Two-level Morphological Analysis of Ottoman Texts

Ali Burak ERDOGAN
Comp. Eng. Dept.
Hacettepe University
Ankara, Turkey
erdoganaliburak@gmail.com

***Abstract*— This study aims to perform two-level morphological analysis of Ottoman texts written with Ottoman script which is a mixture of Persian and Arabic alphabet. This study may help efforts of transliteration of Ottoman texts with Arabic script to Latin scripts because of the ambiguities occurring. We collected a lexicon of Ottoman words by gathering some Ottoman dictionaries from and defined Turkish verb and noun suffixes to build our morphological analyzer.**

Keywords—morphological analysis, two-level morphological analysis, Ottoman text

I. Introduction

Although there are many studies in the NLP field regarding the modern Turkish language, there are still few ones regarding Ottoman Turkish which has been used for many centuries in Anatolia until 1928. The earlier texts written in Ottoman script need to be understood by researchers, students, readers and historians. This study aims to perform morphological analysis of words in Ottoman script.

Since Turkish is an agglutinative language, the most essential part of Turkish morphology is the analysis of suffixes. Sometimes a conjugated verb in Turkish gives a meaning which can be expressed with many words in another language. (e.g. *al-abil-ir-dik = we could have bought*) Turkish suffixes are not static, they are conjugated and changed depending on context. For example, the first letter of the same suffix may change with respect to the last letter of the root word. These kinds of changes bring more complexity to programmatic morphological analysis and synthesis of words. However, this procedure becomes easier in Ottoman script because the suffixes remain the same in the text and they are not transformed as they do in Latin script, as demonstrated in Table I.

Latin	uzun-**dur**	araba-**dır**	kitap-**tır**
Ottoman	اوزون - **در**	آرابه - **در**	كتاب - **در**

Table I. Comparison of some suffix conjugations in Latin script and Ottoman script

II. Related Work

Although natural language processing in Turkish is an active area in recent years, Ottoman texts have been rarely studied because of some difficulties in reading and writing by researchers. Although the number of studies is small, there are remarkable efforts of shedding a light on valuable textual resources of Ottoman history and literature. Information retrieval from Ottoman text archive images are studied in [1,2]. These studies aimed at converting text-images to text files in Ottoman script. There are also some projects performing Ottoman-script to Latin-script transcription and also providing useful dictionary information [7, 8]. The method of those both projects are following a brute-force approach. They prepared a large list of suffix conjugations which tries to cover the majority of possible suffix combinations in Turkish. The list (as seen in Figure 1) includes 98,161 different suffix combinations (without root words) and 96,570 conjugated words (mostly verbs).

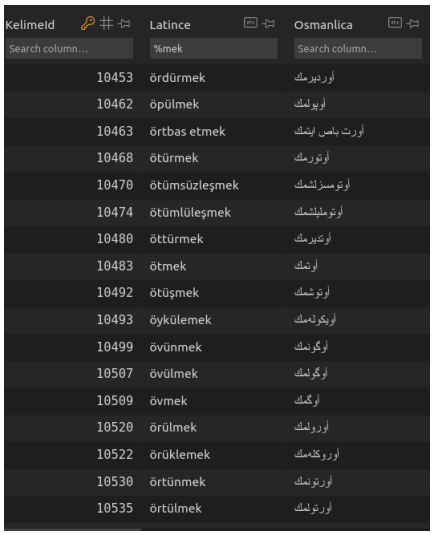
Figure 1. A screenshot from Tevakku Ottoman Dictionary's list.

Converting Ottoman-script to Latin-script is also studied by Güngör et al. [5] and Korkut [6]. Both studies preferred detecting the root word by means of a large Ottoman lexicon and converting the suffixes using Ottoman-Latin suffix pairs instead of using a two-level morphological analyzer.

III. METHODS

A. Lexicon Collection

We made use of numerous Ottoman-Turkish dictionaries from online sources [7, 8] in order to collect Ottoman transcriptions of Turkish words and build a lexicon. Hayrat Nesriyat's dictionary holds nearly 110,000 different words and their translations to Ottoman script, as seen in Figure 2. Instead of using Tevakku's conjugated verb list, we made use of root verbs and nouns from [7].



KelimeId	Latince	Osmanlica
	%mek	
10453	ördürmek	اوردیرمك
10462	öpülmek	اوپولمك
10463	örtbas etmek	اورت باص ایتمك
10468	ötürmek	اوتورمك
10470	ötümsüzleşmek	اوتومسزلشمك
10474	ötümlüleşmek	اوتوملیلشمك
10480	öttürmek	اوتتیرمك
10483	ötmek	اوتمك
10492	ötüşmek	اوتوشمك
10493	öykülemek	اویکولمك
10499	övünmek	اوگونمك
10507	övülmek	اوگولمك
10509	övmek	اوگمك
10520	örülmek	اورولمك
10522	örüklemek	اوروکلمك
10530	örtünmek	اورتونمك
10535	örtülmek	اورتولمك

Figure 2. A sample from Hayrat Nesriyat's Ottoman dictionary [7]

B. Morphological Analysis Method

Morphological analysis of Ottoman text may be performed with two methods:

1. Detecting the possible root words by using a dictionary and trying to extract remaining suffixes by comparing syllables to a suffix list.
2. Building a finite state machine with a lexicon composed of verbs, nouns and suffixes for two-level morphological recognition and generation.

We did not prefer the first method because it does not recognize the morphotactics of Turkish language in a structured way. Also, it does not allow morphological generation, that is picking a root word and defining some suffixes and synthesizing a surface level word.

We followed the second approach and therefore we used *foma* [10], a well-known and common tool for finite-state machine compiling as it is mostly suitable for our requirements. We inserted root verbs and nouns to a lexicon file and described inflection rules by defining a set of verb and noun suffixes. A sample rule set of noun inflections can be seen in Figure 3.

For describing verb and noun inflection rules in foma environment, we are inspired by Coltekin's study

```
LEXICON NPluralSingular

+N:0   NPossesive;
+N+Pl:^لر  NPossesive;


LEXICON NPossesive

#;
+P1S:^م NCase; !(I)m
+P2S:^ڭ NCase; !(I)n
+P3S:^ى NCase;  !(s)I
+P1P:^مز NCase; !(I)mIz
+P2P:^كز NCase; !(I)nIz
+P3P:^لرى NCase; !lArI
```

```
LEXICON NCase
#;
+Acc:^ى #;  !(y)I
+Dat:^ه #;  !(y)A
+Abl:^دن #;   !DAn
+Loc:^ده #;   !DA
+Gen:^ڭ #;   !(n)In
+Inst:^له #;  !(y)lA
```

Figure 3. A sample from Ottoman noun inflection rules in *foma* environment

C. Transformations on suffixes

In Turkish, inflection of a word is not always straightforward. Same suffix may take different forms when being appended to different words as shown in Table I. Therefore we needed to define these kinds of suffix transformations in foma environment by means of functions.

For example, "+P3S" the possessive 3rd singular person suffix (-I) is preceded with the letter "s" if the root word's last letter is a vowel. (kitab-**ı**, araba-**sı**). Same transformation is also applied in Ottoman script by prepending the letter "س" to the suffix "-ى".

Another case is the necessity of inserting a vowel (-I) before possessive suffixes (-m, -n, -mIz, -nIz) when the root word ends with a consonant. This happens in Ottoman script by prepending the letter "-ي" before +P1S, +P1P, +P2P possessive suffixes. We describe these transformations in the "*PossesiveKaynastirmaInsertion*" function in foma as described below:

```
define Vowels [ و | ي | ى | ه | ا ];

define Consonants [ ح | چ | ج | ث | ت | پ | ب
ط | ض | ص | ش | س | ژ | ز | ر | ذ | د | خ |
ن | م | ل | ك | گ | ق | ف | غ | ع | ظ |
];
```

```
! (I)m, (I)n, (I)mIz, (I)nIz, (s)I insertion

define PossesiveKaynastirmaInsertion

    [..] -> ي || NonVowels  "^" _ [ ك ز | م | م ز ] ,,

     [..] -> س || Vowels  "^" _ [ ى ];
```

After *-(s)I* and *-lArI*, the suffixes of *-DA* and *-DAn* become *-nDA* and *-nDAn*, respectively, and *-(y)A* and *-(y)I* become *-nA* and *-nI*, respectively. These insertions are performed in Ottoman script by inserting the letter "-" as described in the "*NInsertion*" function below:

```
define NInsertion

    [..] -> [| ه | د ن ] _ "^" [ ى | ل ر ى ] "^" || [ن
ى | ه ] ,,

     [..] -> ن || Vowels "^" _ ك ; ! (n)In
```

If the root word ends with a vowel, the letter "y" ("-ي" in Ottoman) is inserted in suffixes such as: *-(y)I, -(y)A, -(y)lA*. This rule is described in "*YInsertion*" function below:

```
define YInsertion

    [..] -> ي || Vowels "^" _ [ ه ل | ه | ى  ] ;
```

IV. Experiments

At this stage of our project, we only implemented and experimented inflections of noun words in Ottoman. In order to clearly demonstrate consonant and vowel insertion scenarios of noun inflections we selected two example words:

1. چانطه - Çanta (ends with vowel)
2. كتاب - Kitab (ends with consonant)

All remaining results could be found in Appendix.

A. Surface to Lexical

We performed surface to lexical experiments in order to evaluate the accuracy of FSM. Since inflectional forms of nouns might sometimes cause ambiguities, our FSM should be able to detect possible ambiguities and list all probable parsings according to the rules of Turkish. Figure 4 depicts such a scenario with our two sample words:

```
Done!
2.1 kB. 40 states, 54 arcs, 431 paths.
redefined Lexicon: 2.1 kB. 40 states, 54 arcs, 431 paths.
redefined Grammar: 3.3 kB. 58 states, 93 arcs, 431 paths.
3.3 kB. 58 states, 93 arcs, 431 paths.
foma[3]: up
apply up> چانطەلرىندە
چانطە+N+P3P+Loc
چانطە+N+Pl+P3S+Loc
apply up> كتابلرىندن
كتاب+N+P3P+Abl
كتاب+N+Pl+P3S+Abl
```

Figure 4. Listing noun inflectional ambiguities using FSM in foma environment

چانطەلرنده (Çantalarında) might express two different meanings:

1. Çanta+N+P3P+Loc (at their bag)
2. Çanta+N+Pl+P3S+Loc (at his/her bags)

B. Lexical to Surface Level

Another experiment is giving the lexical description of an inflection and checking the surface level result. As seen in Figure 5, the inflection of word *"كتاب - Kitab"* is given accurately in two scenarios:

1. كتاب +N+Pl+P2P+Gen: Plural noun, Possessive 2nd plural person and genitive inflection is translated to surface level as كتابلريكزڭ (Kitablarınızın - … of your books) accurately.

2. كتاب+N+P1P+Inst:Singular noun, Possessive 1st singular person, instrumental inflection is translated to surface level as كتابيمزله (Kitabımızla = with our book) accurately.

```
apply down> كتاب+N+Pl+P2P+Gen
كڭزكڭيرلرىڭ
apply down> كتاب+N+P1P+Inst
كتابيمزلە
```

Figure 5. Morphological generation with word "كتاب - Kitab"

V. Evaluation

We listed a brief part of our experiments in the previous section. However there were also some wrong inflections that we observed. For example, چانطه+N+Pl+P3P (Plural noun and Possessive 3rd plural person) is translated as چانطەلرلرى (Çantalarları) although it should be translated as (Çantaları). This happens because the ruleset in our FSM description lacks this exceptional case. These kinds of errors should be observed and corrected in the remaining part of our project.

VI. Conclusion

Although some exceptional cases and some other suffixes are obsolete in our current FSM description, we performed the major parts of noun inflections in Ottoman. We demonstrated some example results of two-level morphological recognition and generation (bottom-up and top-down translation). After this midway report we will focus on verb inflections which is a more difficult procedure because of the complicated inflectional nature of verbs in Turkish.

References

[1] Yalniz, I. Z., Altingovde, I. S., Güdükbay, U., & Ulusoy, Ö. (2009). Ottoman archives explorer. Journal on Computing and Cultural Heritage, 2(3), 1–20. doi:10.1145/1658346.1658348.

[2] Ethem F. Can, Pınar Duygulu, A line-based representation for matching words in historical manuscripts, Pattern Recognition Letters, Volume 32, Issue 8, 2011, Pages 1126-1138, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2011.02.013.

[3] Ahmet Afşın Akın, Mehmet Dündar Akın. Zemberek, an open source NLP framework for Turkic languages. https://github.com/ahmetaa/zemberek-nlp Accessed on 03.12.2021.

[4] Çağrı Çöltekin. A freely available morphological analyzer for Turkish. In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), pages 820–827, 2010.

[5] Ayşegül Güngör and İ. Emre Şahin. Dervaze: A spelling dictionary for digital translation of Ottoman documents. International Journal of Languages' Education and Teaching, 5(3):78–84, 2017. doi: 10.18298/ijlet.1824.

[6] Joomy Korkut. Morphology and Lexicon-based Machine Translation of Ottoman Turkish to Modern Turkish. Draft, May 2019. https://www.cs.princeton.edu/~ckorkut/papers/ottoman.pdf Accessed on 03.12.2021.

[7] Hayrat Neşriyat. "Osmanlıca İmla Kılavuzu" Android application. https://play.google.com/store/apps/details?id=com.hayrat.imlakilavuzu Accessed on 03.12.2021

[8] Tevakku. "Osmanlıca Sözlük" Android application. https://play.google.com/store/apps/details?id=com.tevakku.ostr Accessed on 03.12.2021

[9] Oflazer, Kemal. (1993). Two-level Description of Turkish Morphology. Literary and Linguistic Computing. 9. 10.1093/llc/9.2.137.

[10] Mans Hulden. 2009. Foma: a finite-state compiler and library. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session (EACL '09). Association for Computational Linguistics, USA, 29–32.