşman” olduklarını söylemeyecekler. Sürekli havlayan köpek zar zor bulduğu yağ kesimini aldı ve kediyi bulmaya çıktı. Epeydir hiçbir şeyden habersiz güneşlenen kedi köpeğin ayak sesini önceden hissetmiş gibi, birden zıplayarak kaçmaya başladı.	İŞENBÖÖ İttin içsi ırahat , bügün kediyle ekösö dostoşot. Mından kiyin biri-birini körgönlördö burun bükmeyecek , eç kim ondarını kas dep süylöböyt. Baya üren it bulduğu may kesimini aldı cana kediyi bulmaya çıktı . Epeydir eç bir şeyden kabarsız kingö kaktangan mışık ittin but dobuşnu murundan hissetmiş öŋtöl, birden zımırap kaçıp baştadı .

The preliminary test achieved an overall BLEU score 35.08 with a NIST score 4.94 before the suggestion system was activated. The system performed better after the suggestion system was activated and achieved an overall BLEU score 47.77 with a NIST score 5.62. The results of the preliminary evaluation are listed in Table 5.5.

Table 5.4 Preliminary evaluation results of Kirghiz - Turkish translation on 35 sentence text

	Without Suggestion	**With Suggestion**
BLEU	35.08	47.77
NIST	4.94	5.62

However the secondary tests performed badly mainly due to missing lexicon entries. The lexicon is enhanced by manual entries but requires more data. The results of the evaluation with and without suggestion system are listed in Table 5.5.

Table 5.5 Evaluation results of Kirghiz to Turkish translation

	Without Suggestion	**With Suggestion**
BLEU	15.12	21.71
NIST	4.64	4.77

A sample Kirghiz sentence and the translation output in Turkish are given together with the two reference sentences below.

Kapçıgay ördögön cüktüü maşina taş moynoktu aylana berip tık toktodu.	Kirghiz Text
The loaded car that moves at the canyon stopped just after turning the rocky bend.	
Kanyonda ilerleyen yük arabası Taş Moynok tan döner dönmez hemen durdu.	Reference 1
Dağ geçidine doğru ilerleyen yüklü araba, taşlı dönemeçten geçerken tık durdu.	Reference 2
Kanyon ilerledikçe yüklü araç taş dönemeci dönüverip tık tavuktaydı.	No suggestion
Kanyon ilerleyen yüklü araba taş dönemeci dönüverip tık durdu.	With suggestion

When the outputs of the translation is studied, it is seen that the first word “kapçıgay” were translated with the correct stem but missing a suffix, the reason for this is that there is no “kapçı” or “kapçıgay” in any of the dictionaries but the grammar book (Çengel, 2005) contains the word “kapçıgay” and is translated as “kanyon” (canyon), thus “kapçıgay” is stored in the lexicon as “kanyon”. The second, third, fourth and fifth words were translated correctly. The sixth word is translated with a different suffix and although both of the translators used *-ten*, the suffix correspondent of *-dı* (accusative) in Kirghiz is listed as *-I* in Turkish by the grammar books. The word group “aylana berip” is translated as “dönüverip” as “ber” is the auxiliary verb which has the correspondent “-iver” in Turkish. The last two words are translated correctly.

Turkish to Kirghiz translation performance is lower than the Kirghiz to Turkish due to higher number of lexicon in Turkish and also as the manual entries to the lexicon was formed by the correspondences of Kirghiz words, not focusing on forming a full lexicon. Furthermore, the BLEU score is also affected badly from the fact that there is only one reference to evaluate with. The results of the evaluation are listed in Table 5.6.

Table 5.6 Evaluation results of Turkish to Kirghiz translation

	Without Suggestion	**With Suggestion**
BLEU	8.65	12.34
NIST	3.57	4.48

5.4.2 Kazan Tatar and Turkish

Kazan Tatar to Turkish translation scores are slightly lower than the Kirghiz scores as a result of a lower coverage of the lexicon and also suffix entities. The results of the evaluation are listed in Table 5.7.

Table 5.7 Evaluation results of Kazan Tatar to Turkish translation

	Without Suggestion	**With Suggestion**
BLEU	9.52	14.87
NIST	3.60	4.63

As for Turkish to Kirghiz, the Turkish to Kazan Tatar achieved lower scores that Kazan Tatar to Turkish. . Furthermore, similar to Turkish to Kirghiz, the BLEU score is also affected badly from the fact that there is only one reference to evaluate with. The results of the evaluation with suggestion system are listed in Table 5.8.

Table 5.8 Evaluation results of Turkish to Kazan Tatar translation

	Without Suggestion	**With Suggestion**
BLEU	5.04	7.20
NIST	3.12	3.52

5.4.3 Kirghiz and Kazan Tatar

Kirghiz to Kazan Tatar translation scores are slightly higher than Turkish to Kazan Tatar due to closeness of the languages. The results of the evaluation are listed in Table 5.9.

Table 5.9 Evaluation results of Kirghiz to Kazan Tatar translation

	Without Suggestion	**With Suggestion**
BLEU	13.46	18.14
NIST	4.59	4.69

Kazan Tatar to Kirghiz translation achieved slightly lower scores that Kirghiz to Kazan Tatar as Kirghiz lexicon is larger. The results of the evaluation system are listed in Table 5.10.

Table 5.10 Evaluation results of Kazan Tatar to Kirghiz translation

	Without Suggestion	With Suggestion
BLEU	14.23	20.19
NIST	4.62	4.76

5.4.4 Comparison With Similar Studies

BLEU is a metric which is independent from the source language, and although there are studies reporting BLEU scores are not sufficient enough to be used as a comparison technique between machine translation systems (Zhang, Vogel, & Waibel, 2004) and although BLEU scores are not very efficient for agglutinative languages as a mistranslated suffix can produce a total mismatch (Tantuğ, 2007b), some studies on Turkish that reported their BLEU scores are:

- An English-to-Turkish statistical machine translation system achieved a BLEU score of 27.64. They proposed a tool for computing BLEU score of evaluating morphologically rich languages more accurately, BLEU+, and computed an improved BLEU score of 33.03 (Tantuğ, Oflazer, & El-Kahlout, 2008).
- Another study on Turkish, a Turkmen-to-Turkish machine translation system achieved a BLEU score of 33 and BLEUr score (an improved BLEU score for morphologically rich languages) of 38 (Tantuğ et al., 2009).

Furthermore, for the purposes of checking the integrity of the BLEU scores, the translations of the original Kirghiz text were evaluated with reference to each other, selecting the second reference as the candidate translation and the BLEU score was evaluated as 10.31 whereas NIST score was evaluated as 3.69. Although it must be noted that the BLEU score is low as there is only one reference translation, it is an indication of how the BLEU scores are insufficient for evaluating translation between Turkic dialects.

CHAPTER SIX
CONCLUSION

Machine translation is one of the first application areas of computational linguistics; nevertheless there is still work to be done as it is a really hard task to achieve. However, it is more successful in closely related languages.

In the scope of this thesis, an infrastructure for a rule-based translation system between Turkic languages (MT-Turk) is designed and implemented. MT-Turk is designed using transfer approach in combination with a semi-interlingua approach for automatic machine translation.

Translation is achieved in three main levels in MT-Turk, analysis, transfer and generation. The first step to analysis is sentence separator. Then, each sentence is analysed by multi-word expression pre-processor and multi-word expressions are extracted and combined together. The morphological analyser is the final step of the analysis and a semi-interlingua is produced as the output. In transfer level, the stem and suffix replacements and morpheme order modifications are achieved. At the final level the output text is generated according to the constraints of target language.

MT-Turk is tested on Turkish, Kirghiz and Kazan Tatar. Kirghiz was selected as the second language by means of its closeness to Turkish. Kirghiz and Turkish are not either too close to each other or too distant. Considering the problems occurred during the resource retrieval of Kirghiz and closeness to Turkish, Kazan Tatar was selected as the third language.

The grammatical information and lexicon for three Turkic languages, Turkish, Kirghiz and Kazan Tatar, are gathered from two digital and one hardcopy dictionary and different grammar books in collaboration with researchers, and stored in MT-Turk infrastructure. The digital dictionary for Kirghiz includes 7398 stems, whereas the hardcopy includes 7195 stems. Although current size of the Kirghiz lexicon is

5407 stems, only 2514 of them could be stored automatically and the rest of the lexicon was stored manually. The size of the available digital Kazan Tatar lexicon is 6986 stems. Likewise Kirghiz, although the current size of the Kazan Tatar lexicon is 4515 stems, only 3047 stems could be stored automatically and the rest was stored manually.

The success of the translation is evaluated on bilingual texts using two metrics, BLEU (Papineni et al., 2001) and NIST (Doddington, 2002a). The bilingual texts are retrieved from Kültür Bakanlığı (n.d.) and are translated by native speaker linguists as a second reference translation. The original texts for Kirgiz-Turkish and Kazan Tatar-Turkish translations are in Kirghiz and Kazan Tatar respectively and the translations are in Turkish. The Kirghiz evaluation set is consisted of 263 sentences and 3492 words whereas Kazan Tatar evaluation set is consisted of 127 sentences and 2388 words. The texts used for Kirgiz-Kazan Tatar evaluation are originally in Turkish and translations are in Kirghiz and Kazan Tatar. The Kirghiz-Kazan Tatar evaluation set is consisted of 17 sentences and 98 words.

In MT-Turk, when Kirghiz is selected as the source language, BLEU scores of 15.12 and 13.46 with unsupervised translation and 21.71 and 18.14 with semi-supervised translation were obtained for target languages Turkish and Kazan Tatar respectively. When Kazan Tatar is selected as the source language, BLEU scores of 9.52 and 14.23 with unsupervised translation and 14.87 and 20.19 with semi-supervised translation were obtained for target languages for Turkish and Kirghiz respectively. Furthermore, when Turkish is selected as the source language, BLEU scores of 8.65 and 5.04 with unsupervised translation and 12.34 and 7.20 with semi-supervised translation were obtained for target languages Kirghiz and Kazan Tatar respectively.

Consequently, the semi-supervised suggestion system increases the success of the translation. For translation between Turkic languages, there is only one reported BLEU score on a large set, it is a one way translation from Turkmen to Turkish and is reported as 33. There is no reported score to compare on Kirghiz or Kazan Tatar and the success of the translation is affected by the characteristics and closeness of the languages. Besides, it should be noted again that the two translations of the same Kirghiz text to Turkish that are made by human translators performed a BLEU score of 10.31 and this is an indication of how poor BLEU metric performs on agglutinative Turkic languages on the basis of Kirghiz-Turkish language pair.

The idea of a perfect machine translation system with no input from user is seen impossible. It is a fact that even human translation can fail as translation depends on one’s background knowledge which also explains the BLEU score between two reference translations. Therefore, pragmatic knowledge should be stored on the computer in addition to knowledge of the current topic to achieve a perfect translation. However, when the aim is to communicate or get a grip of what is happening in the world, in most cases a translation that is not perfect is also acceptable. If required, the translation can be improved by post editing processes.

In MT-Turk, all possible translations are listed instead of choosing one with disambiguation techniques as some ambiguities cannot be resolved at sentence level and also as most of the disambiguation studies require a corpus, which reveals a problem for under resourced languages like Kirghiz and Kazan Tatar. Therefore, the system is empowered by a suggestion system for disambiguation purposes. The ambiguities in the translation can be eliminated by the human and the suggestions can be used for further translations. Consequently, MT-Turk is a semi-supervised translation infrastructure and the disambiguation is achieved and learned by human interaction.

MT-Turk provides a complete rule-based infrastructure for machine translation between Turkic dialects, therefore; adding a new Turkic dialect can be achieved by just adding the lexicon of roots/stems, suffixes, and the rules. Furthermore, it is bidirectional and open to extension by suggestion. Consequently, the scope and extendibility of MT-Turk will help improve the unity of Turkic communities on written work of art and obtain fusion of Turkic communities.

The success of MT-Turk translation system can be improved by supplying more resources to the system; i.e. increasing the size of the lexicon, the number of suffixes and the number of rules. Also the success of the translation can be improved by additional analysis of phrases and translation of phrase structures. Furthermore MT-Turk will improve the success of the translation autonomously by the use of the translation system through suggestions.

REFERENCES

Ahmed, A., & Hanneman, G. (2005). Syntax-based statistical machine translation: A review. *Computational Linguistics*. Retrieved from http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/cmt-5/OldFiles/OldFiles/lti/Courses/734/Spring-08/Amr+Greg-survey-SSMT.pdf

Aktaş, Ö. (2006). Türkçe için verimli bir cümle sonu belirleme yöntemi. In *Akademik Bilişim 2006 Bilgi Teknolojileri Kongresi IV*. Denizli.

Aktaş, Ö., & Çebi, Y. (2010). *Rule-Based natural language processing methods: For Turkish*. Saarbrücken: LAP LAMBERT Academic Publishing.

ALPAC. (1966). *ALPAC Report*.

Altintas, K., & Çiçekli, İ. (2002). A machine translation system between a pair of closely related languages. In *Proceedings of the 17th International Symposium on Computer and Information Sciences (ISCIS 2002)*(pp192-196). Orlando, Florida: CRC Press.

Altıntaş, K. (2001). *Turkish to Crimean Tatar Machine Translation System*. *Language*. Bilkent University.

Apptek. (2012). *Apptek*. Retrieved from http://www.apptek.com/

Bar-Hillel, Y. (1960). The present status of automatic translation of languages . *Advances in Computers*, *1*, 91–163.

Birant, Ç. C. (2008). *Root-Suffix seperation of Turkish words*. M.Sc. Thesis. İzmir: Dokuz Eylül Üniversitesi.

Birant, Ç. C., Aktaş, Ö., & Çebi, Y. (2010). *Root-Suffix seperation of Turkish words - basics, design and method*. Saarbrücken: LAP LAMBERT Academic Publishing.

Bozkurt, F. (2002). *Türklerin dili* (2nd ed.)(50-51). Ankara: Kültür Bakanlığı Yayınları.

Brown, P., & Cocke, J. (1988). A statistical approach to French/English translation. *Proceedings of Second TMI Conference*. Retrieved from http://mt-archive.info/TMI-1988-Brown.pdf

Brown, P. E., Pietra, S. A. Della, Pietra, V. J. Della, & Mercer, R. L. (1993). The mathematics of statistical machine translation : Parameter estimation. *Computational Linguistics, 10598*.

Brown, P. F., Cocke, J., Pietra, S. A., Della Pietra, V. J. Della Jelinek, F., Lafferty, J. D.,Watson, T. J. (1990). Statistical approach to machine translation, *Computational Linguistics*, *16*(2), 79–85.

Carnegie Mellon University. (1997). *Generalized example-based machine translation.* Retrieved January 30, 2012, from http://www.cs.cmu.edu/~ralf/ebmt/ebmt.html

Chen, K., & Chen, H. (1996). A Hybrid Approach to Machine Translation System Design, *Computational Linguistics and Chinese Language Processing, 1*(1), 159–182.

Cieślak, M. (2011). The scope and limits of machine translation. *eLingUp*, *3*(1988), 156–174.

Çengel, H. K. (2005). *Kırgız Türkçesi grameri - ses ve şekil bilgisi*. Ankara: Akçağ Yayınları.

Çiçekli, İ. (2005). Türkçe ve Kırım Tatarcası arasında bir çeviri sistemi. In *Bilgisayar Destekli Dil Bilimi Çalıştayı Bildirileri* (pp.123– 132). Ankara.

DEU CSE. (2004). *Dokuz Eylül University Natural Language Processing Research Group.* Retrieved April 14, 2013, from http://nlp.cs.deu.edu.tr

Doddington, G. (2002a). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the Second International Conference on Human Language Technology Research*, 138. doi:10.3115/1289189.1289273

Doddington, G. (2002b). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. *Proceedings of the Second International Conference on Human Language Technology Research* -, 138. doi:10.3115/1289189.1289273

Dorr, B., Hovy, E., & Levin, L. (2006). Machine Translation: Interlingual Methods. In E.-C. K. Brown (Ed.), *Encyclopedia of Language and Linguistics (Second Edition)* (2 Ed.)(383-394). Oxford: Elsevier. doi:10.1016/B0-08-044854-2/00939-1

Douglas, A., Balkan, L., Meijer, S., Humphreys, R. ., & Sadler, L. (1993). *Machine translation: An introductory guide*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:MACHINE+TRANSLATION+An+Introductory+Guide#5

Drozdek, A. (1989). Interlingua in machine translation. In *Proceedings of the 17th conference on ACM Annual Computer Science Conference* (pp. 434–434). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1030260

TBD (Informatics Association of Turkey). (2000). *Türk Cumhuriyetleri Bilgi Teknolojileri Çalışma Grubu*. Retrieved March 27, 2013, from http://www.tbd.org.tr/index.php?sayfa=calisma_gruplari&grup=3

Ercilasun, A. B. (1992). *Karşılaştırmalı Türk Lehçeleri Sözlüğü*. Ankara: Kültür Bakanlığı Yayınları.

Fatih University. (2013). *DİLMAÇ Project*. Retrieved April 10, 2013, from http://datamining.ceng.fatih.edu.tr:8080/dilmac/

Fidan, Ö., & Koşaner, Ö. (2007). *Turkish Suffixes*. İzmir.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A. et.al. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, *25*(2), 127–144. doi:10.1007/s10590-011-9090-0

GƏCİL. (2012). *Tatarca-Törekça Süzlek.* Retrieved February 12, 2013, from http://www.gajil.20m.com

Gedikli, Y. (1993). *Türk Lehçelerinin karşılaştırmalı dilbilgisi*. İstanbul: Cem Yayınevi.

Google. (2012). *Google Translate.* Retrieved January 30, 2012, from http://translate.google.com/

Göksel, A., & Kerslake, C. (2005). *Turkish : A Comprehensive grammar* (p. 1,9). Abingdon, Oxon: Routledge Taylor & Francis Group.

GrammarSoft ApS, & Kaldera Språkteknologi AS. (2006). *GramTrans.* Retrieved January 15, 2012, from http://gramtrans.com/

Gülensoy, T., & Sağınbayeva, B. (2004). *Kırgız Türkçesi – Türkiye Türkçesi ve Türkiye Türkçesi – Kırgız Türkçesi sözlük*. Kayseri: Erciyes Üniversitesi Yayınları.

Hajič, J., Hric, J., & Kubon, V. (2000). Machine translation of very close languages. In *Sixth Conference on Applied Natural Language*.(pp.7-12) Retrieved from http://dl.acm.org/citation.cfm?id=974149

Hamzaoğlu, İ. (1993). *Machine Translation from Turkish to Other Turkish Languages and an Implementation for the Azeri Language*. M.Sc. Thesis. İstanbul: Boğaziçi University.

Homola, P., & Kuboˇ, V. (2008). Improving machine translation between closely related Romance languages. *Architecture*, (September), 22–23.

Hovy, E. (2001a). Machine Translation. In Michael I. Jordan & S. Russell (Eds.), *The MIT Encyclopedia of the Cognitive Sciences (MITECS).*(pp.498-501). Mit Press. Retrieved from http://ai.ato.ms/MITECS/Entry/hovy1.html

Hovy, E. (2001b). Machine Translation. In Michael I. Jordan & S. Russell (Eds.), *The MIT Encyclopedia of the Cognitive Sciences (MITECS).* (pp.498-501). Mit Press.

Hutchins, J. (2004). The first public demonstration of machine translation: the Georgetown-IBM system, 7th January 1954. *AMTA conference*, (January 1954). Retrieved from http://sts.bdtf.hu/btk/flli/romanisztika/OKTATSARS DOCENDI/ TANANYAGOK (OKTAT SZERINT)/ANTONIO SCIACOVELLI/TRADUTTORI INTERPRETI/corrispondenza commerciale/traduzione elettronica_articolo_english.pdf

Hutchins, J. (2005). Towards a definition of example-based machine translation. In *Proceedings of the 2nd Workshop on Example-Based Machine Translation at MT Summit X* (pp. 63–70). Retrieved from http://www.iai-sb.de/carl/EBMT2/EBMT2_proceedings.pdf#page=71

Hutchins, W J. (1994). Machine translation : History and general principles. In R. E. Asher (Ed.), *The Encyclopedia of Languages and Linguistics* (Vol. 5, 2322–2332). Oxford: Pergamon Press.

Hutchins, W J. (1986). *Machine translation: Past, present, future*. Chichester: Ellis Horwood. Retrieved from http://www.hutchinsweb.me.uk/PPF-TOC.htm

IBM. (1954). *IBM Archives: 701 Translator*. Retrieved August 28, 2012, from http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html

ISUG (International SGML/XML Users' Group). (2010). *A Gentle introduction to SGML.* Retrieved December 22, 2010, from http://www.isgmlug.org/sgmlhelp/g-index.htm

Jonsay. (2013). *Most Spoken Languages in the World.* Retrieved April 26, 2013, from http://www.jonsay.co.uk/Articles/Language/Most_Spoken_Languages_in_the_World.html

Jurafsky, D., & Martin, J. H. (2006). Machine translation. In *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*.

Kit, C., Pan, H., & Webster, J. (2002). Example-based machine translation: A new paradigm. In *Translation and Information Technology* (pp. 57-78). Chinese University of HK Press. Retrieved from http://openstorage.gunadarma.ac.id/pub/books/MachineTranslation/EBMT-review-CUHK.pdf

Koehn, P. (2007). *Statistical machine translation.* Retrieved from http://www.mt-archive.info/MTS-2007-Koehn-3.pdf

Koehn, P., Hoang, H. Birch, A., Callison-burch, C., Federico, M., Bertoldi et.al. (2007). Moses : Open Source Toolkit for Statistical Machine Translation. *Computational Linguistics,* (June), 177–180.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical Phrase-Based Translation. *Main*, (June), 48–54.

Kulesza, A., & Shieber, S. M. (2004). A learning approach to improving sentence-level MT evaluation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*. Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.143.3469&amp;rep=rep1&amp;type=pdf

Kültür Bakanlığı. (n.d.). *Türkiye Dışındaki Türk Edebiyatları Antolojisi.* Retrieved April 19, 2013, from http://ekitap.kulturturizm.gov.tr/belge/1-27606/turkiye-disindaki-turk-edebiyatlari-antolojisi.html

Kültür Bakanlığı. (2001a). Tatar Edebiyatı (XVII. Cilt). In Proje Yöneticisi: N. Köroğlu (Ed.), *Başlangıçtan Günümüze Kadar Türkiye Dışındaki Türk Edebiyatları Antolojisi*. Ankara: T.C. Kültür ve Turizm Bakanlığı.

Kültür Bakanlığı. (2001b). Tatar Edebiyatı (XVIII. Cilt). In Proje Yöneticisi: N. Köroğlu (Ed.), *Başlangıçtan Günümüze Kadar Türkiye Dışındaki Türk Edebiyatları Antolojisi*. Ankara: T.C. Kültür ve Turizm Bakanlığı.

Kültür Bakanlığı. (2001c). Tatar Edebiyatı (XIX. Cilt). In Proje Yöneticisi: N. Köroğlu (Ed.), *Başlangıçtan Günümüze Kadar Türkiye Dışındaki Türk Edebiyatları Antolojisi*. Ankara: T.C. Kültür ve Turizm Bakanlığı.

Kültür Bakanlığı. (2005a). Kırgız Edebiyatı I (XXXI. Cilt). In Proje Yöneticisi: N. Köroğlu (Ed.), *Başlangıçtan Günümüze Kadar Türkiye Dışındaki Türk Edebiyatları Antolojisi*. Ankara: T.C. Kültür ve Turizm Bakanlığı.

Kültür Bakanlığı. (2005b). Kırgız Edebiyatı II (XXXII. Cilt). In Proje Yöneticisi: N. Köroğlu (Ed.), *Başlangıçtan Günümüze Kadar Türkiye Dışındaki Türk Edebiyatları Antolojisi*. Ankara: T.C. Kültür ve Turizm Bakanlığı.

Lavie, A., Sagae, K., & Jayaraman, S. (2004). The significance of recall in automatic metrics for mt evaluation. *Machine Translation: From Real Users to Research*, 134–143. Retrieved from http://www.springerlink.com/index/lx22t5fuwkn2gvy9.pdf

Lewis, G. L. (1967). *Turkish grammar*. New York: Oxford University Press.

Lewis, M. P. (Ed.). (2009). *Ethnologue: Languages of the World* (Sixteenth .). Dallas, Tex.: SIL International. Retrieved from http://www.ethnologue.com/

Liddy, E. D. (2001). Natural Language Processing. In *Encyclopedia of Library and Information Science* (2nd ed.). New York: Marcel Decker, Inc.

Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys*, *40*(3), 1–49. doi:10.1145/1380584.1380586

Marrafa, P., & Ribeiro, A. (2001). Quantitative evaluation of machine translation systems: sentence level. In *Proceedings of the MT Summit VIII Fourth ISLE workshop* (pp. 39–43). Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.849&amp;rep=rep1&amp;type=pdf

Mayor, A. (2007). *Matxin: Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estal- dura handiko baliabide linguistikoak berrerabiliz (Matxin: construction of a rule-based MT system reusing wide coverage linguistic resources).* University of the Basque Country.

Mayor, A., Alegria, I., Díaz de Ilarraza, A., Labaka, G., Lersundi, M., & Sarasola, K. (2011). Matxin, an open-source rule-based machine translation system for Basque. *Machine Translation*, *25*(1), 53–82. doi:10.1007/s10590-011-9092-y

Microsoft. (2012). *Bing Translator*. Retrieved January 20, 2012, from http://www.microsofttranslator.com/

Nagao, M. (1984). A framework of a mechanical translation between Japanese and English by analogy principle. In A. Elithorn & R. Banerji (Eds.), *Artificial And Human Intelligence*. Elsevier Science Publishers. B.V. Retrieved from http://books.google.com/books?hl=en&amp;lr=&amp;id=yx3lEVJMBmMC&amp;oi=fnd&amp;pg=PA351&amp;dq=A+FRAMEWORK+OF+A+MECHANICAL+TRANSLATION+BETWEEN+JAPANESE+AND+ENGLISH+BY+ANALOGY+PRINCIPLE&amp;ots=sb4zjPHJEq&amp;sig=XQ1Xyqbxoh_DJKmZ1zxDIt1TyJI

NIST (National Institute Of Standards And Technology). (2010). *NIST Tools*. Retrieved December 10, 2010, from http://www.itl.nist.gov/iad/mig//tools/

Och, F., & Ney, H. (2000). Improved statistical alignment models. *Proceedings of the 38th Annual Meeting of the Association of Computational Linguistics (ACL).* Retrieved from http://dl.acm.org/citation.cfm?id=1075274

Oflazer, K., Çetinoğlu, Ö., & Say, B. (2004). Integrating morphology with multi-word expression processing in Turkish. *Proceedings of the Workshop on Multiword Expressions Integrating Processing - MWE ’04*, (July), 64–71. doi:10.3115/ 1613186.1613195

Orhun, M., Adali, E., & Tantuğ, A. C. (2011). Uygurcadan Türkçeye bilgisayarlı çeviri. *İstanbul Üniversitesi Mühendislik Dergisi*, (212), 3–14.

Öner, M. (1998). *Bugünkü Kıpçak Türkçesi*. Ankara: Türk Dil Kurumu Yayınları.

Öner, M. (2007a). Tatar Türkçesi. In *Türk Lehçeleri Grameri*.(pp. 679-748). Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:TATAR+ TÜRKÇESİ#2

Öner, M. (2007b). *Türk Lehçeleri Grameri*. (Ahmet Bican Ercilasun, Ed.). Akçağ Yayınları.

Papineni, K., Roukos, S., Ward, T., Zhu, W., & Heights, Y. (2001). *IBM Research Report Bleu : a Method for Automatic Evaluation of Machine Translation*. *Science* (Vol. 22176).

Ramanathan, A. (2009). *Statistical machine translation*. *US Patent 7,624,005*. Bombay Mumbai.

Resnik, P., & Park, C. (2006). Word-Based Alignment, *Phrase-Based Translation : What’s the Link?*, (August), 90–99.

Sag, I., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. *In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING*, 1–15.

Sánchez-Cartagena, V., Sánchez-Martínez, F., & Pérez-Ortiz, J. A. (2011). The Universitat d'Alacant hybrid machine translation system for WMT 2011. In *Proceedings of the 6th Workshop on Statistical Machine Translation* (pp. 457–463). Edinburgh, Scotland, UK. Retrieved from http://www.dlsi.ua.es/~fsanchez/pub/pdf/sanchez-cartagena11a.pdf

Sawaf, H., Gaskill, B., & Veronis, M. (2008). Hybrid Machine Translation Applied to Media Montoring. *Machine Translation in the Americas*. Retrieved from http://mt-archive.info/AMTA-2008-Sawaf.pdf

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(July 1948), 379–423. Retrieved from http://dl.acm.org/citation.cfm?id=584093

Shylov, M. (2008). *Turkish and Turkmen Morphological Analyzer and Machine Translation Program.* M.Sc. Thesis. İstanbul: Fatih University.

SIL International. (2004a). *Glossary of Linguistic Terms: What is a Root?* Retrieved from http://www-01.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsARoot.htm

SIL International. (2004b). *Glossary of Linguistic Terms: What is a Stem?* Retrieved July 20, 2013, from http://www-01.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsAStem.htm

SIL International. (2013). *Ethnologue: Languages of the World*. Retrieved April 18, 2013, from http://www.ethnologue.com/family/17-15

Somers, H. (2001). *Review Article : Example-based Machine Translation,* (Tmi 1992), 113–157.

Somers, H. (July, 2004). *Machine Translation And Welsh: The Way Forward,* Retrieved August 20, 2013, from http://mti.ugm.ac.id/~adji/courses/resources/doctor/MT_book/Machine%20Translation%20and%20Welsh%20(PDF).pdf.

SOROSORO. (2009). *Turkic Language Family*. Retrieved January 18, 2012, from http://www.sorosoro.org/en/turkic-language-family

Stroppa, N., Groves, D., Way, A., & Sarasola, K. (2006). Example-based machine translation of the Basque language. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas.*(pp.232-241). Retrieved from http://doras.dcu.ie/15821/

Su, K., & Chang, J. (1992). Why corpus-based statistics-oriented machine translation. In *Proceedings of 4th International Conference on Theoretical and Methodological Issues in Machine Translation.*(pp. 249-262). Retrieved from https://www.info.unicaen.fr/M2-LID/articles-2010-2011/JV1.pdf

Systran. (2011). *Systran.* Retrieved January 30, 2012, from http://www.systran.co.uk/systran/corporate-profile/translation-technology/systran-hybrid-technology

Şahin, E. (2003). Kazan Tatar Türklerinin Latin alfabesi mücadelesi. *Türk Dünyası Tarih Kültür Dergisi*, (199), 42–45.

Şenkal, M. (2000). *An Approach For Machine Translation Between Turkish And Spanish*. Bilkent University.

Tantuğ, A. C. (2007a). *Akraba ve bitişken diller arasında bilgisayarlı çeviri için karma bir model*. Istanbul Technical University.

Tantuğ, A. C. (2007b). A mt system from turkmen to turkish employing finite state and statistical methods. In *In Proceedings of MT Summit XI*. Copenhagen, Denmark. Retrieved from http://research.sabanciuniv.edu/6395/

Tantuğ, A. C., Adali, E., & Oflazer, K. (2008). Türkmenceden Türkçeye bilgisayarlı metin çevirisi. *İstanbul Üniversitesi Mühendislik Dergisi*, *7*(4), 83–94.

Tantuğ, A. C., Oflazer, K., & El-Kahlout, I. (2008). BLEU+: a tool for fine-grained BLEU computation. *Proceeding LREC Marrakech*, (L). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.143.6111&amp;rep=rep1&amp;type=pdf

Türk Dil Kurumu. (2005). *Çalışmalarımız.* Retrieved April 29, 2013, from http://www.tdk.gov.tr/index.php?option=com_content&view=article&id=142:Calismalarimiz--2005&catid=41:calismalar&Itemid=63

Türk Dil Kurumu. (2009). *Kişi Adları Sözlüğü.* Retrieved December 12, 2009, from http://www.tdk.org.tr/TR/Genel/AdArama.aspx?F6E10F8892433CFFAAF6AA849816B2EF0BF5B4755D05B9EB

Türk Dil Kurumu. (2011a). *Güncel Türkçe Sözlük.* Retrieved January 14, 2011, from http://www.tdk.gov.tr/index.php?option=com_gts&view=gts

Türk Dil Kurumu. (2011b). *Türk Lehçeleri Sözlüğü.* Retrieved January 14, 2012, from http://www.tdk.gov.tr/index.php?option=com_lehceler&view=lehceler

Thurmair, G. (2005). Hybrid Architectures for Machine Translation Systems. *Language Resources and Evaluation*, *39*(1), 91–108. doi:10.1007/s10579-005-2698-z

Turian, J. P., Shen, L., & Melamed, I. D. (1995). Evaluation of machine translation and its evaluation. *recall (C| R)*, *100*(1), 2. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.3.388&amp;rep=rep1&amp;type=pdf

Turkic languages. (2012). In *Encyclopædia Britannica*. Encyclopædia Britannica Inc. Retrieved from http://www.britannica.com/EBchecked/topic/609955/Turkic-languages

Uğurlu, M. (2004). Türk Lehçeleri Arasında Kelime Eş Değerliği. *Bilig*, (29), 29–40. Retrieved from http://turuz.info/Kesli toplusu-meqale mecmuesi/0318-Turk Lehceleri arasinda kelime esh degherliyi(12).pdf

Ulitkin, I. (2011). Computer-assisted Translation Tools: A brief review. *Translation Journal*, *15*(1). Retrieved from http://translationjournal.net/journal/55computers.htm

Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. In *Ifip Congress* (1114–1122).

Venkatapathy, S., & Joshi, A. K. (2006). Using Information about Multi-word Expressions for the Word-Alignment Task. *Computational Linguistics*, (July), 53–60.

Way, A. (2010). Machine Translation. In *Computational Linguistics and Natural Language Processing Handbook* (531–574). West Sussex: Blackwell.

Weaver, W. (1949). *Translation*. New York, USA.

Wiechetek, L. (2008). Rule-based MT approaches such as Apertium and GramTrans RBTM approaches, (Bick 2000).

Wu, D. (1997). Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora.

Xuan, H. W., Li, W., & Tang, G. Y. (2012). An Advanced Review of Hybrid Machine Translation (HMT). *Procedia Engineering*, *29*, 3017–3022. doi:10.1016/ j.proeng.2012.01.432

Yamada, K., & Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL ’01* (pp. 523–530). Morristown, NJ, USA: Association for Computational Linguistics. doi:10.3115/1073012.1073079

Zhang, Y., Vogel, S., & Waibel, A. (2004). Interpreting BLEU/NIST scores: How much improvement do we need to have a better system. *Proceedings of LREC*. Retrieved from http://www.lrec-conf.org/proceedings/lrec2004/pdf/755.pdf

Zwarts, S. (2010). *Machine Translation Evaluation.* Retrieved from http://web.science.mq.edu.au/~szwarts/MT-Evaluation.php

APPENDICES

A. PHONOLOGICAL RULES

A.1 Turkish Phonological Rules

```
<?xml version="1.0" encoding="utf-8"  ?>

<phonology xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="XmlPhonology.xsd">
 <alphabet>
  <character index="1" type= "vowel">a</character>
  <character index="1" type= "vowel">â</character>
  <character index="2" type= "consonant">b</character>
  <character index="3" type= "consonant">c</character>
  <character index="4" type= "consonant">ç</character>
  <character index="5" type= "consonant">d</character>
  <character index="6" type= "vowel">e</character>
  <character index="7" type= "consonant">f</character>
  <character index="8" type= "consonant">g</character>
  <character index="9" type= "consonant">ğ</character>
  <character index="10" type= "consonant">h</character>
  <character index="11" type= "vowel">ı</character>
  <character index="12" type= "vowel">i</character>
  <character index="13" type= "consonant">j</character>
  <character index="14" type= "consonant">k</character>
  <character index="15" type= "consonant">l</character>
  <character index="16" type= "consonant">m</character>
  <character index="17" type= "consonant">n</character>
  <character index="18" type= "vowel">o</character>
  <character index="19" type= "vowel">ö</character>
  <character index="20" type= "consonant">p</character>
  <character index="21" type= "consonant">r</character>
  <character index="22" type= "consonant">s</character>
  <character index="23" type= "consonant">ş</character>
  <character index="24" type= "consonant">t</character>
  <character index="25" type= "vowel">u</character>
  <character index="26" type= "vowel">ü</character>
  <character index="27" type= "consonant">v</character>
  <character index="28" type= "consonant">y</character>
  <character index="29" type= "consonant">z</character>
  <character index="30" type= "vowel">A</character>
  <character index="31" type= "vowel">I</character>
  <character index="32" type= "consonant">C</character>
  <character index="33" type= "consonant">D</character>
  <character index="34" type= "vowel">1</character><!-- u -->
  <character index="35" type= "vowel">2</character><!-- ü -->
  <character index="36" type= "vowel">3</character><!-- ı -->
  <character index="37" type= "vowel">4</character><!-- i -->
 </alphabet >
 <substitutes>
  <substitute index="1"  name="C|D"  valid="true" force_match="false"
      force_inside_for_suffix="false">
   <match>
    <suffix loc="any" type="char">
     <pair>
      <lex>C|D</lex>
```

```
            <surf>c|ç|d|t</surf>
          </pair>
        </suffix>
      </match>
      <action>
        <suffix loc="first" type ="char">
          <pair>
            <lex>C</lex>
            <surf>c</surf>
            <surf>ç</surf>
          </pair>
          <pair>
            <lex>D</lex>
            <surf>d</surf>
            <surf>t</surf>
          </pair>
        </suffix>
      </action>
    </substitute>
    <substitute index="2"  name="A|I"  valid="true" force_match="false"
         force_inside_for_suffix="false">
      <match>
        <suffix loc="any" type="char">
          <pair>
            <lex>A|I</lex>
            <surf>a|e|ı|i|u|ü</surf>
          </pair>
        </suffix>
      </match>
      <action>
        <suffix loc="any" type ="char">
          <pair>
            <lex>A</lex>
            <surf>a</surf>
            <surf>e</surf>
          </pair>
          <pair>
            <lex>I</lex>
            <surf>ı</surf>
            <surf>i</surf>
            <surf>u</surf>
            <surf>ü</surf>
          </pair>
        </suffix>
      </action>
    </substitute>
  </substitutes>
  <rules>
    <rule index="1"  name="Yumuşama"  applied_stem_type ="isim" syllable="2" valid="true"
         force_match="false" force_inside_for_suffix="false" stem_based_optional="true">
      <match>
          <stem loc="last" type="char">
          <pair>
            <lex>p|ç|t|k</lex>
            <surf>b|c|d|g|ğ</surf>
          </pair>
        </stem>
  <suffix loc ="first" type ="char">
   <pair>
    <lex>vowel</lex>
    <surf>vowel</surf>
   </pair>
  </suffix>
 </match>
 <action>
  <stem loc="last" type ="chars">
   <pair>
    <lex>p</lex>
    <surf>b</surf>
   </pair>
   <pair>
    <lex>ç</lex>
    <surf>c</surf>
   </pair>
   <pair>
    <lex>t</lex>
    <surf>d</surf>
   </pair>
   <pair>
     <lex>nk</lex>
     <surf>ng</surf>
   </pair>
   <pair>
    <lex>k</lex>
    <surf>ğ</surf>
   </pair>
  </stem>
 </action>
     <exception part="stem">et</exception>
     <exception part="stem">ant</exception>
     <exception part="stem">git</exception>
     <exception part="stem">güt</exception>
     <exception part="stem">tat</exception>
</rule>
  <rule index="11"  name="Yumuşama_3_hece"  applied_stem_type ="isim" syllable="3"
  valid="true" force_match="false" force_inside_for_suffix="false" stem_based_optional="true">
      <match>
             <stem loc="last" type="char">
                    <pair>
                           <lex>k</lex>
                           <surf>g|ğ</surf>
                    </pair>
             </stem>
             <suffix loc ="first" type ="char">
                    <pair>
                           <lex>vowel</lex>
                           <surf>vowel</surf>
                    </pair>
             </suffix>
      </match>
      <action>
             <stem loc="last" type ="chars">
                    <pair>
                           <lex>nk</lex>
                           <surf>ng</surf>
                    </pair>
                    <pair>
                           <lex>k</lex>
                           <surf>ğ</surf>
                    </pair>
             </stem>
      </action>
     </rule>
<rule index="2"  name="Darlaşma_Önceki_Düz"  applied_stem_type ="fiil" valid="true"
   force_match="false" force_inside_for_suffix="false" stem_based_optional="true">
  <match>
   <stem loc="last" type="char">
    <pair>
     <lex>a|e</lex>
     <surf>ı|i|u|ü</surf>
    </pair>
   </stem>
   <stem loc="before_last" type="vowel">
    <pair>
     <lex>a|e|ı|i</lex>
     <surf>a|e|ı|i</surf>
    </pair>
   </stem>
   <suffix loc ="first" type ="char">
    <pair>
     <lex>y</lex>
     <surf>y</surf>
    </pair>
   </suffix>
  </match>
  <action>
   <stem loc="last" type ="char">
    <pair>
     <lex>a</lex>
     <surf>ı</surf>
    </pair>
    <pair>
     <lex>e</lex>
     <surf>i</surf>
    </pair>
   </stem>
  </action>
</rule>
<rule index="3"  name="Darlaşma_Önceki_Yuvarlak" applied_stem_type ="fiil" valid="true"
   force_match="false" force_inside_for_suffix="false" stem_based_optional="true">
  <match>
             <stem loc="last" type="char">
                    <pair>
                           <lex>a|e</lex>
                           <surf>ı|i|u|ü</surf>
                    </pair>
             </stem>
             <stem loc="before_last" type="vowel">
                    <pair>
                           <lex>o|ö|u|ü</lex>
                           <surf>o|ö|u|ü</surf>
```

```
                        </pair>
                </stem>
                <suffix loc ="first" type ="char">
                        <pair>
                                <lex>y</lex>
                                <surf>y</surf>
                        </pair>
                </suffix>
        </match>
        <action>
                <stem loc="last" type ="char">
                        <pair>
                                <lex>a</lex>
                                <surf>u</surf>
                        </pair>
                        <pair>
                                <lex>e</lex>
                                <surf>ü</surf>
                        </pair>
                </stem>
        </action>
 </rule>
 <rule index="4"  name="Yuvarlaklaşma" applied_stem_type ="fiil" valid="true" force_match="false"
force_inside_for_suffix="false" stem_based_optional="true">
        <match>
               <stem loc="last" type="char">
     <pair>
      <lex>a|e|ı|i</lex>
      <surf>u|ü</surf>
     </pair>
    </stem>
    <stem loc="before_last" type="vowel">
     <pair>
      <lex>o|ö|u|ü</lex>
      <surf>o|ö|u|ü</surf>
     </pair>
    </stem>
    <suffix loc ="first" type ="vowel">
     <pair>
      <lex>o|ö|u|ü</lex>
      <surf>o|ö|u|ü</surf>
     </pair>
    </suffix>
   </match>
   <action>
    <stem loc="last" type ="char">
     <pair>
      <lex>a|ı</lex>
      <surf>u</surf>
     </pair>
     <pair>
      <lex>e|i</lex>
      <surf>ü</surf>
     </pair>
    </stem>
   </action>
  </rule>
```

```
            <rule index="6"  name="Küçük_Ünlü_Uyumu" applied_stem_type ="hepsi" valid="true"
force_match="true" force_inside_for_suffix="false" stem_based_optional="false">
                        <match>
                                <stem loc="last" type="vowel">
                                        <pair>
                                                <lex>â|e|i|a|ı|4|3</lex>
                                                <surf>â|e|i|a|ı|4|3</surf>
                                        </pair>
                                </stem>
                                <suffix loc ="first" type ="vowel">
                                        <pair>
                                                <lex>â|e|i|a|ı|4|3</lex>
                                                <surf>â|e|i|a|ı|4|3</surf>
                                        </pair>
                                </suffix>
                        </match>
                        <match>
                                <stem loc="last" type="vowel">
                                        <pair>
                                                <lex>u|o|ü|ö|1|2</lex>
                                                <surf>u|o|ü|ö|1|2</surf>
                                        </pair>
                                </stem>
                                <suffix loc ="first" type ="vowel">
                                        <pair>
                                                <lex>â|a|e|u|ü|1|2</lex>
                                                <surf>â|a|e|u|ü|1|2</surf>
                                        </pair>
                                </suffix>
                        </match>
                        <exception part="suffix">ken</exception>
                        <exception part="suffix">yor</exception>
                        <exception part="suffix">mtrak</exception>

            </rule>
  <rule index="6"  name="Küçük_Ünlü_Uyumu_içerde" applied_stem_type ="hepsi" valid="true"
force_match="true" force_inside_for_suffix="true" stem_based_optional="false">
    <match>
      <stem loc="last" type="vowel">
        <pair>
          <lex>â|e|i|a|ı|4|3</lex>
          <surf>â|e|i|a|ı|4|3</surf>
        </pair>
      </stem>
      <suffix loc ="first" type ="vowel">
        <pair>
          <lex>â|e|i|a|ı|4|3</lex>
          <surf>â|e|i|a|ı|4|3</surf>
        </pair>
      </suffix>
    </match>
    <match>
      <stem loc="last" type="vowel">
        <pair>
          <lex>u|o|ü|ö|1|2</lex>
          <surf>u|o|ü|ö|1|2</surf>
        </pair>
    </stem>
    <suffix loc ="first" type ="vowel">
     <pair>
      <lex>â|a|e|u|ü|1|2</lex>
      <surf>â|a|e|u|ü|1|2</surf>
     </pair>
    </suffix>
   </match>
              <exception part="suffix">ken</exception>
              <exception part="suffix">Abil</exception>
              <exception part="suffix">Iver</exception>
              <exception part="suffix">Agel</exception>
              <exception part="suffix">Adur</exception>
              <exception part="suffix">Akal</exception>
              <exception part="suffix">Ayaz</exception>
              <exception part="suffix">yor</exception>
              <exception part="suffix">Iyor</exception>
              <exception part="suffix">mtrak</exception>
              <exception part="suffix">Imtrak</exception>
  </rule>
  <rule index="7"  name="Büyük_Ünlü_Uyumu" applied_stem_type ="hepsi" valid="true"
force_match="true" force_inside_for_suffix="false" stem_based_optional="false">
   <match>
    <stem loc="last" type="vowel">
     <pair>
      <lex>a|ı|u|o|1|3</lex>
      <surf>a|ı|u|o|1|3</surf>
     </pair>
    </stem>
    <suffix loc ="first" type ="vowel">
     <pair>
      <lex>a|ı|u|1|3</lex>
      <surf>a|ı|u|1|3</surf>
     </pair>
    </suffix>
   </match>
   <match>
    <stem loc="last" type="vowel">
     <pair>
      <lex>â|e|i|ü|ö|2|4</lex>
      <surf>â|e|i|ü|ö|2|4</surf>
     </pair>
    </stem>
    <suffix loc ="first" type ="vowel">
     <pair>
      <lex>â|e|i|ü|2|4</lex>
      <surf>â|e|i|ü|2|4</surf>
     </pair>
    </suffix>
   </match>
              <exception part="suffix">ken</exception>
              <exception part="suffix">yor</exception>
              <exception part="suffix">mtrak</exception>
  </rule>
        <rule index="7"  name="Büyük_Ünlü_Uyumu_içerde" applied_stem_type ="hepsi"
valid="true" force_match="true" force_inside_for_suffix="true" stem_based_optional="false">
        <match>
```

```
                <stem loc="last" type="vowel">
                        <pair>
                                <lex>a|ı|u|o|1|3</lex>
                                <surf>a|ı|u|o|1|3</surf>
                        </pair>
                </stem>
                <suffix loc ="first" type ="vowel">
                        <pair>
                                <lex>a|ı|u|1|3</lex>
                                <surf>a|ı|u|1|3</surf>
                        </pair>
                </suffix>
        </match>
        <match>
                <stem loc="last" type="vowel">
                        <pair>
                                <lex>â|e|i|ü|ö|2|4</lex>
                                <surf>â|e|i|ü|ö|2|4</surf>
                        </pair>
                </stem>
                <suffix loc ="first" type ="vowel">
                        <pair>
                                <lex>â|e|i|ü|2|4</lex>
                                <surf>â|e|i|ü|2|4</surf>
                        </pair>
                </suffix>
        </match>
                <exception part="suffix">ken</exception>
                <exception part="suffix">Abil</exception>
                <exception part="suffix">Iver</exception>
                <exception part="suffix">Agel</exception>
                <exception part="suffix">Adur</exception>
                <exception part="suffix">Akal</exception>
                <exception part="suffix">Ayaz</exception>
                <exception part="suffix">yor</exception>
                <exception part="suffix">Iyor</exception>
                <exception part="suffix">mtrak</exception>
                <exception part="suffix">Imtrak</exception  >
        </rule>
  <rule index="8"  name="Sert_Sessiz_Uyumu" applied_stem_type ="hepsi" valid="true"
force_match="true" force_inside_for_suffix="false" stem_based_optional="false">
   <match>
    <stem loc="last" type="char">
     <pair>
      <lex>f|s|ş|h|p|ç|t|k</lex>
      <surf>f|s|ş|h|p|ç|t|k</surf>
     </pair>
    </stem>
    <suffix loc ="first" type ="char">
     <pair>
      <lex>f|s|ş|h|p|ç|t|k|l|m|n|r|v|y|z|a|e|ı|i|u|ü</lex>
      <surf>f|s|ş|h|p|ç|t|k|l|m|n|r|v|y|z|a|e|ı|i|u|ü</surf>
     </pair>
    </suffix>
   </match>
   <match>
    <stem loc="last" type="char">
```

```
     <pair>
      <lex>a|b|c|d|e|g|ğ|ı|i|j|l|m|n|o|r|u|ü|v|y|z</lex>
      <surf>a|b|c|d|e|g|ğ|ı|i|j|l|m|n|o|r|u|ü|v|y|z</surf>
     </pair>
    </stem>
    <suffix loc ="first" type ="char">
     <pair>
      <lex>a|b|c|d|e|g|ğ|ı|i|j|l|m|n|o|r|u|ü|v|y|z</lex>
      <surf>a|b|c|d|e|g|ğ|ı|i|j|l|m|n|o|r|u|ü|v|y|z</surf>
     </pair>
    </suffix>
   </match>
  </rule>
  <rule index="9"  name="Ses_Uyumu" applied_stem_type ="hepsi" valid="true" force_match="true"
force_inside_for_suffix="false" stem_based_optional="false">
   <match>
    <stem loc="last" type="char">
     <pair>
      <lex>vowel</lex>
      <surf>vowel</surf>
     </pair>
    </stem>
    <suffix loc ="first" type ="char">
     <pair>
      <lex>consonant</lex>
      <surf>consonant</surf>
     </pair>
    </suffix>
   </match>
   <match>
    <stem loc="last" type="char">
     <pair>
      <lex>consonant</lex>
      <surf>consonant</surf>
     </pair>
    </stem>
    <suffix loc ="first" type ="char">
     <pair>
      <lex>vowel</lex>
      <surf>vowel</surf>
     </pair>
    </suffix>
   </match>
  </rule>
 </rules>
</phonology>
```

A.2 Kirghiz Phonological Rules

```xml
<?xml version="1.0" encoding="utf-8" ?>
<phonology xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="XmlPhonology.xsd">
  <alphabet>
    <character index="1" type= "vowel">a</character>
    <character index="2" type= "consonant">b</character>
    <character index="3" type= "consonant">v</character>
    <character index="4" type= "consonant">g</character>
    <character index="5" type= "consonant">d</character>
    <character index="6" type= "vowel">e</character>
    <character index="7" type= "consonant">yo</character>
    <character index="8" type= "consonant">j</character>
    <character index="9" type= "consonant">c</character>
    <character index="10" type= "consonant">z</character>
    <character index="11" type= "vowel">i</character>
    <character index="12" type= "consonant">y</character>
    <character index="13" type= "consonant">k</character>
    <character index="14" type= "consonant">l</character>
    <character index="15" type= "consonant">m</character>
    <character index="16" type= "consonant">n</character>
    <character index="17" type= "consonant">ŋ</character>
    <character index="18" type= "vowel">o</character>
    <character index="19" type= "vowel">ö</character>
    <character index="20" type= "consonant">p</character>
    <character index="21" type= "consonant">r</character>
    <character index="22" type= "consonant">s</character>
    <character index="23" type= "consonant">t</character>
    <character index="24" type= "vowel">u</character>
    <character index="25" type= "vowel">ü</character>
    <character index="26" type= "consonant">f</character>
    <character index="27" type= "consonant">x</character>
    <character index="28" type= "consonant">ts</character>
    <character index="29" type= "consonant">ç</character>
    <character index="30" type= "consonant">ş</character>
    <character index="31" type= "consonant">şç</character>
    <character index="32" type= "vowel">ı</character>
    <character index="33" type= "consonant">e</character>
    <character index="34" type= "consonant">yu</character>
    <character index="35" type= "consonant">ya</character>
    <character index="36" type= "consonant">'</character>
    <character index="37" type= "consonant">^</character>
    <character index="38" type= "vowel">A</character>
    <character index="39" type= "vowel">I</character>
    <character index="40" type= "vowel">U</character>
    <character index="41" type= "vowel">O</character>
    <character index="42" type= "consonant">L</character>
    <character index="43" type= "consonant">K</character>
    <character index="44" type= "consonant">G</character>
    <character index="45" type= "consonant">B</character>
    <character index="46" type= "consonant">D</character>
    <character index="47" type= "consonant">N</character>
    <character index="48" type= "consonant">M</character>
  </alphabet>
  <substitutes>
```

```
<substitute index="1"  name="L|K|G|B|D|N|M"  valid="true" force_match="false"
force_inside_for_suffix="false">
   <match>
    <suffix loc="first" type="char">
     <pair>
      <lex>L|K|G|B|D|N|M</lex>
      <surf>l|d|t|k|g|ğ|b|p|n|m</surf>
     </pair>
    </suffix>
   </match>
   <action>
    <suffix loc="first" type ="char">
     <pair>
      <lex>L</lex>
      <surf>l</surf>
      <surf>d</surf>
      <surf>t</surf>
     </pair>
     <pair>
      <lex>K</lex>
      <surf>k</surf>
      <surf>g </surf>
      <surf>ğ</surf>
     </pair>
     <pair>
      <lex>G</lex>
      <surf>g</surf>
      <surf>k</surf>
     </pair>
     <pair>
      <lex>B</lex>
      <surf>b</surf>
      <surf>p</surf>
     </pair>
     <pair>
      <lex>D</lex>
      <surf>d</surf>
      <surf>t</surf>
     </pair>
     <pair>
      <lex>N</lex>
      <surf>n</surf>
      <surf>d</surf>
      <surf>t</surf>
     </pair>
     <pair>
      <lex>M</lex>
      <surf>m</surf>
      <surf>d</surf>
      <surf>t</surf>
     </pair>
    </suffix>
   </action>
  </substitute>
  <substitute index="2"  name="A|I|O|U"  valid="true" force_match="false"
force_inside_for_suffix="false">
```

```
   <match>
    <suffix loc="any" type="char">
     <pair>
      <lex>A|I|O|U</lex>
      <surf>a|e|o|ö|ı|i|u|ü</surf>
     </pair>
    </suffix>
   </match>
   <action>
    <suffix loc="any" type ="char">
     <pair>
      <lex>A</lex>
      <surf>a</surf>
      <surf>e</surf>
      <surf>o</surf>
      <surf>ö</surf>
     </pair>
     <pair>
      <lex>I</lex>
      <surf>ı</surf>
      <surf>i</surf>
      <surf>u</surf>
      <surf>ü</surf>
     </pair>
     <pair>
      <lex>U</lex>
      <surf>u</surf>
      <surf>ü</surf>
     </pair>
     <pair>
      <lex>O</lex>
      <surf>o</surf>
      <surf>ö</surf>
     </pair>
    </suffix>
   </action>
  </substitute>
 </substitutes>
 <rules>
  <rule index="1"  name="Yumuşama"  applied_stem_type ="isim" valid="true" force_match="false"
force_inside_for_suffix="false"  stem_based_optional="true">
   <match>
    <stem loc="last" type="char">
     <pair>
      <lex>p|k</lex>
      <surf>b|g</surf>
     </pair>
    </stem>
    <suffix loc ="first" type ="char">
     <pair>
      <lex>vowel</lex>
      <surf>vowel</surf>
     </pair>
    </suffix>
   </match>
   <action>
```

```
        <stem loc="last" type ="char">
          <pair>
            <lex>p</lex>
            <surf>b</surf>
          </pair>
          <pair>
            <lex>k</lex>
            <surf>g</surf>
          </pair>
        </stem>
      </action>
    </rule>
    <rule index="2"  name="Ünsüz_Düşmesi_tek_heceli"  syllable="1" applied_stem_type ="hepsi"
valid="true" force_match="false" force_inside_for_suffix="false"  stem_based_optional="true">
      <match>
        <stem loc="last" type="char">
          <pair>
            <lex>p</lex>
            <surf></surf>
          </pair>
        </stem>
        <suffix loc="first" type ="chars">
          <pair>
            <lex>Ip</lex>
            <surf>ıp|ip|up|üp</surf>
          </pair>
        </suffix>
      </match>
      <action>
        <stem loc="last" type ="char">
          <pair>
            <lex>p</lex>
            <surf></surf>
          </pair>
        </stem>
      </action>
    </rule>
    <rule index="5"  name="Ses_Düşmesi" applied_stem_type ="hepsi" syllable="2" valid="true"
force_match="false" force_inside_for_suffix="false" stem_based_optional="true">
      <match>
        <stem loc="last" type="vowel">
          <pair>
            <lex>u|ü|ı|i</lex>
            <surf>1|2|3|4</surf>
          </pair>
        </stem>
        <stem loc="last" type="char">
          <pair>
            <lex>consonant</lex>
            <surf>consonant</surf>
          </pair>
        </stem>
        <suffix loc="first" type ="char">
          <pair>
            <lex>vowel</lex>
            <surf>vowel</surf>
          </pair>
    </suffix>
   </match>
   <action>
    <stem loc="last" type ="vowel">
     <pair>
      <lex>u</lex>
      <surf>1</surf>
     </pair>
     <pair>
      <lex>ü</lex>
      <surf>2</surf>
     </pair>
     <pair>
      <lex>ı</lex>
      <surf>3</surf>
     </pair>
     <pair>
      <lex>i</lex>
      <surf>4</surf>
     </pair>
    </stem>
   </action>
  </rule>

  <rule index="7"  name="Büyük_Ünlü_Uyumu" applied_stem_type ="hepsi" valid="true"
force_match="true" force_inside_for_suffix="false" stem_based_optional="false">
   <match>
    <stem loc="last" type="vowel">
     <pair>
      <lex>a|ı|u|o</lex>
      <surf>a|ı|u|o</surf>
     </pair>
    </stem>
    <suffix loc ="first" type ="vowel">
     <pair>
      <lex>a|ı|u</lex>
      <surf>a|ı|u</surf>
     </pair>
    </suffix>
   </match>
   <match>
    <stem loc="last" type="vowel">
     <pair>
      <lex>e|i|ü|ö</lex>
      <surf>e|i|ü|ö</surf>
     </pair>
    </stem>
    <suffix loc ="first" type ="vowel">
     <pair>
      <lex>e|i|ü</lex>
      <surf>e|i|ü</surf>
     </pair>
    </suffix>
   </match>
   </rule>
   <rule index="7"  name="Büyük_Ünlü_Uyumu_içerde" applied_stem_type ="hepsi" valid="true"
force_match="true" force_inside_for_suffix="true" stem_based_optional="false">
     <match>
       <stem loc="last" type="vowel">
         <pair>
           <lex>a|ı|u|o</lex>
           <surf>a|ı|u|o</surf>
         </pair>
       </stem>
       <suffix loc ="first" type ="vowel">
         <pair>
           <lex>a|ı|u</lex>
           <surf>a|ı|u</surf>
         </pair>
       </suffix>
     </match>
     <match>
       <stem loc="last" type="vowel">
         <pair>
           <lex>e|i|ü|ö</lex>
           <surf>e|i|ü|ö</surf>
         </pair>
       </stem>
       <suffix loc ="first" type ="vowel">
         <pair>
           <lex>e|i|ü</lex>
           <surf>e|i|ü</surf>
         </pair>
       </suffix>
     </match>
   </rule>
   <rule index="8"  name="Ünsüz_Uyumu"   applied_stem_type ="hepsi"  valid="true"
force_match="true" force_inside_for_suffix="true" stem_based_optional="false">
     <match>
       <stem loc="last" type="char">
         <pair>
           <lex>ç|f|x|k|p|s|ş|t|ts|şç</lex>
           <surf>ç|f|x|k|p|s|ş|t|ts|şç</surf>
         </pair>
       </stem>
       <suffix loc ="first" type ="char">
         <pair>
           <lex>ç|f|x|k|p|s|ş|t|ts|şç</lex>
           <surf>ç|f|x|k|p|s|ş|t|ts|şç</surf>
         </pair>
       </suffix>
     </match>
     <match>
       <stem loc="last" type="char">
         <pair>
           <lex>b|c|d|g|j|l|m|n|ŋ|r|v|y|z|a|e|ı|i|u|ü|o|ö</lex>
           <surf>b|c|d|g|j|l|m|n|ŋ|r|v|y|z|a|e|ı|i|u|ü|o|ö</surf>
         </pair>
       </stem>
       <suffix loc ="first" type ="char">
         <pair>
           <lex>b|c|d|g|j|l|m|n|ŋ|r|v|y|z|a|e|ı|i|u|ü|o|ö</lex>
```

```
      <surf>b|c|d|g|j|l|m|n|ŋ|r|v|y|z|a|e|ı|i|u|ü|o|ö</surf>
     </pair>
    </suffix>
   </match>
   <exception part="suffix">çA</exception>
   <exception part="suffix">çAk</exception>
   <exception part="suffix">çAn</exception>
   <exception part="suffix">çAr</exception>
   <exception part="suffix">çI</exception>
   <exception part="suffix">çIk</exception>
   <exception part="suffix">çIl</exception>
   <exception part="suffix">nçI</exception>
   <exception part="suffix">çAk</exception>
   <exception part="suffix">ke</exception>
   <exception part="suffix">kay</exception>
   <exception part="suffix">key</exception>
   <exception part="suffix">tAy</exception>
  </rule>

  <rule index="9"  name="Küçük_Ünlü_Uyumu" applied_stem_type ="hepsi" valid="true"
force_match="true" force_inside_for_suffix="false">
   <match>
    <stem loc="last" type="vowel">
     <pair>
      <lex>e|i|a|ı</lex>
      <surf>e|i|a|ı</surf>
     </pair>
    </stem>
    <suffix loc ="first" type ="vowel">
     <pair>
      <lex>e|i|a|ı</lex>
      <surf>e|i|a|ı</surf>
     </pair>
    </suffix>
   </match>
   <match>
    <stem loc="last" type="vowel">
     <pair>
      <lex>u|o|ü|ö</lex>
      <surf>u|o|ü|ö</surf>
     </pair>
    </stem>
    <suffix loc ="first" type ="vowel">
     <pair>
      <lex>a|e|u|ü|1|2</lex>
      <surf>a|e|u|ü|1|2</surf>
     </pair>
    </suffix>
   </match>
   <exception part="suffix">ken</exception>
   <exception part="suffix">yor</exception>
   <exception part="suffix">mtrak</exception>

  </rule>
  <rule index="10"  name="Küçük_Ünlü_Uyumu_içerde" applied_stem_type ="hepsi" valid="true"
force_match="true" force_inside_for_suffix="true">
   <match>
```

```
    <stem loc="last" type="vowel">
     <pair>
      <lex>e|i|a|ı</lex>
      <surf>e|i|a|ı</surf>
     </pair>
    </stem>
    <suffix loc ="first" type ="vowel">
     <pair>
      <lex>e|i|a|ı</lex>
      <surf>e|i|a|ı</surf>
     </pair>
    </suffix>
   </match>
   <match>
    <stem loc="last" type="vowel">
     <pair>
      <lex>u|o|ü|ö</lex>
      <surf>u|o|ü|ö</surf>
     </pair>
    </stem>
    <suffix loc ="first" type ="vowel">
     <pair>
      <lex>a|e|u|ü</lex>
      <surf>a|e|u|ü</surf>
     </pair>
    </suffix>
   </match>
  </rule>
  <rule index="11"  name="Ses_Uyumu" applied_stem_type ="hepsi" valid="true" force_match="true" force_inside_for_suffix="false" stem_based_optional="false">
   <match>
    <stem loc="last" type="char">
     <pair>
      <lex>vowel</lex>
      <surf>vowel</surf>
     </pair>
    </stem>
    <suffix loc ="first" type ="char">
     <pair>
      <lex>consonant</lex>
      <surf>consonant</surf>
     </pair>
    </suffix>
   </match>
   <match>
    <stem loc="last" type="char">
     <pair>
      <lex>consonant</lex>
      <surf>consonant</surf>
     </pair>
    </stem>
    <suffix loc ="first" type ="char">
     <pair>
      <lex>vowel</lex>
      <surf>vowel</surf>
     </pair>
    </suffix>
```

```
      </match>
    </rule>

  </rules>
</phonology>
```

A.3 Kazan Tatar Phonological Rules

```
<?xml version="1.0" encoding="utf-8" ?>
<phonology>
  <alphabet>
    <character index="1" type= "vowel">a</character>
    <character index="2" type= "vowel">ä</character>
    <character index="3" type= "consonant">b</character>
    <character index="4" type= "consonant">c</character>
    <character index="5" type= "consonant">ç</character>
    <character index="6" type= "consonant">d</character>
    <character index="7" type= "vowel">e</character>
    <character index="8" type= "consonant">f</character>
    <character index="9" type= "consonant">g</character>
    <character index="10" type= "consonant">ğ</character>
    <character index="11" type= "consonant">h</character>
    <character index="12" type= "vowel">ı</character>
    <character index="13" type= "vowel">i</character>
    <character index="14" type= "vowel">í</character>
    <character index="15" type= "consonant">j</character>
    <character index="16" type= "consonant">k</character>
    <character index="17" type= "consonant">l</character>
    <character index="18" type= "consonant">m</character>
    <character index="19" type= "consonant">n</character>
    <character index="20" type= "consonant">ñ</character>
    <character index="21" type= "vowel">o</character>
    <character index="22" type= "vowel">ö</character>
    <character index="23" type= "consonant">p</character>
    <character index="24" type= "consonant">q</character>
    <character index="25" type= "consonant">r</character>
    <character index="26" type= "consonant">s</character>
    <character index="27" type= "consonant">ş</character>
    <character index="28" type= "consonant">t</character>
    <character index="29" type= "vowel">u</character>
    <character index="30" type= "vowel">ü</character>
    <character index="31" type= "consonant">v</character>
    <character index="32" type= "consonant">w</character>
    <character index="33" type= "consonant">x</character>
    <character index="34" type= "consonant">y</character>
    <character index="35" type= "consonant">z</character>
    <character index="36" type= "vowel">A</character>
    <character index="37" type= "vowel">I</character>
    <character index="38" type= "consonant">G</character>
    <character index="39" type= "consonant">D</character>

  </alphabet>
  <substitutes>
    <substitute index="1"  name="G|D"  valid="true" force_match="false"
force_inside_for_suffix="false">
      <match>
        <suffix loc="first" type="char">
          <pair>
            <lex>G|D</lex>
            <surf>k|g|d|t</surf>
          </pair>
        </suffix>
      </match>
      <action>
        <suffix loc="first" type ="char">
          <pair>
            <lex>G</lex>
            <surf>g</surf>
            <surf>k</surf>
          </pair>
          <pair>
            <lex>D</lex>
            <surf>d</surf>
            <surf>t</surf>
          </pair>
        </suffix>
      </action>
    </substitute>
    <substitute index="2"  name="A|I"  valid="true" force_match="false"
force_inside_for_suffix="false">
      <match>
        <suffix loc="any" type="char">
          <pair>
            <lex>A|I</lex>
            <surf>a|e|ı|í</surf>
          </pair>
        </suffix>
      </match>
      <action>
        <suffix loc="any" type ="char">
          <pair>
            <lex>A</lex>
            <surf>a</surf>
            <surf>e</surf>
          </pair>
          <pair>
            <lex>I</lex>
            <surf>ı</surf>
            <surf>í</surf>
          </pair>
        </suffix>
      </action>
    </substitute>
  </substitutes>
  <rules>
    <rule index="1"  name="Yumuşama"  applied_stem_type ="isim" valid="true" force_match="false"
force_inside_for_suffix="false">
      <match>
        <stem loc="last" type="char">
          <pair>
            <lex>p|k</lex>
            <surf>b|g</surf>
          </pair>
        </stem>
        <suffix loc ="first" type ="char">
          <pair>
            <lex>vowel</lex>
            <surf>vowel</surf>
          </pair>
        </suffix>
      </match>
      <action>
        <stem loc="last" type ="char">
          <pair>
            <lex>p</lex>
            <surf>b</surf>
          </pair>
          <pair>
            <lex>k</lex>
            <surf>g</surf>
          </pair>
        </stem>
      </action>
    </rule>
    <rule index="6"  name="Daralma"  applied_stem_type ="fiil"   valid="true" force_match="false"
force_inside_for_suffix="true">
      <match>
        <stem loc="last" type="char">
          <pair>
            <lex>a|e</lex>
            <surf>ı|i</surf>
          </pair>
        </stem>
        <suffix loc ="first" type ="char">
          <pair>
            <lex>y</lex>
            <surf>y</surf>
          </pair>
        </suffix>
      </match>
    </rule>
    <rule index="5"  name="Ses_Düşmesi" applied_stem_type ="hepsi" syllable="2" valid="true"
force_match="false" force_inside_for_suffix="false" stem_based_optional="true">
      <match>
        <stem loc="last" type="vowel">
          <pair>
            <lex>u|ü|ı|i|í</lex>
            <surf>1|2|3|4|5</surf>
          </pair>
        </stem>
        <stem loc="last" type="char">
          <pair>
            <lex>consonant</lex>
            <surf>consonant</surf>
          </pair>
        </stem>
        <suffix loc="first" type ="char">
          <pair>
            <lex>vowel</lex>
            <surf>vowel</surf>
          </pair>
        </suffix>
      </match>
      <action>
        <stem loc="last" type ="vowel">
          <pair>
            <lex>u</lex>
        <surf>1</surf>
       </pair>
       <pair>
        <lex>ü</lex>
        <surf>2</surf>
       </pair>
       <pair>
        <lex>ı</lex>
        <surf>3</surf>
       </pair>
       <pair>
        <lex>i</lex>
        <surf>4</surf>
       </pair>
       <pair>
        <lex>í</lex>
        <surf>5</surf>
       </pair>
      </stem>
     </action>
    </rule>

    <rule index="7"  name="Büyük_Ünlü_Uyumu" applied_stem_type ="hepsi" valid="true"
force_match="true" force_inside_for_suffix="false">
     <match>
      <stem loc="last" type="vowel">
       <pair>
        <lex>a|ı|u|o</lex>
        <surf>a|ı|u|o</surf>
       </pair>
      </stem>
      <suffix loc ="first" type ="vowel">
       <pair>
        <lex>a|ı|u|o</lex>
        <surf>a|ı|u|o</surf>
       </pair>
      </suffix>
     </match>
     <match>
      <stem loc="last" type="vowel">
       <pair>
        <lex>e|i|ü|ö|í</lex>
        <surf>e|i|ü|ö|í</surf>
       </pair>
      </stem>
      <suffix loc ="first" type ="vowel">
       <pair>
        <lex>e|i|ü|ö|í</lex>
        <surf>e|i|ü|ö|í</surf>
       </pair>
      </suffix>
     </match>
    </rule>
    <rule index="7"  name="Büyük_Ünlü_Uyumu_içerde" applied_stem_type ="hepsi" valid="true"
force_match="true" force_inside_for_suffix="true">
     <match>
      <stem loc="last" type="vowel">
```

```
          <pair>
            <lex>a|ı|u|o</lex>
            <surf>a|ı|u|o</surf>
          </pair>
        </stem>
        <suffix loc ="first" type ="vowel">
          <pair>
            <lex>a|ı|u|o</lex>
            <surf>a|ı|u|o</surf>
          </pair>
        </suffix>
      </match>
      <match>
        <stem loc="last" type="vowel">
          <pair>
            <lex>e|i|ü|ö|í</lex>
            <surf>e|i|ü|ö|í</surf>
          </pair>
        </stem>
        <suffix loc ="first" type ="vowel">
          <pair>
            <lex>e|i|ü|ö|í</lex>
            <surf>e|i|ü|ö|í</surf>
          </pair>
        </suffix>
      </match>
    </rule>

    <rule index="8"  name="Ünsüz_Uyumu"   applied_stem_type ="hepsi"  valid="true"
force_match="true" force_inside_for_suffix="false" stem_based_optional="false">
      <match>
        <stem loc="last" type="char">
          <pair>
            <lex>ç|f|h|k|p|s|ş|t</lex>
            <surf>ç|f|h|k|p|s|ş|t</surf>
          </pair>
        </stem>
        <suffix loc ="first" type ="char">
          <pair>
            <lex>ç|f|h|k|p|s|ş|t</lex>
            <surf>ç|f|h|k|p|s|ş|t</surf>
          </pair>
        </suffix>
      </match>
      <match>
        <stem loc="last" type="char">
          <pair>
            <lex>b|c|d|g|j|l|m|n|ñ|r|v|y|z|q|w|x|a|e|ı|i|u|ü|o|ö|í|ä</lex>
            <surf>b|c|d|g|j|l|m|n|ñ|r|v|y|z|q|w|x|a|e|ı|i|u|ü|o|ö|í|ä</surf>
          </pair>
        </stem>
        <suffix loc ="first" type ="char">
          <pair>
            <lex>b|c|d|g|j|l|m|n|ñ|r|v|y|z|q|w|x|a|e|ı|i|u|ü|o|ö|í|ä</lex>
            <surf>b|c|d|g|j|l|m|n|ñ|r|v|y|z|q|w|x|a|e|ı|i|u|ü|o|ö|í|ä</surf>
          </pair>
        </suffix>
      </match>
    </rule>

  </rules>
</phonology>
```

B. EVALUATION DATA

B.1 Kırghiz - Turkish Translation (Tale 1 - Ayıldın Baldarı: Village Children)

Kapçıgay ördögön cüktüü maşina taş moynoktu aylana berip tık toktodu. Kabinanın sol cak kaalgası açılıp, andan beri tura kalgan marçaygan ulan beri üstündögülörgö kolun cañsap buyurdu:

- Catkıla! Kömköröñördön soylop kalgıla! - Kuzovdo caymalangan arpa, arpa üstündö tigige teñtuş eki bala bar ele. Munu ugarı menen biri-birine köz kısıp kımıñ külgön eköö arpaga katar soyloştu. Ubada uşunday ele.

Künü keçke maşina tosup, çarçap suy cıgılganda mına uşuga carmaşkan baldar. Ertelep col boyuna çıkkan bular cogorton çañ körüngön sayın dır koyup tura kala berip suy cıgılbadıbı. Kol kötörüp katar turgan baldarga toktoboy ötüp cattı maşinalar. Oo, beşimge oogondo mına uşu maşina ötö berip toktop, terezesinen beri başbakkan şofor özünö carışa çurkap cetken uşul üçöönön suradı:

- Kayda barasıñar?

- Frunzege! Okuuga! - Carışa süylöp, carışa entigip turgan baldarga bir sıyra köz toktotup algan kabelteñ şofor carılıp ketken erdin calap mintti:

- Bir gana orun bar. - Mına uşu tapta kabinaga çap carmaşa kalgan bala entige, şaşa süylödü:

- Bayke! Biz çoguu barat elek.

- Emne kılayın? - Şofor iynin kuuşurdu. – Başka orun cok.

- Abılbek! - dedi baldardın biri tigi kabinaga carmaşıp turgan balag - Sen kete berseñçi anda, biz artıñdan…

- Cok, çoguu barabız. - Kabinaga carmaşkan bala kese süylödü.

- Boluptur anda, çoguu kelgile. Arı tur bala! - Şofor cönöy turgan boldu ele Abılbektin kolu tutkadan ketpey, kabinaga carmaşkan kalıbında zarıldı:

- Bayke, ala ketseñiz, bayke? Institutka barat elek. Üstünö ele oturup alabız, bayke?

Özün karmap turganga bir ese açuusu kelgen şofor akırı bul akidey asılgan cubarımbekterden kutulbasına közü cettibi, Abılbekti karap kıcırdana burk etti:

- Kel! Anda sen mında otur. Birdeke bolso tigiler üçün sen coop beresiñ! Ey! A siler üstünö çıkıla! Arpanın üstünö… Azır kabinaga daldalanıp otura bergile. Nepaada mına bul, - dedi canındagı Abılbekti tigilerge körgözüp, - silerge belgi berse, oşo zamat arpanın üstünö booruñardan soylogula da üstüñördön mınabu kendirdi çümkönüp algıl Kayra tur demeyin dımıñar çıkpasın! - Baldar süyünüp ketti. – Makul bayke.

- Tük baş kötörböybüz. - Çınında, bulardı mına uşul arpanın astına tirüülöy köömp taştasa da makul boluşmak. Arpa üstünö cabılgan kaldagay kendirdin bir çeti menen kurcun-keçegin caap, bir çetin özdörü tizesine çeyin capkan baldar, tigi şofor aytkanday kabinaga cölönüp arttı karap oturuştu.

Tigine, keçke col toskon bulardın ene-atasınday eerçiy karagan kıştak kaldı artt Tuulgandan bulardın oyun-külküsünö ortok, soguş künündö bular menen Birge ıylap, bir soorongon kıştagı. Bular keçke mal cayıp, bular keçke türkün oyun oynogon tetigi çoñ sunun kök araldarı, capan talduu tokoyu, toonun caşıl çıbırları kaldı. Kekçi kün magdırap, balkıp catkan caykı kıştaktı, caşıl çıbırlardı köz ayırbay karap baratıştı maşina üstündögü eki bal Bular körgöndü uşul azır körböy baratkan tigi şofordun canındagı Abılbek. Köñülü abalap uçat.

Caşıl cibek köynöktün büyürmösüdöy cazgı çıbırlar bul kıştaktın malı-canın camgırdan kiyin özünö çakırçu. Koy-kozu kök kuup, cıbıt-cılgaga cabalaktap, baldar koy tekey terip çıbırlardı kırdap çurkap, kırdankır aşkan baldar kızuu menen tee uluu too köödönünö çeyin çıgıp ketişçü da keede caanga kalıp, taştın cansarına korgoloşçu. Toolordu bir-birine kabıştırgan kün kökürök, tünörgön asmandı oyku-kaykı tilip, caracara çapkan çagılgan, şarkırata tökkön nöşör… Tekeyden karargan oozdoru carım açılgan baldar çagılgan çart etken sayın kulaktarın basa, közdörün cılk cumup ciberişçü.

Baldardın köz aldında mına uşular. Baldar menen ayıldın arkıragan şamalı koşo kelet, tigi kök kaşka dayra koşo kelet çalkıp. Cazgı camgırdın ilebi kelet, koy tekey menen kozu kulaktın daamı kelet. Şaar cönündö, okuu cönündö ukkandarın estep içteri uygutuyu. Ene-atasının koburugu ugulat kulakka: " Oo, şaarda kıyın…", "Ee, degi könö alışar beken şaarga?" "Degele oşo okuşka ötö alışar beken?" - Iyri oturuşup munu koburaşkan ene-ata baldarına özünçö da kündöptündöp kulagına kuygan. "Emi çoñ şaarga baratasıñ, mındagıday enöö bolbo. Men seni eerçiy cürböym,

boyuña tıkan bol. Anan, canındagı tıyın-tıpırıñdı aldırıp iybe. Şaar ısık. Mındagıday eençilik cok. Kokuy, anan çañkaganda muzdak suu içe körbö, bezgek bolosuñ. Darbızdın danegine da sak bol, al ketse da oşondoy ooruga çaldıgat deçü". Aytor mınday akıldı ugup oturgan balanın başı mañ. "Şaar kaynagan ısık, el köp bolso, anan oşol ısıkta mındagıday ayran, kımız tügül suu da cutuuga bolboso kişi degi kantip tirüü gana kalar eken?" Mına azır da uşunu oylogon baldar arkıragan celden kere cutuşup, kere dem alıp, arttagı caşıl çıbırlardı sugula tiktep kelatıştı: Munun baarı erteñ cok. Kaktagan ısıkta suu tappay çañkap turabız go". Bulardın oyun toktogon maşina üzdü:
- Bat soylop kalgıla booruñardan! Abılbektin buyrugu bulardı şaştırıp saldı. Boortoktoy soylop, kendirdi budalaktay çümkönüştü.

- Catıştı, kettik, bayke. - Abılbek şofordu karadı. Uşunu kütkön maşina ordunan kozgoldu. Öz buyrugu menen baldardı catkırıp, öz buyrugu menen maşina cürgüzgönünö korston Abılbek marçayıp oñdono oturdu. Irdap iyçüdöy kakırınıp ünün casap koydu. Uşul azır mınabu şofor kokus sözgö aralaşıp: " Kana cigit, ırdap koyboysuñbu?" dese, Abılbek tartınmak emes. Ayılda baya kan kakşap cürçü ırlardın birin koyo bermek. Abılbekti antip sözgö almak tügül mınabu şofor ün katıp koyboyt.

Abılbek uşu kişini büşürköy karap kelattı. Biyik kabak, kalbagay erindüü, oñkogoy murun kişinin dem alganı da cay belem, anda-mında gana töşü saal kötörülö tüşüp basınat. Sol uurtunda tırtıgı bar eken. Mayluu-köölü barbaygan balban mancanı rulga birotolo celimdep taştaganday, karmagan ordunda lapıyat. Abılbek munun kolun tiktep, kolun büşürköp kelattı. Kudu ele Abılaydın kolu. Cazında traktor aydap, küzündö kombayn aydap cürçü Abılaydın çorluu kabelteñ kolunun ele özü.

B.2 Kırghiz - Turkish Translation (Tale 2 - Iyık Sezim: Sacred Emotion)

Iyık sezim - bul uçuna köz cetpegen biyik çoku!

Köz aldımda kök tiregen biyik aska turat. Aga karaym da kızıgam, kızıgam da oylonom: bul diynödö seni körbögön, sana suktanbagan kişi az çıgaar çındıgınd Antkeni, saga çıguu tilegi ar bir kişinin tabiyatına bütkön şirin eñsöölörü menen cuurulgan. Adatta, kişilerdin tunuk sezimderi biyikke umtulup. Seni mara kılat. Cerdin dal özögünön önüp çıkkansıp, asman künügö bagıt algan ukmuştay zalkar bolsoñ da, sende bayırkı adamdardın cıluu cürögünün cönököy izderi catat...

Eki adam batpagan tar çıyır beline orolo öödölöyt. Bir taalay çıyırı, bakıt çıyırı. Ar bir kişi bul çıyırga öz aldınça iz kaltıruuga tiyiş. Al emi, saga kötörülüp, senin töböñö tamanın tiygizgender kaygı - kasiret, ıza - korduk, üröy - korkunuç degenderden müldö arılat. Töbölükkö beykuttuk ornop, çoktuktun, cetişpestiktin kişeni bıt - çıt bolgonsuyt. Andan arı tımızın sagınıçtın, kubanıçtın sezimderi oygonup, bütkön boy balkıyt.

Düynögö adam bolup caraluunun sıymıgı terendep baş kötöröt.

O, sıykırduu, aska! Sende bulutsuz asman köp. Kün nuru birinçi tiyip, akırkı nurları seni menen koştolot. Sende kün nuru mol. Töböndö salkın cel cüröt da, çerler cazılıp, köñüldör cumşarat. Denege cagımduu kan tarap, tattuu çımırkanuu payda bolot, Kişige bütkön marttıktın, erdiktin, meerimdin, berilgendiktin calını aloolonot. Mına oşondo, ubakıt toktop, büt düynö bir saam es ala tüşkönsüyt...

Iyık sezim - bul nazik gül!
Caz kündörü açılgan bul güldün ömür küçü cıldar boyu sozulat. Kişinin can düynösün tolkutkan naziktiktin, koozduktun, köñüldü ergitken cagımduu cıttın soolbas bulagı öñtölöt al. Köp sandagan kooz köpölöktör anı tegerene uçat, alardın da kıpınday armanının bir maarası oşol öñdölöt.

Kırmızı barkut, kat - kabat celekçeler tenselgen caşıldık arasınan boy kötöröt. Adatta anın şoola çaçıp turganın körösüz. Karap tursañız, ömürdün tübölük öçpös şamı öñdüü boygo kubat, kıymılga medet. Al kişilerdin küügüm tartkap kökürök boştuguna caygaşıp, nur çaçıp turgan kan kızıl gauhar taştı elestetet. Anın sıykırına kabılgandar tiriçilik tüyşügün unutat. Karagan sayın karagısı kelet. Oşondo al başka bardık kızıkçılıktan keçüügö macburlayt. Anın cıtı adam seziminin eñ tereñine cetet. Kişide anı menen ömür boyu ırakattanuu tilegi özünön özü payda bolot.

Birok, al nazik, arı sezgiç gül. Sizden çıdamduuluktu, adal kütümdü talap kılat. Al aram oy menen cılmaygan kişi aldında sooluy baştayt. Nurlanıp kulpunuusun toktotot.

Iyık sezim - bul kaarduu deñiz!

Çeksiz ketken kaarduu deñiz özünçö köölgüyt. Adam balası kımbat baalagan şuru menen bermettin esepsiz baylıgı munun tereñine caşırıngan deşet. Birok, andakı bermet menen şor suu miñ sandagan adamdardın köz caşın tüşündürüügö tiyiş...

Adamdar aga tüşünüügö kumar. Antkeni, adatta al meerimdüü deñiz. Aga tunuk bulaktar, dartka dabaa araşan suular kuygandıktan tüşkön adamdardı külgündöy tazartıp, ömürgö degen süyüüsün tebeñdetet.

Anın kaardanuusunun sebebi emgiçe açılbagan sır. Al kaarına alganda çeekterine ak köbük bürküp, tüyülö cıyrılıp, cindene baştayt. Tüsü kara küröñ tartıp buzulat. Üstündö oor buluttardın kerbeni kırkalayt. Çagılgandar ot kamçısın üyrüp, kara tumandı tilgilep çartıldayt. Oor deñiz tüp kötörülö kozgolup, bir bel aşıp ıldıy kulagan, toodoy tolkundar örköçün cönötöt. Üröydü uçurgan dülöy kürüldök terenden ugulat, candan tüñültöt. Büt düynönü capırıp, tebelep ötçüdöy küülönöt, ceekterin tıtkılayt. Turuştuk bere albagan toolor urap, cerler cemirilip catkansıyt. Aga kabılgan adam çabal oydon tez kutulat, kamgaktay uçup, tıyanaksız, ümütsüz aldastayt. Aga carık düynö menen koştoşuu baarınan ceñil sezilet!

Tabiyat küçünün uluu kocosu bolso da, adam balası özün alsız sezip, çenemsiz kaygıga bataar uçuru oşondo payda bolot.

Iyık sezim - bul süyüü!

B.3 Kırghiz - Turkish Translation (Tale 3 – İşenböö: Not Believing)

İttin köñülü cay, büün mışık eköö dostoşot. Mından kiyin eköö birin-biri körgöndö murundarı tırışpayt, eç kim da alardı bir-birine "kas" dep aytışpayt. Bassa-tursa ele "av-av" deçü it şimşip cürüp tapkan bir kesim mayın aldı da mışıktı izdep cönödü.
Bayatan kaparsız küngö kaktangan mışık ittin şıbırtın alda kaydan sezdibi, cerge andan-mından bir tiyip zımıradı.
- Tokto… saga dostoşkonu keldim… Tokto… - It mışıktın artınan çurkadı.
"Meni çındap kubalagan eken" - dep mışık andan beter kaçtı. A it bolso: "Bügün çıgınganda dostoşoyun, erteñ unutup kalam ce mayımdı özüm cep salsambı" - dep artınan buçkaktadı. Mışık tikendi aralasa, tikendi araladı, teşikten ötsö teşikten öttü, emneden sekirse oşondon sekirdi…
Mışık can talaşıp terektn başına çıga kaçtı. Karasa ittin tili salañdap, közü kızargan.
- It! Ittigiñ koyboduñ… Emne kubalaysıñ?
- Sen emne kaçasıñ?
- Sen emne kubalaysıñ?
- Saga dos boloyun dep…
- Kalp… kubalap dostoşosuñbu?
Kıykırışıp catıp eköönün ündörü da büttü.
- Aldagı agargan emne? - dedi mışık terektin başınan.
- May… Saga bereyin dep, - dedi it tömöndön.
- "Bul aldap atat. It kantip maga dos bolsun", - dep mışık terekten tüşpödü.
Kün caap kirdi.
"Meyli, kün caasa… Kolumda may turganda dostoşposom, kiyin takır işenbeyt", - dep it ketpedi.
"Ii, bul kündün caaganına karabay meni añdıdı" - dep mışık tereke ulam öydölöp çıktı.
Eköö biri-birine işenbedi.
Caan könöktp tögö berdi.

B.4 Kırghiz - Turkish Translation (Tale 4 – Mebel: Furniture)

"Mebel alabız… Mebel alabız", - dep cürüştü atam, apam. Men da köçödögü baldarga maktandım:

- Biz mebel alabız… Mebel alabız.

Erik anı bilbeyt eken: "al emine, tehnikabı dep suradı.

- Eñ sonun birdeme… Eñ sonun. Saga sözsüz körgözöm, - dedim.

Bir künü kirsem ele üyübüz kooz bolup kalıptır. Cıgaç deyin desem küzgüdöy caltırayt. "Omiyin" dep karmalap ciberdim.

- Kokuy, tokto, bulgaysıñ, koluñdu cuugun, - dedi apam.

Cuunup, aarçıngandan kiyin mebeldi köröt eken.

Oşentip, men mebel menen taanışkam.

Mebel turgan törkü üygö kirip baratsam ele apam "şımıñdı çeçip kir, koluñda mık cokpu, çiyip iybe, çukup iybe" dep tekşerip turat. Emne üçün antet, bilbeym. Menin törkü üygö takır kirgim kelbey kaldı. Al caktın terezesi sonun bolçu: daraktardın başında çımçıktar ırdap olturçu, cazında çıyırçıktar kelip konçu, çabalakeyler kubalaşıp oynoçu. Alar emi mebeldin ar cagında kalıştı. Bir künü murdum kıçışıp, tanoomon suu aktı. "Kokuy, etiñ ot menen çok, oorup kalıpsıñ" dep apam mebel-komnataga catkırıp koydu. Men ooruganıma ayabay süyündüm. Akırın turup terezege bardım: kursaktarı çoñ-çoñ çımçıktar oturuptur - tündö Indiyadan uçup kelişse kerek. Baldar suu keçip carışıp cürüşöt. Erik başına cırtık çaka kiyip çurkap baratıptır, beline calbıraktardı baylanıp alıptır.

- Erik! Erik, berkel. Meni kördüñbü?

- Anıy, barbaym, apañ uruşat, - dedi Erik.

- Apam üydö cok!

Erik kirip keldi. Al mebeldi körgön cok. Kolu suu, şımına batkak cabışıp kalıptır.

- Tokto, şımıñdı çeç, - dedim men Erikke.

Al eçteme tüşüngön cok.

- Koluñda mık cokpu? Çiyip iybe, çukup iybe… Bizdin mebel…

- Kanakey mebel? - dedi al dagı körböy.

Añgıça apam kirip keldi, al Erikti kolunan karmap eşikke çıgardı.

Oşondon kiyin Erik bizdikine baş bakpayt.

Menin apam eñ cakşı, apamdı eç kimge teñebeym, caltıragan mebelge dagı.

Apam, keede atam da mebelin kebez menen sürüşöt. Baarı bir men mebelge abdan kapamın. Köp mebelden körö maga törkü üydün terezesi cakşı boluçu.

Özümdün bul oyumdu apama, atama ayta albay cüröm.

B.5 Kırghiz - Turkish Translation (Tale 5 - At Cakşı Körgön Bala: The boy horse loves)

Asıltay bacırañdap süyünüp üygö kirdi:

- Meni at cıttap koydu. Meni at cakşı köröt!

- Kalpıçı, - dedi Abıltay aga işenbey. - At cakşı körgöndü bilbeyt.

- Bilet! Cürçü, körgözöyün.

Eköö ceteleşip mamıga baylanuu attın canına barıştı. Asıltay akırın aldıga öttü da attın tizginin karmap, kökülünön sıladı. At bılk etpedi, oozdugun şıldırata başın salañdatıp Asıltaydın iymeygen kolun cıttadı, tumşugun iyinine süyödü.

- Saga süylöp atabı? - dedi Abıltay, attın kıpkızıl tanoosunan, arsaygan tişterinen korkup.

- Süylögön cok. Cakşı körüp catat.

- Emne üçün cakşı köröt?

- Men anı sugargam. Arıktın boyunan otkozgom. - Asıltay derdeyip koydu. Attı mamıdan çeçip, arkı dümürgö tartıp minip aldı.

Abıltay at kayakka bassa eerçiy çurkap, at menen Asıltay eköönün biri-birin cakşı körüşkönünö süyünüp, özü da oşondoy bolsom dep kıykırıp cürdü:

- Asıltaydı at cıttap koydu! Asıltaydı at cakşı körüp koydu.

Bir künü Abıltay attın canına bardı da, kulagına şıbıradı.

- Oy, at. Meni cakşı kör, - dedi söömöyü menen közgö sayıp. At selt etip başın kötörüp, kayradan ürgülöp turup kaldı.

- Oy! Meni cıtt Cakşı kör, - dedi attı keketip.

At dagı ele unçukpadı. Abıltay anın murduna muştumun takap çañırdı.

- At! Seni öltüröm!

At kulagın tikçiytip, oozdugun kaçır-kaçır çaynadı. " A-a, korkutup saldım" dep oylodu Abıltay:

- Meni cakşı körbösöñ öltüröm! - Al çöntögüñdögü mıgın aldı da, attı murunga sayıp iydi.

At tanoosun tarsaytıp, közün çakçarılta tikesinen sekirdi. Abıltay bakırıp cıgıldı.

Oşondon kiyin Abıltay atka cakındabadı. Cakındasa ele kulagın çunañdatıp, koşkurup, közün çakçañdatat.

Asıltayga bolso anpeyt. Al kayakka cetelese ce minse dele kaalagan cagına kete beret. Anı bu köçödögü baldar, al turgay çoñ kişiler da "at cakşı körgön bala" dep ataşat. Mına, Abıltaydın inisi Asıltay! Şaarga ketkiçe çoñ atasının atın minip, köp cerge bastırıp cürdü. Abıltay bir da colu özünçö minip bastırgan cok.

Kızık, emne üçün Asıltaydı at cakşı kördü? Emne üçün Abıltaydı caman kördü?

B.6 Kazan Tatar - Turkish Translation (Tale 6 – Kiyim: Stone Dress)

Bonn zamanda bir han bulıp, elle ni öçin gine üzinifi halkına açulanmış. Bir comga könni mescidten çıkkaç han hezretleri mescid yanındagı bir zur taşka kulı bilen kürsetip, cıyılgan halkına:
-Kileçek comgaga çaklı şuşı taştan minim öçin kiyim tigifüz. Tigip ölgirte almasafuz, barlığınız m ütirteçekmin, -diyip öyine kaytıp kitti. Cemegat ise heyran kalıp, taştan kiyim tigü mömkin tügil idikinden her kayusı bir-birsine karap, ni kifieş iterge de bilmeyince, küz yaşlerini agızıp, öylerine kayttılar.
Bolar arasında citmiş yaşlerinde bir bik yarlı kart bar idi. Öyine kaytıp kirdi de, urmdıkka ultırıp, kart üksip-üksip iglarga başladı. Karçıgı sebebini sorasa da, karşı bir süz de eytmiyçe, kart iglavmda gına buldı.
Kartnm buyga çitken bik matur, gakıllı, zirek bir kızı bar idi. Kız, işikten kirip, atasının iglaganını kürgeç, sebebine soradı. Kart, eytirge tilemese de, kızının küfiili kalmasın öçin, mescid aldındagı işni söylerge başladı.

-İy kızım! Kart atannın gomiri tik cidi gine kön kaldı... Kileçek comgada han bizni ültirteçek...-dip, tagı iglıy başladı.

-Sabır it, atam, ni öçin sizni han ültirteçek? Zinhar, söylep birsene!
Kart bik kaygılı tavış bilen söyliy, kız ise çm künilinden atasının süzlerin tıfilıy idi. Kartnıfi süzleri bitkeç, kız kölip hem biraz uylanıp torgaç:
-Kunkmamz, atam, bu bir de kurkırlık iş tügil? -didi.
Kart:
-Niçik alay bir de kurkırlık eş tügil? Taştan kiyim tigip bulamı sofi?
-Mine, atam, kileçek comga kön han: "Taştan kiyim tiktinizmi?"-dip soragaç: "Han hezretlere, biz ülçev almainça tige almadık. Ülçevsiz tigilgen kiyim yeki bik ozm, yeki bik kıska bulıp çıgar idi. Şuran öçin de biz hezir buyınıznı ülçep alıyk ta, siz üziniz möbarek kulmız bilen kisip biriniz! biz, baş östi, lkileçek comgaga çaklı kiyiminizni tigip bitireçekbiz!"-diyip eytigiz.
-Yuk, kızım kurkam... Niçik ittirip hanga karşı süz kaytarıp torırga?
-Bir de kurıkmagız, atam alla tilese, birnerse de bulmas!..
2
Comga kön çitti. Halik mescidke cıyılıp bitkeç, han kildi. Comganı ukıp mescidten çıkkaç, han halıkka taba karap:
-Kiyim temam buldımı?-dip soradı.
Bir kim de cavab birmedi. Şundagı kişiler, başlarını bögip, aldılarına karap tik tordılar. Birniçe minuttan sofi, eliği kart alga çıgıp, kaltırap kına kızı öyretkenni söyledi. Han:
-Bolay eytirgi sina kim öyretti? dip soradı. Kart aptıradı... Ni dip eytirge de bilmedi. Tagın birniçe minut ütkeç, kart kurıkkan tavış bilen gine "kızım" diyip cavab birdi.
-Alay iken. Eyde minim artımdan,-diyip, han öyine kitti hem kartka üzi artından barırga kuştı. Biçara kartnın başımnı kiserge ilte diyip, kotı oçtı. Şulay da hannın, boyırıgmdan çığa almadı, akrın gına han kapkasma taba kitti... "İndi bittim!" diyip ükinip, içinden gine imanın ukıp, tevbesin kıylıp bara idi.
Han, öyge kirgeç, hizmetçilerine aşhaneden yigirmi dane pisken tavik yomırkası kitirirge kuştı. Yomırkalar kilgeç:
-Mine, kart, şuşı yomırkalarnı kızma alıp bar, mina yigirmi dane çibiş çıgarsm, digin. Eğer çibişler çıkmasa, kızın ile ikeviniznin de başlarmıznı kistireçekmin, didi.
Kart miskin, bir süz de cavap kaytarmainça, öyini kaytıp kitti. Anın yözi ülikniki tösli agargan, yöregine elle nindi kurku urnaşkan idi. Ul akrın gma kayta idi. Başı eylene hem küzleri üzinifi basacak cirin kürmiler idi. Ul kayttı. İşikten kirgeç te:
-Mine, kızım, sinin süzin bilen üzimni gine tügil, sini de belage töşirdim, dip, han boyırgan hizmetni söyledi.
Kız atası kulından yomırkalarnı alıp östelge kuydı da:
-Rehmet han hezretlerine, eydeniz, didi, atam anam! Ultırıp yahşi gına aşıyık!

İy kızım, kotırmasana! Han sina bu yomır-kalarnı çibiş çıgarırga ciberdi iç!-ülsim de, aşımnı aşagan, yeşimni yeşegin üzimden artık, kızım sini ayıym.

Kızı kölimsiredi de:

-Eydeniz, utırınız. Üzim erçip birim eli,-diyip, yomırkalarnı çat-çat vatıp, ata-anası aldına kuydı.

Kart bilen karçık evvel aşarga batırçılık kıylmasalar da, kızları aşagannı kürip, ni bulsa bulır diyip, aşarga totındılar. Kız:

-Mine indi niçik hanga rehmet eytmessiz? Bügin ikmekten başka aşıbız yuk idi. İndi ikmegibizni yomırka bilen aşıybız!

-Bir de borçılmafüz Çibiş öç comgasız çıkmıy. Ul vakıtka kader cavabımız hezir bulır...

Bir un könler çaması ütkeç, kız anasına:

-Anakay, bügin mifia bir çülmik tan butkası pişiriniz, -didi.

Butka pişip citkeç, kız çülmekni bir yavlıkka tördi de atasına birip:

Şuşı butkanı hanga alıp barmız hem, çibişler tiz çıgaçak, diyiniz. Han hezretleri şuşı butkanı çeçip, çibişler öçin yana yarma hezirlesin idi, diyiniz.

-Yuk, kızım, bara almıym, başımm kiser.

-Atam, bir de kurıkma, minim künilim öçin bar,-digeç, kart butkanı alıp kitti.

Hanga kirip kartnın kilgenin bilgirttiler. "Kirsin" digeç, kart, kirip, kızınnın süzini eytip, butkanı han aldına kuydı. Han kölip ciberdi de:

-Yarar, -kart, irtege kızınız mina üzi kilsin. Bırak, kilgende yul bilen de kümesin, yulsız da kümesin; bulekler de kitirmesin, büleksiz de kümesin: kiyimli de kümesin, kiyimsiz de kümesin; atlı da, ceyev de bulmasın!-didi.

Kart yane öyine iglıy-iglıy kaytıp, kızına süzlerini söyledi.

-Yarar, atam, irten min hanga barırmın...

İrte torgaç, kız kiyimlerini salıp, balık tota torgan cetmege çolganıp, keçe östine atlandı hem kulına bir çıpçık alıp yulga çıktı. Yul bilen tugrı barmainça, eli uranının bir yağına, eli bir yağına çıgıp, borgalanıp barıp, han kapkasma çitti. Han aldma hem keçesine atlangan kiliş kirip, hanga kulındagı çıpçıknı suzdı da:

-Siznifi öçin alıp kilgen bülegim!-didi.

Han almak bulıp kulnı suzganda kız çıpçıknı oçırıp ciberdi. Şundan son kız süzge başlap:

-Han hezretleri, boyırganınızça karşıfuzga kil-dim. Mini çakıruvınızdan moradınız eliği çibişlerge çeçe io~z*n tan tugrısındadır dip uylıym,-didi.

Han bir kürü bilen kartnın matur kızma gaşıyk buldı.

-İy matur kız, sin razıy bulsan, min sini üzimi hatınlıkka alır idim, -didi.

-Üzimni bik behitlilerden sanap, şatlık ile kabul item, -dip cavap birdi (kız).

Şunnan sofi bik zur tuy yasap, han kartnın kızın üz öyine aldırdı da rehet tora başladılar.

Bir-iki yıldan sofi bir cirge kiterge yulga çıktı. Han hatmi bilen isenleşken vakıtta:

-Min yuk vakıtta hiçbir işke de katışm Bir kiçirek kine iş çıksa da, hökim itken bulıp mataşm Eğer de üzinni tota almıyça, bir-bir hökim ite kalsan, minim öyimnen kitersin,-diyip, yulga çıktı.

Han kitkeç, küpmi-azmı vakit ütkeç, öç kişi hökimge kildiler. Alamın üz aralarındagı kıç-kırışları bu idi: By öç kişi birsinin atın-ikinçisinift arbasma, öçinçisinin dugası bile cigip yulga çıkkannar iken. Yulda barganda bolarga bir tay iyergen. İndi şul tay kaysına tiyişli?

Hanım küp karap tormainça:

Atnı arbadan tugarınız da: At iyesi atnı iyarttip, arba iyesi arbasın tartıp, duga iyesi dugasmı küterip, öçiniz öç yakka yöriniz. Kaysınız artından tay iyerse, tay şuna bulacak, -didi.

Elbette, tay atka iyerdi hem kıçkırış şunıfi bilen bitti.

Şundan son baytak vakit ütti. Han kayttı. Eliği öç kişi arasında bulgan kıçkırış hanga işitilgen. Han ha tınına:

-Sin atan yanına kayt!-digeç:

-Yarar, kaytırmın. Bırak, bu öyden bir-bir nerse alıp kiterge yararmı?

- Tilegen nerseftni al. Bırak, kit süzimni tınla-madın.

-Tagı bir süzim bar: üzin de bilesin, bir-iki il birge tatulıkta künilli gine könleribizni ütkerdik. İndi atam öyine kaytam. Bügin sofigı kiçebiz şul songı kiçibizni içmasam keyiflenip, aşap-içip ütkerik!

Han razıy buldı. Aşap-içip ultırganda hanım hannı baytak kına sıyladı. Yuldan bik arıp kayt-kanga küre, han ultırgan cirinde yokıga da kitti.

Hanım naçar gına bir çana cigerge kuştı hem iki hizmetçi çakırıp aldı da hannı kürsetip:

-Küteriniz de çanağa çıgarıp sahnız!-didi. Hizmetçiler ilik şakkatsalar da, hanım kuşkaç, tınlamıy çaraları bulmadı, yoklagan hannı çıgarıp çanağa saldılar.

Hanım üzi çıgıp ultırdı da, üzi atnı totıp, söyikli hanın atası öyine alıp kitti. İrte birle han uyanıp kuzini açıp karasa-üzinin bir tebenek kine öyde yatkanın hem hatının yanında kürdi. Şakkattı.
-Bu ni hel, min nik monda?
-Borçılmanız, söyikli han! Sizni min alıp kildim.
-Min safta: üzin kit, didim iç!

-Şulay, siz mina "tilegen nersenni alıp kit" dip te eyttigiz. İndi minim tilegen nersem-siz idiniz, siz-ni hem alıp kittim,-diyip, hannm iki kuzinden üpti.
-Atan barıp bizni kilip alsınlar dip heber itsin,-didi.
Kart tiz gine barıp heber itti. Andan altı at cigip kilip, han ile hanımnı alıp kittiler. Mondan son han ile hanım bik tmıç könler ütkerdiler, tatu tordılar hem birge üldiler de.

B.7 Kazan Tatar - Turkish Translation (Tale 7 - Ölüf,Yaki Güzel Kız Hediçe (Part I): Aleph or Beautiful Girl Hatice)

Hikeyemez bir gaceep güzel hikeyeder. Hikeyemezi yahşi anlatmak öçin, evvelen vokugısı ne cirde ve ne vakıtta uldıgını yazamız. Hikeye-mezden ukılıp anlaşılacak iş berniçe sene mokatdem beldeyi Kazanda vakıyg ulmış iştir.

Sabah segat unda, Kazan mösafirhanelerennen bir olug ve mogteber mösafirhane yanına küp adem-ner cemgı ulmışlar. Cemegarnin mabeennerennen bu kebi süzler işittiler idi:

-Ne var, ne karıysız, ne ulmış?

-Ne ulsrn, bu kön bu mösafirhanede bir mösafire hatırını katel itmişler.

-Gaceep! Kem katel itmeş, katil totılmışu?

-Ul kaderlese meglüm tügel, hezer politsiya hem sudebnıy sledovatelne köteler, teftiş ulır, belkem, katil de anlaşılır.

Ve ma eşbehe zalike, cemegat arasında küp törlü süzler söyleneder.

Yene biş-un dekıyka keçdekden sonra: "Keleler! Keleler!" -digen süz cemegat telende tekrar kılına başladı. Kelüçeler politsiya ile sudebnıy sledovatel idi. Nomerlerge mektüle hatın hosusında teftiş öçin keleerler idi. Atlar çitti. Teftişçeler, garebelerennen töşep, nomerge kirip ketteler. Cemgı ulmış cemegat nomer içene kererge nekader kasıd ittiler ise de, nomerge kerüden mengı kılındılar. İçerü kirüvden mengı kılınsalar da, cemegat nomir yanınnan taralmaymça, ne heber ulır iken dip, teftişçilernin çıguvına montazıyr bulıp toralar.

Teftişçeler ehvalene yazalım.

Teftişçeler nomerlerge kerdeklerinnen sonra: "Ülterelmeş hatın kaysı nomerde?" -diyü nomirler sahibinnen teftişçilernen berse soal kıldığında, nomirler sahibi Gabdullin: "11 nçe nomerde", -deyu cavap virip, nomirlernin hucasile teftişçeler 11 nçe nomereyi kerdeler. Şimdi nomeresene tegrif idelem: Bu 11 nçe nomere taş pulatnın 2 nçe etajında-kahnda ike bülmele nomire ulıp, uram tarafında öç terezese var. Kış könendege kebi, terezelere ikişer kat. Nomer içende ike östel, dürt stul, ike kreslo, bir divan, bir karavat. İden urtasma kanga buyalmış mektüle hatın igılmış. Hatırının başında mıltık pul cerahete kebi cerahet. Hatanın, yese egerme iki, egerme öçten ziyade bulmaska kerek, üze hem sahibi camal. Sudebnıy sledovatel, nomer hucasına karap: "Bu harın ne vakıttan birle seznen nomerefiezde toradır?" -deyu söal kılgaç, nomer hucası: "Ber comga", -deyü cavap virde.

"İsimi niçik?" didikle,

"Zöleyha", -diyü huca efendeden işetelde.

"Bu mektülenen in evvel ültereldekene kem belmeş?" -deyu sledovatel söal kılgaç, huca efende: "Gorniçnaya-hadime belmeş", -didi. "Hadimene bu yire degvet kılınız", -didiklerende, hadimene degvet kılıp kelterdeler. Sudebnıy sledovatel bu tarıyka teftiş kılmağa şorug kıldı:

- Siz bu hatırının ültereldekene ne tarıyka beldeftez?

Hadime bu tarıyka cevapka şorug kıldı: "Ben sabahta hezmet ilen yördekemde bu nomer yanınnan uzıp bardıgımda, bu nomer işekene yartılaş açık kürep, ni bulsa hezmet yukmu iken deyu nomereya kerdekemde, bu hatanı bu keyfiyitte kürep, bik heveflendem. Hetta nomerden ne tarıyka çıkdıgımnı dabelmiymin".

Hadimenen cevabı bu tarıyka temam bulgaç, teftişçeler isek yanma kelep karasalar, işekte yozaknı açkıcı ile kürep, teftişçeler tegacceplenip: "Bu nin-deleen katil iken, üze ketdekte işekne yozak ile biklemiençe kitmeş, hetta işekne yartılaş açık kaldır-mış, eğer de işekne yozak ile biklep kitse idi, bu hatırının ültereldekene belü kiçegep, katilge gizle-nerge de forsat bulır idi, yuksa bu hatın, üz-üzene ültermeşmü iken", -deyu tefekker kıldıklarında, aralarınnan bir teftişçe: "Yuk, üz-üzini katel itmemiş. Eğer de üz-üzini katil itmeş ulsa idi, yanında rivolvir hem üz-üzimni ültiremin deyu hatı bulır idi", -didikinde, başka teftişçiler: "beli, buhr idi", -didiler. Min begıd teftişçiler şviytsarhadimni degvet kıldılar. Slidovatil şviytsardan: "Bu utken kiçte bu mektüle hatınga kim bulsa keldimü?" -deyu söal kıldığında, hadim: "Beli, mektüle ile beraber bir ir adem keldi".

"Ne vakıtta kelip, ne vakıtta ketti?" -deyu söal kıldıklarında şviytsar: "Ahşam sigez segatte kelip, uniki segatte ketti", -deyu cevap virdi. "Uniki segatte kitüvini yakıynen neden belden?" -deyu söal kıldıkta, hadim-şviytsar bu tarıyka cavap virde: "Çönki merhümi ile beraber kelgen adem nomerelerden kiterken: - "Kaç segat?" - deyu benden söal kıldı. Ben, segatke karap: "Unike segat",-deyü cevap birdim. Sonra befia öç ruble na çay virip çıgıp kitti".

Teftişçeler, şveytsardan bu süzlerne işitkeç, mösafire Zöleyhenifi mektüle buluvı sigez segat ile unike segat mabeynende vakıyg uldıgıru anladılar. Teftişçeler, şveytsardan mektüle ile beraber kilgen irnen kıyafetene, tösene: "Ne kıyafetle, ne tösle?" -deyu soradıklarında: "Neçke adem, ozm buylı, ak yözle, kiçik sarı sakallı", -deyu şveytsar kıyafetene, tösene tegrif kıldığı begdende, teftişçeler nomere içendege eşyanı karamağa başladılar. Nomere içermen teftiş kılıp tabılgan nerseler bonlardır: evvelen, bir dane ir perçatkası sul kuldan; sag kul perçatkası yuk. Elbette, bu tabılgan sul kul perçatkası katilnen perçatkası bulırga kirek, döşerep kaldırgandır. Saniyan, iki tereze arasınnan bir ertkalanmış ir portretı - surete tabılmış. Gerçe ertkalanmış ulsa da, suret iczalarını ba tertip cemgı kılıp karalsa, ne şikelle suret idikene anlamak mömkin. Şveytsarraft anladgına hem tegrifene binaen, bu tabılmış suret mektüle Zöleyhe ile utken kiçte beraber kilgen irnen surete bulırga kirek. Salisen, nomere içermen tabılgan bir mektup. Şuşı tabılgan mektup, nomere hucasınıfi rivayatene binaen, mektüle Zöleyhenen kitabete bulırga kirek, çönki nomere hucası mektüle Zöleyhenen resme hatını biledir iken. Bu tabılgan mektüpte ne yazılmış iken deyu tefekkere dalıp belesenez kilse, hikeye-mizdi ziyade gad kılmayınca, tabılgan mektüpni de tercime kılıp yazamız.

Gıyzzetlü ukuçı, mektüle möslime ulsa da, möselmança yazu yazmak bilmiydir idi. Bu tabılmış mektup rus lisanı ile yazılmış idi. Mektüpnin tercümesi budır:

"Efzalış-şebab ulan gıyzzetlü Musa efendilerine can ve dilden selamleremezne kündiremiz. Bu kön ahşam sigez segatte, Gabdullin nomerlerende 11 nçe nomereyi kerem boyırıp kelesez, 2 nçe mart".

Teftişçeler bu mektüpni ukıgaç, bu mektup, elbette, mektüle ile utken kiçte beraber kelgen ademge yazılgandır, didiler. Bu tabılgan mektup mösafire Zöleyhenen mektül bulgan könende yazıldığına mektüptege çislosı daldır. Mektül bulgan kön mektüpnen yazıluvı ile bir köndeder. Teftişçeler mektüle ile beraber kelgen ademnen, isme Musa uldıgını da mektüpten beldeler. Nomerden bu mezkûr öç nerseye başka bir nerse tabılmadı. Teftiş ile, katilnen isme hem kıyafete meglüm uldı. Şimdi teftişçelere katil Musanı izlep tapmak lazim ulıp, teftişne temam idip, teftişçeler nindeleen Musa iken deyu tefekker kılıp, nomerlerden çıgıp kitteler. Nomerler yanma cemgı ulmış cemegat te bir nerse de bele almayınca nomerler yanınnan taralıp kitteler, İkençe könne mektüle Zöleyhene defen kıldılar.

B.8 Kirghiz – Kazan Tatar Translation

Turkish

- Çolpan, günaydın yavrum, kahvaltı hazır.
- Günaydın anne, geliyorum.

Oğulları Erkin'in bugün dersi yoktu.

Gülcan Hanım Çolpan'a seslendi:

- Ağabeyini uyandırma Çolpan.
- Peki anneciğim.

Murat Bey, Gülcan Hanım ve Çolpan sofrada konuşuyorlardı.

- Erkin uyanmadı mı Gülcan?
- Evet Murat, o akşam geç geldi, biraz uyusun.
- Niçin gecikti akşam?
- Maçları varmış, çok yorgun geldi.
- Babacığım bugün anatomi dersim var.
- Ya öyle mi? İyi hazırlandın mı bari?
- Evet babacığım, fakat çok heyecanlıyım.
- Yavrum yumurtanı yemeden gitme.
- Geç kalıyorum anneciğim, bugün yumurta yemesem olmaz mı?
- Peki yavrum, kendine dikkat et.

Kirghiz

- Çolpan, cakşısıŋbı, kızım, nanüştö dayar.
- Cakşısızbı, apa, azır kelem.

Ülu Erkindin bügün sabağı cok.

Gülcan ayım Çolpanğa ün saldı:

- Akeŋdi oyğotpo, Çolpan.
- Makul, apake.

Murat mırza, Gülcan ayım cana Çolpan dastorkon üstündö süylöşüp oturuştu.

- Erkin oyğonbodubu, Gülcan?
- Oba, Murat, al keçinde keç keldi, biraz uktasın.
- Emme üçün keçikti keçinde?
- Futbolu bar eken, ötö çarçap keldi.
- Atake, bügün anatomiyadan sinöm bar.
- A, oşondoybu? Cakşı dayardandıŋbı?
- Oba, atake, birok ötö tolkundanıp catamın.
- Kızım, cumurtkanı cemeyinçe ketpe.
- Keçirip catamın, apake, bügün cumurtka cebesem bolobu?
- Meyli kızım, özüŋö sak bol.

Kazan Tatar

- Gülcan, heyirli irte, kızım! Irtengi aş ezir.
- Heyirli irte, enkey, hezir kileçekmin.

Ulları Erkinnin bugin derisleri yuk.

Gülcan hanım Çolpanğa eytti:

- Abıyınnı uyatma, Çolpan!
- Yarıy, enkey.

Murat bey, Gülcan hanım hem Çolpan östel artınta söyleşeler.

- Erkin uyanmadımı Gülcan?
- Eyi, Murat, ul kiçe son kayttı, biraz yoklasın.
- Nige kiçe sonğa kaldı iken?
- Matöları (uyınnarı) bulğan, bik arıp kayttı.
- Eti! Bugin anatomiyadan imtihanım bar.
- E, şulaymı? Yahşı hezirlendinmi son?
- Eyi eti, fekat (emma) bik nık dulkınlanamın.
- Kızım, yomırka aşamıyça kitme, yarıymı?
- Sonğa kalamın, enkey, bügin yomırka aşamasam bulmasmı?
- Yahşı, kızım, uramda sak bul!