June, 1 – 4, 2011,
KUSADASI, AYDIN, TURKEY

PROCEEDINGS ABSTRACTS

Gediz University Publications
www.gediz.edu.tr
A Comprehensive Two Level Description of Turkmen Morphology

Mehmet Kara, Istanbul University, Contemporary Turkic Languages and Literatures
Dept, Istanbul, Turkey, mehkara@yahoo.com, Maxim Shilyov, Fatih University, Computer Eng
Dept, Istanbul, Turkey, mshilyov@gmail.com, Atakan Kurt, Fatih University, Computer Eng
Dept, Istanbul, Turkey, akurt@fatih.edu.tr

Keywords: Turkmen, phonology, orthography, two level morphology, machine translation, applied linguistics

INTRODUCTION

In this paper we present a two level description of Turkmen Language. Turkmen is a Turkic language and the official language of Turkmenistan. It is spoken by more than 6 million people mostly in Central Asia. We describe the Turkmen orthography using two level rules of Koskenniemi. These orthographic rules governing the phonology of the language during word formation is essential to morphological parsing and generation. We then represent the Turkmen morphotactics using finite state machines. Turkmen like Turkish is an agglutivative language with a rich set of inflectional and derivational morphemes. Words are formed by affixing these morphemes to the root words successively. The FSMs for nominal, verbal and adverbial morphotactics describe in detail how the words of the language can be formed. The orthographic rules and morphotactics are implemented in the Dilmac Machine Translation Framework by encoding them in XML files. We have created a lexicon of root words in Turkmen to test the morphological parsing. We present a number of nominal, verbal and adverbial word formation examples to demonstrate the systems.

The Turkmen language is the official language of Turkmenistan. The Turkmen language is one of the Turkic languages, belonging to the Oghuz group. In the past Arabic script, Unified Turkish Latin Alphabet (UTLA), and Cyrillic script was used, until 1995, the "Tāže Elipbiyi" or New Alphabet was formally introduced and officially came into use in 1996.

Like the rest of the Turkic languages, Turkmen is agglutinative, meaning that most grammatical functions are pointed out by attaching suffixes to the stems of words. One of the most notable features of the Turkmen language is the vowel harmony. All vowels can be classified as front vowels or back vowels. In the Turkmen
language, if there is a back vowel in the first syllable of the word, back vowels are also used in the following syllables. The same can be said for the front vowels.

In this study we present a description of the Turkmen morphology in the two level morphological model of Koskenniemi. Our purpose is to describe the Turkmen phonology and morphology in a formal and precise manner in order to develop a machine translation system between Turkmen and Turkish. Since both languages belong to the same language family and have similar morphology and grammar, a morphological machine translation is possible between these languages.

RELATED WORK

Machine translation is one of the major problems of natural language processing. Machine translation between related languages is feasible because of the similarities and commonalities in morphology, syntax and lexicon. Machine translation between Czech to Slovak, Czech to Polish, and Spanish to Catalan have recently been developed. These projects are practical proof that successful translation between related languages can be built with not too much effort. In a similar afford machine translation between Turkic languages can be possible. They obey the same morphologic, syntactic structures to a great degree. Turkish morphology and syntax from computational perspective is studied in depth by Oflazer and other scholars making it possible to build morphological machine translation systems. Initial work on Azerbaijani, Uygar, Crimean Tatar languages are produced in different studies.

METHODS

Like other Turkic Languages Turkmen is also agglutinative and employs vowel harmony. It has a rich set of inflectional and derivational suffixes and is able to generate huge number of words. A word is formed by affixing inflectional and derivational morphemes to roots or stems successively as described by the Turkmen morphotactics. During the formation of words phonological changes frequently occur. These changes are governed by the phonological or orthographic rules of Turkmen. These two level orthographic rules define how and when many phonological events such as vowel harmony occur. The surface level word corresponds to the written form of a word, after affixing morphemes to the root at the lexical level. We present a comprehensive list two level rules covering most of the Turkmen phonology. An example is given below:

\[ k: g \delta_ - \rightarrow 0 (@:0)V \]

which states that the last k of a word becomes g whenever a morpheme starting with a vowel is affixed to it. An application of this rule would be

Lexical: kirjimek+nH \quad \text{dirty} = \text{Acc}

Surface: kirjimego0i \quad \text{kirjimegi}

Then we move onto describing Turkmen morphology using finite state machines (FSM). A finite state machine, which in principal is a directed graph, consists of a set of states and a set of transitions among these states as shown in the figure below. Transitions such as +IAr, +Hm, +Cl are the edges of graph labeled with inflectional
or derivational morphemes defining in what order those morphemes can be affixed to a word. The immediate states such as [Plural], in a way, represent partial worlds and their part of speech tagging. The initial states such as [Noun] in the figure represent the roots words from a lexicon and their part of speech such as noun, verb, adverb, adjective, etc. The final states such as [1 Prs Possessive] represent full words created by starting with a root word in an initial state and affixing morphemes on the transitions to the partial words in each intermediate state. We define the nominal, verbal and adverbial morphotactics of the language using this FSM model.

**FINDINGS & CONCLUSION**

Finally we provide a number of examples from Turkmen demonstrating the expressive power of two level rules and finite state machines. We have implemented almost all of the Turkmen morphology in Dilmac Machine Translation Framework. We have formulated 41 orthographic rules and covered about 250 morphemes in morphotactics which is done by encoding the morphology in a specific XML format. This is to our knowledge the most comprehensive study on Turkmen language which is the first precondition for a successful machine translation between Turkmen and Turkish. We provided parsing examples using Dilmac. Dilmac software framework primarily developed for morphological machine translation between Turkic languages. Our ultimately goal in Dilmac project is to translate most Turkic languages to Turkish and vice versa.

**REFERENCES**

12. Tanış A. C., Adalı, E. & Ofazer K., A MT System from Turkmen to Turkish Employing Finite State and Statistical Methods, Proceedings of MT Summit XI, 2007