GTU Department of Computer Engineering
CSE 484/654 - Spring 2023
Homework 1 Report

Emirkan Burak Yılmaz
1901042659

1 Contents

2   Introduction

This homework focuses on utilizing the OpenFST library to implement Turkish morphology rules, specifically the plural and possessive suffixes. The primary objectives are to create Finite State Transducers (FSTs) based on these rules, visualize them, and subsequently test them with various Turkish words to derive their morphological forms. To access more information and the demonstration, locate the **demo.ipynb** file in the notebooks folder.

3   Implementation Details

3.1   The Lexicon

The final state transducer (FST) design is structured around eight distinct word classes: Niv, Nic, Nuv, Nuc, Niiv, Niic, Nuuv, and Nuuc. This design strategically employs knowledge about the last vowel in these classes to handle both plural and possessive suffixes.

In the classification of word classes, 'i' specifically represents the Turkish vowel 'ı', 'ii' represents 'i', 'u' stands for 'u', and 'uu' is indicative of the Turkish vowel 'ü'. The choice of these letters within these classes determines the suffix added to form the possessive structure of a word, which can either precede the word or be embedded within it. For instance, consider the word 'kapı', which ends with the vowel 'ı'. In its second singular possessive form, the suffix also takes the letter 'ı', resulting in 'kapısı'. Another example is the word 'göz', ending with a consonant, yet its last vowel is 'ü'. When forming its first singular possessive form, it changes to 'gözüm'. These examples highlight how the specific vowel used in the possessive suffix directly correlates with the last vowel in the word.

Word Class	Last Vowel	Ends with	Vowel in Possessive Form	Plural Suffix
Niv	{a, ı}	vowel	ı	
Nic	{a, ı}	consonant	ı	
Nuv	{o, u}	vowel	u	+lar
Nuc	{o, u}	consonant	u	
Niiv	{e, i}	vowel	i	
Niic	{e, i}	consonant	i	
Nuuv	{ö, ü}	vowel	ü	+ler
Nuuc	{ö, ü}	consonant	ü	

Following the definition of the word classes, a sample lexicon containing eleven Turkish words is established, and an FST is generated. Each word is linked to its corresponding word class as the final state. Subsequently, it becomes straightforward to append our plural and possessive rules to the existing lexicon FST.

Root Word	Word Class
kapı	Niv
kitap	Nic
piyano	Nuv
kuş	Nuc
kedi	Niiv
ev	Niic
öykü	Nuuv
göz	Nuuc
sarnıç	Nic
kağıt	Nic
toprak	Nic

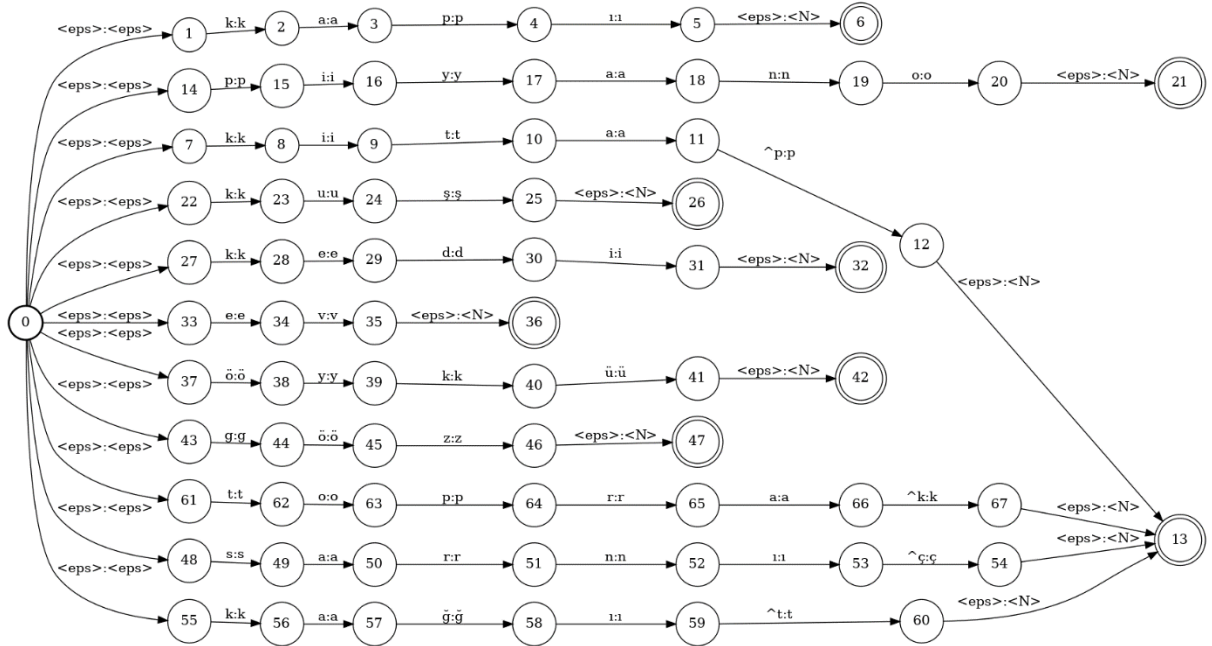| Function	Tag
Noun	<N>
Plural	<pl>
First person singular	<p1s>
Second person singular	<p2s>
Third person singular	<p3s>
First person plural	<p1p>
Second person plural	<p2p>
Third person plural	<p3p>

Figure 1: The Lexicon, Morphological Tags and Lexicon FST

Note that, all word class states have been designated as final states. Additionally, words ending with hard consonants 'p', 'ç', 't', and 'k' are identified with a preceding '^' symbol, indicating the potential alteration in the consonant when it encounters a suffix beginning with a vowel. For example, the consonant 'p' in the word 'kitap' changes to 'b' when attached with the first singular person suffix '-im', resulting in 'kitabım'. This specific rule is referred to as consonant softening.

3.2   Plural Suffix Rule

In Turkish, the plural suffix generally adds "-ler" or "-lar" to the noun base depending on vowel harmony. If the last vowel in the noun is unrounded (a, ı, o, u), you add "-lar" for the plural; if it's rounded (e, i, ö, ü), you use "-ler". Additionally, certain vowels may undergo changes due to vowel harmony rules when the plural suffix is added, ensuring harmony between the suffix and the noun stem.
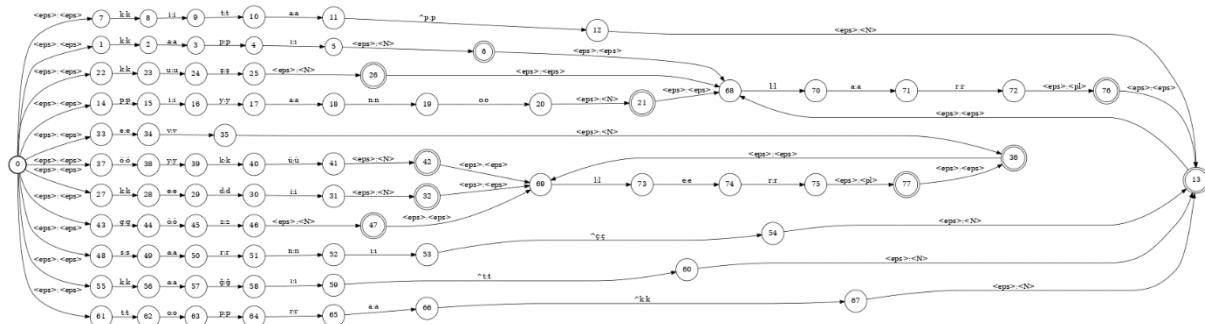

Figure 2: Plural FST

The plural rule operates quite simply: append new arcs as suffixes 'ler' and 'lar' from the word class states, gather them together, and introduce an epsilon transition with the output '<pl>'.

3.3   Possessive Suffix Rule

Possessive suffixes in Turkish add information like "my school" or "his car" to words. They follow Vowel Harmony rules, like the verb "to be". If a word ends in a vowel, "my" gets -m and "your" gets -n **without an extra vowel**. For words ending in certain vowels, the suffix vowel matches the last vowel of the word. Consonant-ending words need a vowel before the suffix. When a word ends in a consonant, "his/her/its" drops the 's' from the suffix. "Their" keeps its suffix as -leri/-ları, regardless of word ending.

The possessive rule relies on both the final vowel and the ending letter, whether it's a vowel or consonant. It involves gathering possessive suffixes based on the word class and adding new arcs from the word class states. The termination point of the suffix node is connected to an epsilon transition to generate the possessive form as output.

```python
def generate_possessive_suffixes(wc: WordClass) -> list:
    plural_suffix = 'lar' if wc in [WordClass.Niv, WordClass.Nic, WordClass.Nuv, WordClass.Nuc] else 'ler'

    vowel = get_suffix_vowel(wc)

    if wc in [WordClass.Niv, WordClass.Nuv, WordClass.Niiv, WordClass.Nuuv]:
        # Vowel ending word, additional consonant 's' for the 3rd singular person
        s1 = 'm'
        s2 = 'n'
        s3 = 's' + vowel
        p1 = 'm' + vowel + 'z'
        p2 = 'n' + vowel + 'z'
        p3 = plural_suffix + vowel
    elif wc in [WordClass.Nic, WordClass.Nuc, WordClass.Niic, WordClass.Nuuc]:
        # Consonant ending word, before suffix there needs to  be additional a vowel
        s1 = vowel + 'm'
        s2 = vowel + 'n'
        s3 = vowel
        p1 = vowel + 'm' + vowel + 'z'
        p2 = vowel + 'n' + vowel + 'z'
        p3 = plural_suffix + vowel

    return [s1, s2, s3, p1, p2, p3]
```
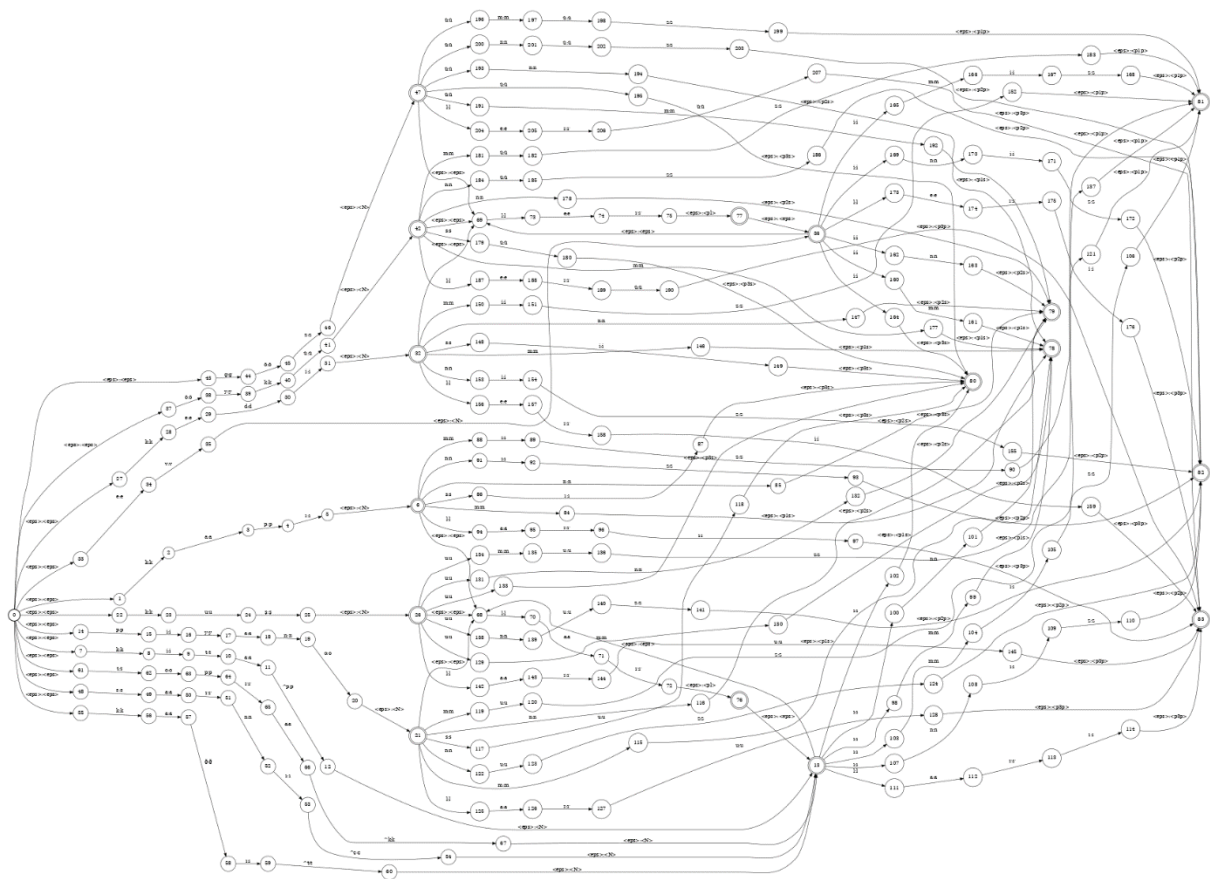
Figure 3: Plural & Possessive FST

4   Testing the Rules

During the testing phase, Turkish possessive suffix rules were validated using ten words. The test dataset comprised valid words conforming to Turkish grammar for potential suffix transformations, alongside invalid words violating Turkish language rules, such as non-Turkish characters, or incorrect suffix combinations. This dataset allowed assessing if the possessive suffix rules produced expected morphological changes.

4.1   Test cases

1. **Plural suffix rule**
2. **Possessive suffix rule**
3. **Combination of plural and possessive suffix rules**
4. **Consonant softening in the root**
5. **İncorrect suffix combinations**
6. **Unrecognized/Unfamiliar Words**

```
test_words = [
    "kapılar",
    "kapıları",
    "kitabım",
    "kitab",
    "kitaplarım",
    "kitablarım",
    "kuşun",
    "kuşumuz",
    "kuşların",
    "kedileri",
    "kedim",
    "kedi",
    "cat",
    "evimiz",
    "evden",
    "evinden",
    "öykümüz",
    "öykünüz",
    "göz",
    "gözlerin",
    "gözü",
    "sarnıçlarımız",
    "sarnıcın",
    "kağıtları",
    "kağıdım",
    "toprakları",
    "toprağımız",
    "toprakımız",
    "TOPRAKLARIMIZ",
    "%cscd",
]
```

4.2 Results

The obtained outcomes align with expectations, indicating that the generated Finite State Transducer accurately performs morphological analysis on the defined lexicon based on the specified rules.

```
1.) kapılar:              11.) kedim:               21.) gözü:
kapı<N><pl>               kedi<N><p1s>              göz<N><p3s>

2.) kapıları:             12.) kedi:                22.) sarnıçlarımız:
kapı<N><pl><p3s>          kedi<N>                   sarnıç<N><pl><p1p>
kapı<N><p3p>
                          13.) cat: REJECT          23.) sarnıcın:
3.) kitabım:                                        sarnıç<N><p2s>
kitap<N><p1s>             14.) evimiz:
                          ev<N><p1p>                24.) kağıtları:
4.) kitab: REJECT                                   kağıt<N><pl><p3s>
                                                    kağıt<N><p3p>
5.) kitaplarım:           15.) evden: REJECT
kitap<N><pl><p1s>                                   25.) kağıdım:
                          16.) evinden: REJECT      kağıt<N><p1s>
6.) kitablarım:
kitap<N><pl><p1s>         17.) öykümüz:             26.) toprakları:
                          öykü<N><p1p>              toprak<N><p3p>
7.) kuşun:                                          toprak<N><pl><p3s>
kuş<N><p2s>               18.) öykünüz:
                          öykü<N><p2p>              27.) toprağımız:
8.) kuşumuz:                                        toprak<N><p1p>
kuş<N><p1p>               19.) göz:
                          göz<N>                    28.) toprakımız: REJECT
9.) kuşların:
kuş<N><pl><p2s>           20.) gözlerin:            29.) TOPRAKLARIMIZ:
                          göz<N><pl><p2s>           toprak<N><pl><p1p>
10.) kedileri:
kedi<N><pl><p3s>                                    30.) %cscd: REJECT
kedi<N><p3p>
```

- 13, unknown word, rejected
- 15, 16, the suffix 'den' is not included in the defined language
- 28, the word does not obey the consonant softening rule. Correct spelling is 'toprağımız'
- 30, unknown word, rejected

5   Conclusion

This homework explored Turkish morphology using OpenFST by implementing rules for plural and possessive suffixes. The report outlines the construction of finite state transducers, their visualization, and testing on Turkish words, showcasing the derivations.