

## A THREE-LAYERED MORPHO-PHONOLOGICAL ANALYZER FOR TURKISH

Yılmaz Kiliçaslan, Önder Açıkgöz, Özlem Aydın

Trakya University, Department of Computer Engineering, 22030, Edirne, Turkey

### Abstract

*This paper presents a three-layered morpho-phonological analyzer for Turkish. These layers correspond to the phonetic, phonological and morphological levels of the language. The layered approach is based on the recognition of the autonomy of the levels of a language. Three types of automata are designed and implemented to carry out the autonomous operations: an automaton without initial and final states for analyzing phonetic dependencies, an acyclic automaton for performing phonological operations and a cyclic automaton for morphological operations. Each automaton is responsible for combining linguistic units relevant to the level implemented into larger fragments. However, in addition to this axis of combination the analyzer also implements an axis of abstraction. Notice that the linguistic levels are argued to be autonomous not entirely independent. The vertical interaction among the levels is carried out by an automaton providing the one higher up with properly abstracted information.*

**Key words:** *Phonetics, Phonology, Morphology, Computational Linguistics, Turkish.*

### 1. INTRODUCTION

The aim of this paper is to present a computational model of Turkish morpho-phonology. The model comprises three levels of language: phonetics, phonology and morphology. Each level is modelled by an automaton of a certain kind.

The linguistic levels to be modelled are analyzed in Section 2. The discussion given in this section centers around the idea that each level is autonomous on a dimension of combination and they are altogether ordered on a dimension of abstraction. Section 3 presents the automata constructed to computationally model the relevant fragment of Turkish grammar from this two-dimensional perspective. Section 4 is a brief conclusion of the paper.

### 2. LINGUISTIC ANALYSIS

The process of understanding of a sentence takes place at two dimensions: the dimension of combination and the dimension of abstraction. At the former, smaller linguistic units are combined together in order to form larger units. At the latter dimension, which is usually ignored, linguistic constructions become more abstract as they get bigger. A holistic analysis of language from this two-dimensional perspective goes beyond the scope of this study. The work presented in this paper is confined to the following layers: phonetics, phonology and morphology.

#### 2.1 Phonetics: Language in nature

The least abstract (most concrete) level of language is that of phonetics. At this level, we hear physical sounds, sounds as occurring in nature.

Phonetic sound is studied with respect to its physical properties concerning its articulation (by the speaker), its transmission through air (from the speaker to the hearer) and its reception and perception (by the hearer).

It is not uncommon to confuse the field of phonetics with that of phonology. However, as Lass (1984) point out in the following quote, a clear boundary can be drawn between the two fields:

Phonology, broadly speaking, is that subdiscipline within linguistics concerned with ‘the sounds of language’. More narrowly, phonology proper is concerned with the function, behaviour, and organization of sounds as LINGUISTIC items; as opposed to phonetics, which is a rather more ‘neutral’ study of the sounds themselves as phenomena in the physical world, and the physiological, anatomical, neurological, and psychological properties of the human beings that make them. Phonology, that is, is ‘linguistic’ in the sense that syntax, morphology, and to a large extent, semantics are; while phonetics shades off at various points into neurophysiology, perceptual psychology, acoustics, and so on. (pp. 1)

To paraphrase Lass, phonological sound serve a linguistic function such as encoding meaning whereas phonetic sound has physical properties irrelevant to linguistic meaning. To give an example, consider the English words *kit* and *skill*. Both words contain the same phoneme /k/. Nonetheless, these two sounds are not identical. In *kit*, the sound is aspirated, whereas in *skill*, it is unaspirated. Despite the difference in spelling, English speakers will tend to think these two sounds are the same. Thus, these two sounds should be considered two allophones of the same phoneme in English.

Which allophone to use depends on the context of utterance. Rules of the following form serve to indicate the context in which an allophone appears:

$$A \rightarrow B / D \_ E.$$

For example, the fact that voiceless stops (e.g. /k/) become aspirated in English when they occur syllable-initially can be captured by the following rule:

$$[\text{voiceless stop}] \rightarrow [\text{aspirated}] / \sigma \_ .$$

It is noteworthy that phonetic alternations such as aspiration are not orthographically represented in ordinary texts. However, some alternations surface in text as different letters. Turkish is particularly rich in this respect. The use of vowels and consonants in Turkish is susceptible to some constraints of morphophonemic harmony. The vowel harmony and consonant harmony are two such morphophonemic processes.

The vowel harmony in Turkish is based on two features:  $\pm$ front and  $\pm$ rounded. In general words, the first vowel of a morpheme has to agree with the last vowel of the preceding morpheme. We can exemplify such harmony with the plural morpheme in Turkish. This morpheme has two allomorphs: *-ler* and *-lar*. The former can follow a +front vowel whereas the latter appears after a -front vowel (e.g. *köpek-ler* (dog-plur) vs. *kitap-lar* (book-plur)).

The consonant harmony in Turkish is more intricate than the vowel harmony. We will not go into these intricacies.<sup>1</sup> Practically speaking, the consonant harmony requires the Turkish consonants to be divided into three groups: (1) {ç, f, h, k, p, s, ş, t} (2) {l, m, n, r, y} and (3) {b, c, d, g, ğ, j, v, z}. A consonant in group (2) can be followed and preceded by any consonant in groups (1) and (3). Furthermore, a consonant can be followed and preceded by any consonant in its group.

## 2.2 Phonology: Sounds of language in cognition

Phonology can be defined as the discipline concerning the qualitative aspects of the sounds of language. That is, rather than physical properties of sounds, the organizations of sounds that signal changes in meaning fall within the scope of phonology.

The most basic units in phonology are phonemes. A phoneme is a group of different sounds perceived to have the same function by speakers of a language. That is, phonemes are discrete and cognitive units which do not necessarily carry the exact same physical properties. They are not only the outputs of a process of conversion from continuous waves of sounds in the physical world but also abstractions from a set of so-called phones, which are physical speech units perceived by the human ear. To consider two different phones to as two distinct phonemes, they need to be employable to form meaningful contrasts between utterances: for example, there is a meaning change between *call* and *tall* as /t/ and /c/ are different phonemes in English.

Phonemes are represented by sequences of letters (commonly placed between slashes). In rather transparent languages like Turkish, more or less each phoneme is represented by a single letter. All languages possess a finite set of phonemes. However, researchers are not in agreement about the exact number of phonemes that appear across the world's languages. Bernstein et al (2011) state that this number varies from a low of thirteen (Hawaiian) to a high of over sixty (Hindi).<sup>2</sup> According to Başkan (1967), Turkish has 37 phonemes, 9 of them being vowels, 23 consonants and 5 prosodic features.<sup>3</sup> Demircan (1977) argues this number to be 38.<sup>4</sup> What is

<sup>1</sup> The interested reader is referred to Clements and Sezer (1982) and Özsoy (2004).

<sup>2</sup> Bernstein, D, Louis AP, Alison CS and Edward, JR 2011, 'Psychology', Cengage Learning, 9th edition, pp. 315.

<sup>3</sup> Başkan, Ö 1967, 'Lengüistik Metodu', İstanbul: Multilingual.

computationally important is that the human cognitive apparatus need not have the capability of productively generating the phonemes. As they constitute a finite set in any case, we may reasonably assume that there is one simple mathematical operation concerning phonemes: identification of a phoneme. Further, phonemes are concatenated to build up morphemes.

### 2.3 Morphology: Linear ordering of morphemes

We have seen above that a phoneme is the smallest (abstract) unit of the sound system of a language that makes a difference in the meaning of its words. The next bigger unit of a language is a *morpheme*. That is the smallest unit of the language that is meaningful by itself. The subfield of linguistics dedicated to the study of morphemes is morphology.

Moreover, just like phonemes have allophones, morphemes have allomorphs. More precisely, just as phonemes are abstractions over allophones, morphemes are abstractions over allomorphs. An allomorph of a morpheme is semantically identical to it but differs in pronunciation. English has allomorphs for the plural. The plural marker of a regular English noun can be pronounced /-z/, /-s/, or /-iz/, depending on the final sound of the noun's singular form. A morpheme can be thought of as a set of sequences of phonemes, i.e. a set of allomorphs.

Turkish is a suffixal language. That is, morphemes are right-attached to the words. The suffixes in Turkish can be divided into two paradigms: a noun paradigm and a verb paradigm. Underhill (1986)<sup>5</sup> describes these paradigms as a finite list of morpheme slots. The description of the verb paradigm and noun paradigm are illustrated, respectively, in **Table 1** and **Table 2**:

#### **MORPHEMES OF THE VERB PARADIGM (IN TURKISH):**

1. **VERB STEM.**
2. **DERIVATION:** reflexive *-In*, reciprocal *-İş*, causative *-Dir*, passive *-Il*.
3. **TENSE:** aorist *-Ir*, progressive *-Iyor*, definite past *-DI*, narrative past *-mİş*, future *-(y)EcEk*, optative *-(y)E*, necessitative *-mElI*, conditional *-sE*  
(Notice that not all of these are tenses in the strict sense of the term.)
4. **AUXILIARY:** past *-(y)mİş*, conditional *-(y)sE*, adverbial *-(y)ken*.
5. **PERSON**

**Table 1.**

#### **MORPHEMES OF THE NOUN PARADIGM (IN TURKISH):**

1. **NOUN STEM.**
2. **PLURAL:** *-lEr*.
3. **POSSESSIVE.**
4. **CASE:** objective/accusative *-(y)I*, genitive *-(n)In*, dative *-(y)E*, locative *-DE*,  
a. ablative *-DEn*, instrumental/comitative *-(y)lE*.
5. **RELATIVE:** *-ki*. (This suffix is added only to genitive or locative suffixes.)

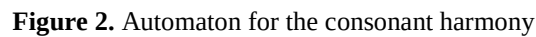
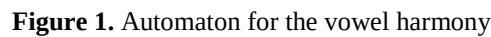
**Table 2.**

<sup>4</sup> Demircan, Ö 1977, 'Türkiye Türkçesinde Kök-Ek Bileşmeleri', Ankara.

<sup>5</sup> Underhill, R. 1986. 'Turkish2, In D.I. Slobin and K. Zimmer, (eds.), Turkish, pp. 7-21.

The design is based on the separate analysis of the linguistic levels that are analyzed in the previous section. Thus, the design consists of three distinct modules, each of which is responsible for combining linguistic units into larger fragments relevant to the corresponding level of language. However, it should be kept in mind that there is a vertical interaction between the levels, which has to be taken into consideration in the implementation phase. Each level is modelled by an automaton of an appropriate type.

Of the phonetic processes in Turkish, our analyzer deals with the vowel and consonant harmony, only. Cebiroğlu and Adalı (2002) propose the automata in **Figure 1** and **Figure 2** for modelling the vowel harmony and consonant harmony, respectively:



```

graph LR
    start((start)) --> q0((q0))
    q0 -- k --> q1((q1))
    q1 -- ɪ --> q2((q2))
    q2 -- p --> q3((q3))
    q3 -- t --> q4((q4))
    q4 -- k --> q5(((q5)))
    q5 -- l --> q6((q6))
    q6 -- ɪ --> q7((q7))
    q7 -- r --> q8(((q8)))
    q0 -- noun --> q5
    q5 -- plur --> q8
    style q0 fill:#fff,stroke:#333,stroke-width:1px
    style q1 fill:#fff,stroke:#333,stroke-width:1px
    style q2 fill:#fff,stroke:#333,stroke-width:1px
    style q3 fill:#fff,stroke:#333,stroke-width:1px
    style q4 fill:#fff,stroke:#333,stroke-width:1px
    style q5 fill:#ccc,stroke:#333,stroke-width:1px
    style q6 fill:#fff,stroke:#333,stroke-width:1px
    style q7 fill:#fff,stroke:#333,stroke-width:1px
    style q8 fill:#ccc,stroke:#333,stroke-width:1px
  
```

In the diagram above, the circled vowels are in harmony as they are both +front. It is this harmony that picks out *-ler* rather than *-lar* as the plural allomorph to use in this example.

### 3.2 Allomorph construction with acyclic automata

Phonemes combine together to form allomorphs. This is a linear and upper-bounded process. That is, the phonemes are sequentially ordered and the sequence formed in this way is non-repeatable in a single allomorph. Moreover, as pointed out in the previous section, each letter in the Turkish alphabet represents, more or less, a distinct phoneme. Mathematically speaking, the allomorphs of each Turkish morpheme can be constructed with an acyclic automaton. Below is the automaton for the plural morpheme (**Figure 4**):

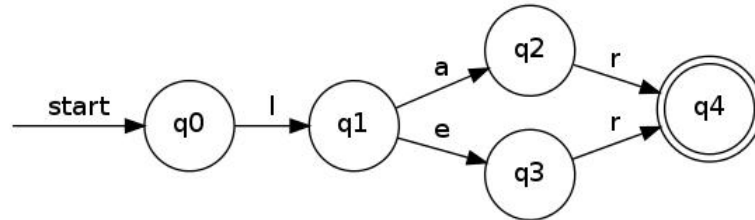


**Figure 4.** An acyclic automaton for the plural morpheme in Turkish

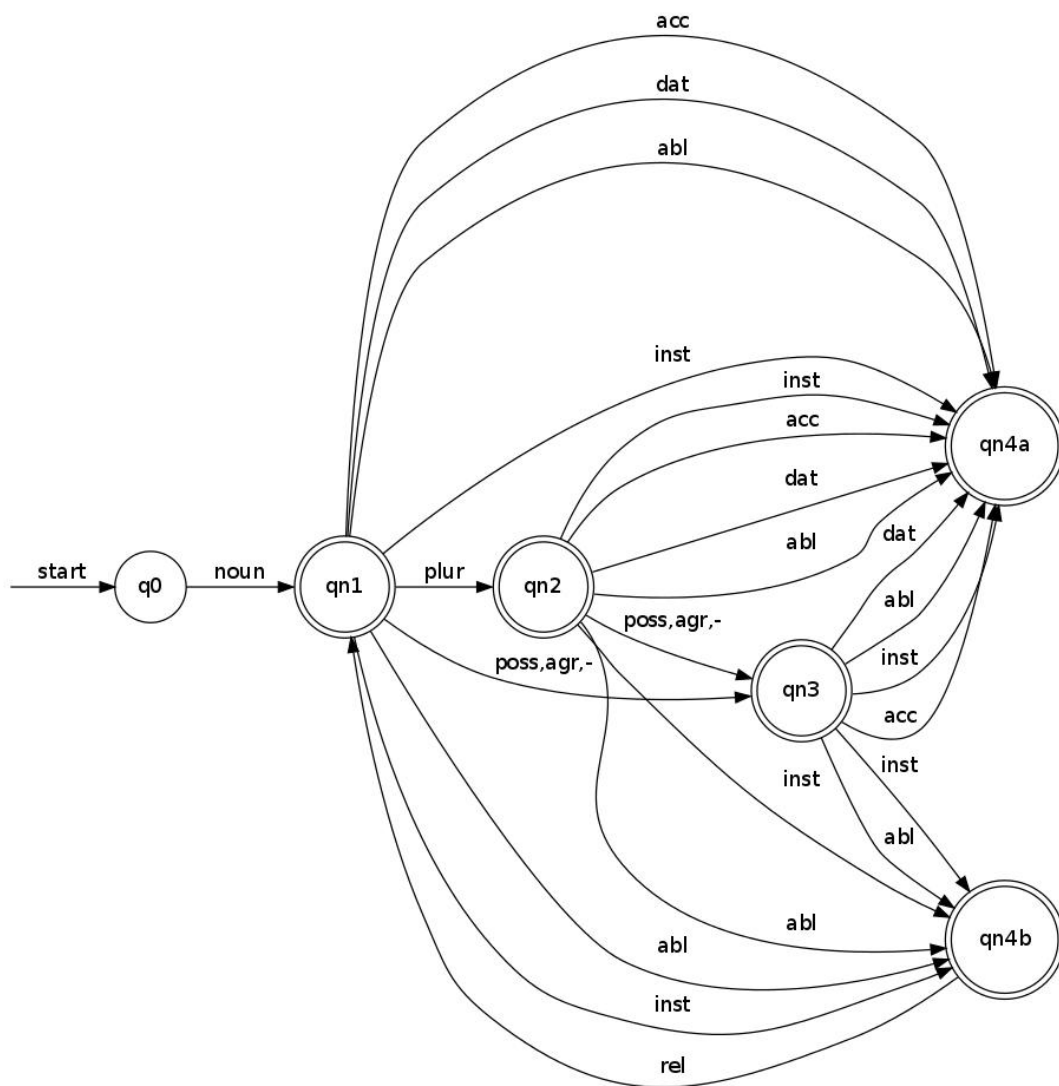
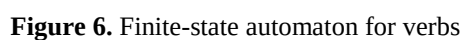


**Figure 5.** Finite-state automaton for nouns



Such acyclic automata are called by the automaton responsible for morphological analysis in order to abstract out a morpheme from the allomorph returned. Let us now turn to that automaton for morphological analysis.

### 3.3. Word formation with a finite-state automaton

Turkish words can be infinitely longer in theory. However, they can still be generated by a finite-state automaton. Below is the finite-state automaton for nouns in Turkish.

The finite-state automaton for verbs in Turkish is as shown in the following diagram.

Turkish has also certain suffixes that turn nouns to verbs and vice versa. The automaton diagrammed below illustrates how to realize such nominalizations and verbalizations.

The automaton above describes all possible word formations in Turkish with inflectional suffixes. The derivational suffixes are left beyond the scope of this study.

## 4. CONCLUSION

The model of grammar presented in this paper is based on two axes: an axis of combination and an axis of abstraction. As the levels of syntax and discourse is beyond the scope of the study, the model is upper-bounded by word structure. Phonemes combine together to form words. However, an orthogonal process of abstraction accompanies the process of combination. Once a series of phonemes have constructed a meaningful unit, an allomorph, the phonological content of this unit is abstracted away to generate a morpheme, an abstract building block for word formation. Similarly, phonemes are abstractions from allophones, physical sounds carrying features irrelevant to linguistic meaning. We have seen that this two-dimensional perspective allows for neatly modelling the phonological and morphological levels of Turkish with relevant sorts of automata, along with some phonetic effects observed on the realization of phonemes.

## REFERENCES

- Adalı, E, Cebiroğlu, G 2002, 'Sözlüksüz köke ulaşma yöntemi', *In Proceedings of the 19th TBD Bilişim Kurultayı*, pp. 155-160.
- Başkan, Ö 1967, 'Lengüistik Metodu', İstanbul: Multilingual.
- Bernstein, D, Louis AP, Alison CS and Edward, JR 2011, 'Psychology', Cengage Learning, 9th edition, pp. 315.
- Clements, GN, Sezer, E 1982, 'Vowel and consonant disharmony in Turkish', *In The structure of phonological representations*, Harry van der Hulst & Norval Smith, eds., Dordrecht: Foris, pp. 213–255.
- Demircan, Ö 1977, 'Türkiye Türkçesinde Kök-Ek Bileşmeleri', Ankara.
- Lass, R 1984, 'Phonology: an introduction to basic concepts', Cambridge University Press.
- Özsoy, S. 2004, 'Türkçenin Yapısı 1 Sesbilim', İstanbul: Boğaziçi Üniversitesi Yayınevi.
- Underhill, R. 1986. 'Turkish', *In D.I. Slobin and K. Zimmer, (eds.), Turkish*, pp. 7-21.