Chapter 2
Morphological Processing for Turkish

Kemal Oflazer

Abstract This chapter presents an overview of Turkish morphology followed by the architecture of a state-of-the-art wide coverage morphological analyzer for Turkish implemented using the Xerox Finite State Tools. It covers the morphophonological and morphographemic phenomena in Turkish such as vowel harmony, the morphotactics of words, and issues that one encounters when processing real text with myriads of phenomena: numbers, foreign words with Turkish inflections, unknown words, and multi-word constructs. The chapter presents ample illustrations of phenomena and provides many examples for sometimes ambiguous morphological interpretations.

2.1 Introduction

Morphological processing is the first step in natural language processing of morphologically complex languages such as Turkish for downstream tasks such as document classification, parsing, machine translation, etc. In this chapter, we start with an overview of representational issues, and review Turkish morphophonology and morphographemics including phenomena such as vowel harmony, consonant assimilation, and their exceptions. We then look at the root word lexicons and morphotactics, and describe *inflectional groups*, first mentioned in Chap. 1 and are quite important in the interface of morphology with syntax. We then provide numerous examples of morphological analyses highlighting morphological ambiguity resulting from root word part-of-speech ambiguity, ambiguity in segmentation of a word into morphemes, and homography of morphemes.

We then briefly discuss the internal architecture of the finite state transducer that has been built using the two-level morphology approach (Koskenniemi 1983; Beesley and Karttunen 2003), that is the underlying machinery that can be

K. Oflazer (✉)
Carnegie Mellon University Qatar, Doha-Education City, Qatar
e-mail: ko@cs.cmu.edu

K. Oflazer, M. Saraçlar (eds.), *Turkish Natural Language Processing*,
Theory and Applications of Natural Language Processing,
https://doi.org/10.1007/978-3-319-90165-7_2

customized to provide many different analysis representations: for example, as surface morphemes, or as lexical morphemes or as a root word followed by morphological feature symbols. We can also generate more complex representations that encode both the phonological structure of a word (phonemes, syllables, and stress position) and its morphological structure as morphological feature symbols.

Subsequently we discuss issues that arise when processing real texts where one encounters many tokens that present different types of complications. Examples of such phenomena are acronyms, numbers written with digits but then inflected, words of foreign origins but inflected according to Turkish phonological rules and unknown words where the root words are not known but some morphological features can be extracted from any suffixes. We do not cover issues that occur in seriously corrupted sources of text such as tweets where vowels and/or consonants are dropped, capitalization and/or special Turkish characters are haphazardly used or ignored, and character codes that do not occur in Turkish are widely used when text is typed through smartphone keyboards from users in various countries especially across Europe. However our morphological analyzer is very robust in handling many cases that one encounters even in such sources.

Finally we conclude with an overview of multi-word processing covering compound verbs, lexicalized collocations, and non-lexicalized collocations.

2.2 Overview of Turkish Morphology

Morphologically Turkish is an agglutinative language with word forms consisting of morphemes concatenated to a root morpheme or to other morphemes, much like "beads on a string" (Sproat 1992). Except for a very few exceptional cases, the surface realizations of the morphemes are conditioned by various regular morphophonological processes such as vowel harmony, consonant assimilation, and elisions. The morphotactics of word forms can be quite complex when multiple derivations are involved as it is quite possible to construct and productively use words which can correspond to a multiple word sentence or phrase in, say, English. For instance, the derived modifier *s*ağlamlaştırdığımızdaki[1] would be represented as:

```
sağlam+Adj^DB
      +Verb+Become^DB
      +Verb+Caus+Pos^DB
      +Noun+PastPart+A3sg+Pnon+Loc^DB
      +Adj+Rel
```

[1]Literally, "(the thing existing) at the time we caused (something) to become strong." Obviously this is not a word that one would use everyday. Turkish words (excluding non-inflecting high-frequency words such as conjunctions, clitics, etc.) found in typical running text average about 10 letters in length. The average number of bound morphemes in such words is about 2.

Starting from an adjectival root *sağlam*, this word form first derives a verbal stem *sağlamlaş*, meaning "to become strong," with the morpheme *+laş*. A second suffix, the causative surface morpheme *+tır* which we treat as a verbal derivation, forms yet another verbal stem meaning "to cause to become strong" or "to make strong." The immediately following participle suffix *+dığ* produces a nominal, which inflects in the normal pattern for nouns (here, for 1st person plural possessor and locative case with suffixes *+ımız* and *+da*). The final suffix, *+ki*, is a relativizer, producing a word that functions as a modifier in a sentence, whose overall semantics was given above modifying a noun somewhere to the right.

The feature form representation above has been generated by a two-level morphological analyzer for Turkish (Oflazer 1994) developed using XRCE finite state tools (Karttunen and Beesley 1992; Karttunen 1993; Karttunen et al. 1996; Beesley and Karttunen 2003). This analyzer first uses a set of morphographemic rules to map from the surface representation to a lexical representation in which the word form is segmented into a series of lexical morphemes. For the word above, this segmented lexical morphographemic representation would be:

```
sağlam+lAş+DHr+DHk+HmHz+DA+ki
```

In this representation, lexical morphemes except the lexical root utilize meta-symbols that stand for a set of graphemes. These graphemes are selected on the surface by a series of morphographemic processes which are originally rooted in the morphophonological processes of the language. We will discuss some of these processes below.

For instance, `A` stands for back and unrounded vowels *a* and *e* in orthography, `H` stands for high vowels *ı*, *i*, *u*, and *ü*, and `D` stands for *d* and *t*, representing alveolar consonants. Thus a lexical morpheme represented as `+DHr` actually represents 8 possible allomorphs, which appear as one of *+dır, +dir, +dur, +dür, +tır, +tir, +tur, +tür* depending on the local morphophonemic/morphographemic context.

Once all segmentations of a word form are produced, they are then mapped to a more symbolic representation where root words are assigned part-of-speech categories from any relevant lexicons, and morphemes are assigned morphosyntactic feature names including default features for covert or zero morphemes, (e.g., if there is no plural morpheme on a noun, then we emit a feature name `+A3sg`, indicating that word is singular.)

A short listing feature names are provided in Appendix.

2.3 Morphophonology and Morphographemics

Overviews of Turkish phonology can be found in Clements and Sezer (1982), van der Hulst and van de Weijer (1991), and Kornfilt (1997). Turkish has an eight vowel inventory which is symmetrical around the axes of backness, roundness, and height: /i, y, e, 2, a, o, 1, u/ which correspond to *i*, *ü*, *e*, *ö*, *a*, *o*, *ı*, and *u* in

Turkish orthography.[2] Suffix vowels typically harmonize in backness, and (if high) in roundness to the preceding stem vowel (compare, e.g., *ev+ler* /evler/ "houses" to *at+lar* /atlar/ "horses"). But there are several suffixes, e.g., the relativizer *+ki*, whose vowels do not harmonize, as well as others, e.g., progressive suffix *+Hyor*, in which the first vowel harmonizes but the second does not. Many roots are internally harmonic but many others are not; these include loan words (e.g., *kitap* /citap/ "book", from Arabic) as well as some native words (e.g., *anne* /anne/ "mother"). Furthermore, vowel harmony does not apply between the two components of (lexicalized) compounds.

Turkish has 26 consonants: /p, t, tS, k, c, b, d, dZ, g, gj, f, s, S, v, w, z, Z, m, n, N, l, 5, r, j, h, G/. However, orthography uses only 21 letters for consonants: /g/ and its palatal counterpart /gj/ are written as *g*, while /k/ and its palatal counterpart /c/ are written as *k*, /5/ and its palatal counterpart /l/ are written as *l*, /v, w/ are written as *v* and /n/ and its nasal counterpart /N/ are written as *n*. Palatalized segments (/gj, c, l/) contrast with their nonpalatalized counterparts only in the vicinity of back vowels (thus *sol* is pronounced /so5/ when used to mean "left" vs. /sol/ when used to mean the musical note G). In the neighborhood of front vowels, palatality is predictable (*lig* /ligj/ "league"). /G/, written as *ğ*, represents the velar fricative or glide corresponding to the historical voiced velar fricative that was lost in Standard Turkish. When it is syllable-final, some speakers pronounce it as a glide (/w/ or /j/) and others just lengthen the preceding vowel. In morphological processing we treat it as a consonant when it is involved in morphologically induced changes.

Root-final plosives (/b, d, g/) typically devoice when they are syllable-final (thus *kitab+a* /ci-ta-ba/ "to the book," but *kitap* /ci-tap/ "book," *kitap+lar* /ci-tap-lar/ "books".[3] Suffix-initial plosives assimilate in voice to the preceding segment (thus *kitap+ta* /ci-tap-ta/ "in the book" but *araba+da* /a-ra-ba-da/ "in the car".

Velar consonants (/g/ and /k/) reduce to /G/ at most root-suffix boundaries; thus *sokak* /sokak/ "street" *sokak+ta* /so-kak-ta/ "on the street" but *so-ka-ğa* /so-ka-Ga/ "to the street." For more details on the phonology of Turkish words including details of syllable structure and stress patterns, we refer the reader to Oflazer and Inkelas (2006).

We now present relatively informally, a reasonably complete list of phonological phenomena that are triggered when morphemes are affixed to root words or stems. These rules can be implemented in many different ways depending on the

[2]For phonological representations we employ the SAMPA representation. The Speech Assessment Methods Phonetic Alphabet (SAMPA) is a computer-readable phonetic script using 7-bit printable ASCII characters, based on the International Phonetic Alphabet (IPA) (see en.wikipedia.org/wiki/Speech_Assessment_Methods_Phonetic_Alphabet (Accessed Sept. 14, 2017) and www.phon.ucl.ac.uk/home/sampa/ (Accessed Sept. 14, 2017)). The Turkish SAMPA encoding convention can be found at www.phon.ucl.ac.uk/home/sampa/turkish.htm (Accessed Sept. 14, 2017).

[3]In this chapter, we use - to denote syllable boundaries and + to denote morpheme boundaries wherever appropriate.

underlying engine for implementing a morphological analyzer.[4] We present our examples through aligned lexical and surface forms in the usual convention of two-level morphology to point out the interactions between phonemes. In the examples below, the first row (L) shows the segmentation of the lexical representation of a word into its constituent lexical morphemes, the second row (S) shows the (aligned) surface form with any morphologically–induced changes highlighted in boldface (where we also use `0` to indicate the empty string resulting from the deletion of a lexical symbol) and the third row indicates the actual orthographical surface form (O) as written in text. At this stage all our representations employ letters in the Turkish alphabet although all these changes are phonologically motivated.[5]

(a) **Vowel Harmony-1**: The lexical vowel `A` (representing a back and rounded vowel) in a morpheme is realized on the surface as an `a` if the last vowel on the surface is one of `a`, `ı`, `o`, `u`, but is realized as an `e` if the last vowel on the surface is one of `e`, `i`, `ö`, `ü`. For example:

L	masa+l**A**r	okul+l**A**r	ev+l**A**r	gül+l**A**r
S	masa**0**l**a**r	okul**0**l**a**r	ev**0**l**e**r	gül**0**l**e**r
O	masalar	okullar	evler	güller

(b) **Vowel Harmony-2**: The lexical vowel `H` (representing a high-vowel) in a morpheme is realized on the surface as an

- `i` if the last vowel on the surface is one of `e, i`,
- `ı` if the last vowel on the surface is one of `a, ı`,
- `u` if the last vowel on the surface is one of `o, u`
- `ü` if the last vowel on the surface is one of `ö, ü`

For example:

L	masa+y**H**	okul**+yH**	ev**+yH**	sürü+y**H**	gül+l**A**r**+yH**
S	masa**0**y**ı**	okul**00u**	ev**00i**	sürü**0**y**ü**	gül**0**l**e**r**00i**
O	masayı	okulu	evi	sürüyü	gülleri

There are a couple of things to note here. Clearly there are other morphographemic processes going in the second, third, and fifth examples: for example, a lexical `y` is (concurrently) deleted on the surface (to be discussed below). The fifth example actually shows three processes happening concurrently: the mutually dependent vowel harmony processes take place along with the `y` in the third morpheme being deleted.

[4]For example, Xerox Finite State Tools, available at www.fsmbook.com (Accessed Sept. 14, 2017), FOMA, available at fomafst.github.io/ (Accessed Sept. 14, 2017), HFST available at hfst.sf.net (Accessed Sept. 14, 2017) or OpenFST available at www.openfst.org (Accessed Sept. 14, 2017).

[5]Note that we also explicitly show the morpheme boundary symbol, as in implementation, it serves as an explicit context marker to constrain where changes occur.

While these vowel harmony rules are the dominant ones, they are violated in quite many cases due to vowel quality being modified (usually) as a result of palatalization. For example:

L	hilal+l**A**r		alkol**+yH**	
S	hilal**0**l**e**r	(not hilal**0**l**a**r)	alkol**00ü**	(not alkol**00u**)
O	hilaller		alkolü	

These cases are for all practical purposes lexicalized, and the internal lexical representations of such cases have to mark them with alternative symbols so as to provide contexts for overriding the default harmony rules.

(c) **Vowel Deletion**: A morpheme initial vowel is deleted on the surface when affixed to a stem ending in a vowel, unless the morpheme is the present progressive morpheme +Hyor in which case the vowel in the stem is deleted. For example,

L	masa**+H**m	ağl**a+H**yor
S	masa**00**m	ağl**00ı**yor
O	masam	ağlıyor

(d) **Consonant Deletion**: Morpheme-initial s, y, and n are deleted when either of the accusative morpheme +yH or the possessive morpheme +sH or the genitive case morpheme +nHn is attached to a stem ending in a consonant. For example:

L	kent**+sH**	kent**+yH**	kent**+nH**n
S	kent**00i**	kent**00i**	kent**00i**n
O	kenti	kenti	kentin

Note that this can also be seen as insertion of a y, s or an n on the surface when the stem ends in a vowel. As long as one is consistent, this ends up being a representational issue which has no bearing on the computational implementation.

(e) **Consonant Voicing**: A morpheme initial dental consonant (denoted by D representing d or t) will surface as a voiced d, when affixed to a stem ending in a surface vowel or the consonants other than h, ş, ç, k, p, t, f, s. For example:

L	kalem**+DA**	kale**+DA**
S	kalem**0de**	kale**0de**
O	kalemde	kalede

(f) **Consonant Devoicing**: A morpheme-initial D will surface as an unvoiced t, when affixed to a stem ending in the consonants h, ş, ç, k, p, t, f, s. Furthermore stem-final voiced consonants b, c, d with unvoiced counterparts will assimilate by surfacing as their unvoiced counterparts p, ç, t. For example:

L	kitab**+DA**	tad**+DH**k	saç**+DA**	kitab
S	kitap**0**ta	tat**0**tık	saç**+**ta	kitap
W	kitapta	tattık	saçta	kitap

(g) **Consonant Gemination**: This is a phenomenon that only applies to a set of words imported from Arabic but that set is large enough so that this phenomenon warrants its own mechanism. For this set of words, the root-final consonant is geminated when certain morphemes are affixed. For example:

L	tıb**0+yH**	üs**0+sH**	şık**0+yH**	hak**0+nH**n
S	tıb**b00ı**	üs**s00ü**	şık**k00ı**	hak**k00ı**n
W	tıbbı	üssü	şıkkı	hakkın

(h) **Consonant Changes**: A stem-final k will surface as ğ or g depending on the left context, when followed by the accusative case morpheme +yH, the possessive morpheme +sH or the genitive case morpheme +nHn. A stem-final g will surface as ğ under the same conditions. For example:

L	tara**k+yH**	ren**k+sH**	psikolog**+yH**
S	tara**ğ00ı**	ren**g00i**	psikoloğ**00u**
W	tarağı	rengi	psikoloğu

The phenomena discussed above have many exceptions to them and these are too numerous to cover here in detail. These exceptions are mostly lexicalized and many times some of the rules do not apply when the roots are monosyllabic. For example, even though gök and kök are very similar as far as the affixation boundary is concerned, we have gök+sH → göğü but kök+sH → kökü and not köğü. There are also a set of words, again from Arabic, but ending in vowels where the consonant deletion rule optionally applies and then only in one context but no in another context; e.g., cami+sH could surface as either camisi or as camii, but cami+yH would always surface as camiyi. The orthographic rules for proper nouns also have some bearing on the changes that are reflected to the written forms but they do not impact the pronunciation of those words. The proper noun affix separator ' blocks form changes in the root form. For instance, Işık'+nHn will surface as Işık'ın when written but will be pronounced as /1-S1-G1n/ (note also that when used as a common noun ışık+nHn will surface as ışığın when written and will have the same pronunciation.)

In state-of-the-art finite state formalisms for implementing these rules computationally, one can use either the two-level rule formalism or the cascade-rule formalism, to implement transducers that can map between surface and lexical forms. To implement the exceptions to the rules and many other rare phenomena that we have not covered, one needs to resort to representational mechanisms and tricks to avoid over- and undergeneration. The interested reader can refer to Beesley and Karttunen (2003) for the general formalism-related background and to Oflazer (1994) for details on Turkish two-level implementation.

2.4 Root Lexicons and Morphotactics

In this section we present an overview of the structure of the Turkish words of different root parts-of-speech. Turkish has a rather small set of root words from which very large number of word forms can be generated through productive inflectional and derivational processes. The root parts-of-speech used in Turkish are as follows:

- Nouns
- Verbs
- Numbers
- Postpositions
- Onomatopoeia Words
- Pronouns
- Adjectives
- Adverbs
- Conjunctions
- Determiners
- Interjections
- Question Clitics
- Punctuation

2.4.1 Representational Convention

The morphological analysis of a word can be represented as a sequence of tags corresponding to the overt (or covert) morphemes. In our morphological analyzer output, the tag `^DB` denotes derivation boundaries that we also use to define what we call `inflection groups` (IGs). If we represent the morphological information in Turkish in the following general form:

$$\texttt{root+IG}_1 + \texttt{\^{}DB+IG}_2 + \texttt{\^{}DB+}\cdots + \texttt{\^{}DB+IG}_n.$$

`root` is the basic root word from a root word lexicon and each $\texttt{IG}_i$ denotes the relevant sequence of inflectional features including the part-of-speech for the root (in IG_1) and for any of the derived forms. A given word may have multiple such representations depending on any morphological ambiguity brought about by alternative segmentations of the word, and by ambiguous interpretations of morphemes.

For instance, the morphological analysis of the derived modifier `uzaklaştırılacak` (the one) that will be sent away," literally, "(the one) that will be made far") would be:

```
uzak+Adj
        ^DB+Verb+Become
        ^DB+Verb+Caus
        ^DB+Verb+Pass+Pos
        ^DB+Adj+FutPart+Pnon
```

The five IGs in this word with root `uzak` are: (1) `+Adj`, (2) `+Verb+Become`, (3) `+Verb+Caus`, (4) `+Verb+Pass+Pos`, (5) `+Adj+FutPart+Pnon`.

The first IG indicates that the root is a simple adjective. The second IG indicates a derivation into a verb whose semantics is "to become" the preceding adjective *uzak* "far," (equivalent to "to move away" in English). The third IG indicates that

a causative verb (equivalent to "to send away" in English) is derived from the previous verb. The fourth IG indicates the derivation of a passive verb with positive polarity from the previous verb. Finally, the last IG represents a derivation into future participle which will function as a modifier in the sentence.

2.4.2 Nominal Morphotactics

Nominal stems (lexical and derived nouns) can take up to three morphemes in the order below, that mark

- *Number*: Plural (lack of a number morpheme implies singular—except for mass nouns).
- *Possessive Agreement*: First/second/third person singular/plural (lack of a possessive morpheme implies no possessive agreement).[6]
- *Case*: Accusative, Dative, Ablative, Locative, Genitive, Instrumental, and Equative (lack of a case morpheme implies nominative case).

Thus from a single noun root one can conceivably generate $2 \times 7 \times 8 = 112$ inflected word forms. For instance, the simplest form with the root *ev* "house" is *ev*, which is singular, with no possessive agreement and in nominative case, while one of the more inflected forms would be *evlerimizden* which would be segmented into surface morphemes as `ev+ler+imiz+den` and would be a plural noun with first person plural possessive agreement and ablative case, meaning *from our houses*. In case we need to mark a noun with plural agreement and third person plural possessive agreement (as would be needed in the Turkish equivalent of toys in *their toys* in English as in the fourth case below), we would need to have a form like *oyuncak+lar+ları*. In such cases the first morpheme is dropped with the final word form being *oyuncakları*. But then in a computational setting such surface forms become four ways ambiguous if one analyzes them into possible constituent (lexical) morphemes:

1. `oyuncak+lAr+sH`: his toys
2. `oyuncak+lAr+yH`: toys (accusative)
3. `oyuncak+lArH`: their toy
4. `oyuncak+lArH`: their toys

all of which surface as *oyuncakları*.

Nominal inflected forms can undergo many derivations to create words with noun or other parts-of-speech and each of these can further be inflected and derived.

[6]There are also very special forms denoting families of relatives, where the number and possessive morphemes will swap positions to mean something slightly different: e.g., *teyze+ler+im* "my aunts" vs. *teyze+m+ler* "the family of my aunt."

2.4.3 Verbal Morphotactics

Verbal forms in Turkish have much more complicated morphotactics. Verbal stems (lexical or derived verbs) will inflect, taking morphemes one after the other in (approximately) the following order, marking:

1. *Polarity*: When this morpheme is present, it negates the verb (akin to *not* in English).
2. *Tense-Aspect-Mood*: There can be one or two such morphemes marking features of the verb such as: Past/Evidential Past/Future Tenses, Progressive Aspect, Conditional/Optative/Imperative Moods. Not all combinations of the two morphemes are allowed when both are present.[7]
3. *Person-Number Agreement*: For agreement with any overt or covert subject in the sentence, finite verbs can take a morpheme marking such agreement. The absence of such a morpheme indicates 3rd singular or plural agreement.
4. *Copula*: This morpheme when present adds *certainty/uncertainty* to the verb semantics depending on the verb context.

With just this much a given verb stem can give rise to about 700 inflected forms.

2.4.4 Derivations

Although the number of word forms quoted above are already impressive, it is the productivity of the derivational morphological processes in Turkish that give rise to a much richer set of word forms. However instead of presenting a full set of details on derivations, we will present a series of examples which we hope will give a feel for this richness, after presenting some rather productive derivations involving verbs.

A verb can have a series of voice markers which have the syntactic effect of changing its argument structure. We treat each such voice as a derivation of a verb from a verb. Thus, for example, a verb can have *reflexive*, *reciprocal/collective*, *causative*, and *passive* voice markers.[8] There can be multiple causative markers—two or three are not uncommon, and occasionally, two passive markers. Here is an example of a verbal form that involves four voice markers (with surface morpheme segmentation)

```
yıka+n+dır+t+ıl+ma+mış+sa+m
```

The first morpheme *yıka* is the verbal root meaning "wash/bathe." The next four morphemes mark reciprocal, two causative and passive voice markers. The next four

[7] An example below when we discuss derivation will show a full deconstruction of a complex verb to highlight these features.

[8] Obviously the first two are applicable to a smaller set of (usually) transitive verbs.

morphemes are the inflectional morphemes and mark negative polarity, evidential past, conditional mood and 1st person singular agreement respectively. The English equivalent would be (approximately) "if I were not let (by someone) to cause (somebody else) to have me bathe (myself)." Granted, this is a rather contorted example that probably would not be used under any real-world circumstance, it is nevertheless a perfectly valid example that highlights the complexity of verbal derivations. Verbs can also be further derived with modality morphemes to derive compound verbs with a variety of different semantic modifications. Such morphemes modify the semantics of a *verb* in the following ways[9]:

- able to *verb* (e.g., sür+*ebil*+ir, "she can/may drive")
- keep on *verb*ing (sometimes repeatedly) (e.g., yap+*adur*+du+m "I kept doing (it)")
- *verb* quickly/right away (e.g., yap+*ıver*+se+n, "wish you would do it right away")
- have been *verb*ing ever since (e.g., oku+*yagel*+diğ+iniz "that you have been reading since . . . ")
- almost *verb*ed but didn't (e.g., düş+*eyaz*+dı+m, "I almost fell")[10]
- entered into/stayed in a *verb*ing state (e.g., uyu+*yakal*+dı+m "(literally) I entered into a sleeping state—I fell asleep")
- got on with *verb*ing (e.g., piş+ir+*ekoy*+du+m, " I got on with cooking (it)")

Some of these derivations are very productive (e.g., the first one above) but most are used rarely and only with a small set of semantically suitable verbal roots.

Verbs can also be derived into forms with other parts-of-speech. One can derive a whole series of temporal or manner adverbs with such derivational morphemes having the following semantics:

- after having *verb*ed (e.g., yap+*ıp* "after doing (it)")
- since having *verb*ed, (e.g., yap+*alı*, "since doing (it)")
- when . . . *verb*(s) (e.g., gel+*ince*, "when . . . come(s)")
- by *verb*ing (e.g., koş+*arak* "by running")
- while . . . *verb*ing (e.g., oku+r+*ken* "while reading . . . ")
- as if . . . *verb*ing (e.g., kaç+ar+*casına* "as if . . . running away")
- without having *verb*ed (e.g., bit+ir+*meden* "without having finished")
- without *verb*-ing (e.g., yap+*maksızın*, "without doing")

The final set of derivations from verbs are nominalizations into infinitive or participle forms. After the derivations, the resulting nominalizations inflect essentially like nouns: that is, they can take a plural marker and a possessive marker (which now marks agreement with subjects of the underlying verb), and case marker. Here are some examples:

[9]We present the surface morpheme segmentations highlighting the relevant derivational morpheme with italics.

[10]So the next time you are up on a cliff looking down and momentarily lose your balance and then recover, you can describe the experience with the single verb *düşeyazdım*.

- uyu+*mak* “to sleep,” uyu+*mak*+tan “from sleeping”
- oku+*ma*+m “(the act of) my reading”
- oku+*yuş*+un “(the process of) your reading”
- oku+*duğ*+u “(the fact) that s/he has read”
- oku+*yacağ*+ı+ndan“from (the fact) that s/he will read”

These forms are typically used for subordinate clauses headed by the verb, that function as a nominal constituent in a sentence.

A similar set of derivations result in forms that head clauses acting as modifiers of nouns usually describing the relation of those nouns to the underlying verb as a argument or an adjunct. These correspond to subject-gapped or non-subject-gapped clauses. For example:

- kitap oku+*yan* adam: “The man reading a book”
- kitap oku+*muş* adam “The man who has read a book”
- adam+ın oku+*duğ*+u kitap “The book the man is reading”
- adam+ın oku+*yacağ*+ı kitap “The book the man will be reading”

We mark these derivations as adjectives as they are used as modifiers of nouns in syntactic contexts but add a minor part-of-speech marker to indicate the nature of the derivation. Additionally in the last two cases, a possessive morpheme marks verbal agreement with the subject of the verb.

Although not as prolific as verbs, nouns and to a much lesser extent adjectives can productively derive stems of same or different parts-of-speech. Instead of giving a comprehensive list of these derivations, we would to list some of the more interesting of such derivations:

- Acquire *noun*: para+*lan*+mak “to acquire money”
- Become *adjective*: zengin+*leş*+iyor+uz “we are becoming rich”
- With *noun*: para+*lı* “with money”
- Without *noun*: para+*sız* “without money”

In addition to these more semantically motivated derivations, nouns and adjectives can be derived (sometimes with zero derivation triggered by a tense/aspect morpheme) into forms that function as nominal/adjectival verbs, adverbs, or clauses in a sentence. For example:

- ev+de+*ydi*+k “we were at home”
- ev+de+*yse*+k “if we are at home”
- mavi+*ydi* “it was blue”
- okul+da+*yken* “while he was at school”

2.4.5 Examples of Morphological Analyses

In this section we present several examples of morphological analyses of Turkish words. These examples will also serve to display some of the morphological

ambiguity that ambiguous parts-of-speech, segmentation ambiguity or morpheme homography can cause[11]:

1. *bir*
   - *bir* `bir+Adverb` "suddenly"
   - *bir* `bir+Det` "a"
   - *bir* `bir+Num+Card` "one"
   - *bir* `bir+Adj` "same"
2. *okuma*
   - *ok+um+a* `ok+Noun+A3sg+P1sg+Dat` "to my arrow"
   - *oku+ma* `oku+Verb+Neg+Imp+A2sg` "do not read!"
   - *oku+ma* `oku+Verb+Pos^DB+Noun+Inf2+A3sg+Pnon+Nom` "reading"
3. *koyunu*
   - *koy+u+nu* `koy+Noun+A3sg+P3sg+Acc` "his bay (Accusative)"
   - *koy+un+u* `koy+Noun+A3sg+P2sg+Acc` "your bay (Accusative)"
   - *koyu+n+u* `koyu+Adj^DB+Noun+Zero+A3sg+P2sg+Acc` "your dark (thing) (Accusative)"
   - *koyun+u* `koyun+Noun+A3sg+P3sg+Nom` "his sheep"
   - *koyun+u* `koyun+Noun+A3sg+Pnon+Acc` "sheep (Accusative)"
4. *elmasında*
   - *elma+sı+nda* `elma+Noun+A3sg+P3sg+Loc` "on his apple"
   - *elmas+ı+nda* `elmas+Noun+A3sg+P3sg+Loc` "on his diamond"
   - *elmas+ın+da* `elmas+Noun+A3sg+P2sg+Loc` "on your diamond"
5. *öldürülürken*
   - *öl+dür+ül+ür+ken* `öl+Verb^DB+Verb+Caus^DB+Verb+Pass+Pos+Aor^DB+Adverb+While` "while he is being caused to die"
6. *iyileştirilince*
   - *iyi+leş+tir+il+ince* `iyi+Adj^DB+Verb+Become^DB+Verb+Caus^DB+Verb+Pass+Pos^DB+Adverb+When` "when he is made to become well/good"

[11] Where meaningful we also give the segmentation of the words form into surface morphemes in italics.

7. *ruhsatlandırılamamasındaki*

   - *ruhsat+lan+dır+ıl+ama+ma+sı+nda+ki*

     ```
     ruhsat+Noun+A3sg+Pnon+Nom
       ^DB+Verb+Acquire^DB+Verb+Caus
       ^DB+Verb+Pass^DB+Verb+Able+Neg
       ^DB+Noun+Inf2+A3sg+P3sg+Loc
       ^DB+Adj+Rel
     ```

     "related to (something) not being able to acquire certification"

2.5 The Architecture of the Turkish Morphological Processor

In this section we present a short overview of the implementation of the Turkish morphological processor using the Xerox Finite State Tools (Beesley and Karttunen 2003). These tools take in a description of the morphographemic rules of the language along with the root lexicons and morpheme lexicons and compile them into a (very large) finite state transducer that can map surface forms to multiple analyses. The morphological processor can be customized to produce outputs in different representations, as shown in Fig. 2.1:

- **Morphological Features and Pronunciation**: The output consists of an interleaved representation of both the pronunciation and the morphological features of each possible interpretation of the surface word. For the input word *evinde*, one would get

  ```
  – (e - v )ev+Noun+A3sg(i )+P3sg(n - "d e )+Loc
  – (e - v )ev+Noun+A3sg(i n )+P2sg(- "d e )+Loc
  ```

  Here the parts of the representation between `(...)` encode the pronunciation of the word with phonemes in SAMPA, with `-` denoting syllable boundaries and `"` indicating the syllable with the primary stress. The following shows a more interesting example where we have three analyses for the surface word *okuma* but only two different pronunciations that only differ in the position of the stressed syllable:

  ```
  – (o - k )ok+Noun+A3sg(u - "m )+P1sg(a )+Dat
  – (o - "k u )oku+Verb(- m a )+Neg+Imp+A2sg
  – (o - k u )oku+Verb+Pos(- "m a )
                          ^DB+Noun+Inf2+A3sg+Pnon+Nom
  ```

- **Surface Morphemes**: The output consists of a set of segmentations of the surface word into surface morphemes. So for the input word *evinde*, one would get

  ```
  – ev+i+nde
  – ev+in+de
  ```

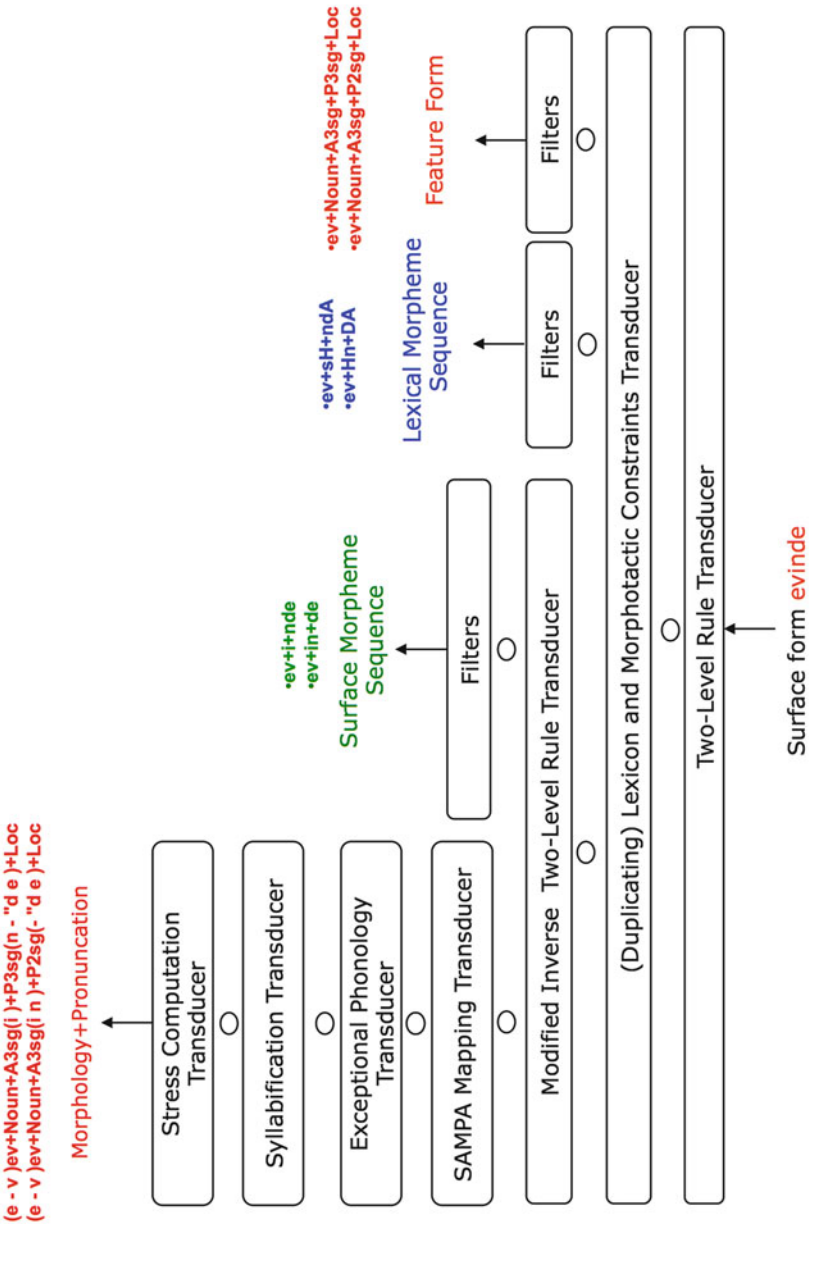

Fig. 2.1 The overall internal transducer architecture of the Turkish Morphological Processor

- **Lexical Morphemes**: The output consists of a set of segmentations of the surface word into lexical morphemes possibly involving meta-symbols denoting miscellaneous subset of phonemes. For the input word *evinde*, one would get
  - `ev+sH+ndA`
  - `ev+Hn+DA`
- **Morphological Features**: The output consists of a set of morphological analyses consisting of the root word followed by a sequence of morphological features encoding the morphological information. So for the input word *evinde*, one would get
  - `ev+Noun+A3sg+P3sg+Loc`
  - `ev+Noun+A3sg+P2sg+Loc`

The root lexicons of the morphological processor comprise about 100,000 root words—about 70,000 being proper names. When fully compiled, the transducer that maps from surface forms to the feature-pronunciation representation ends up having about 10 million states and about 16 million transitions. All transducers are *reversible*, that is, they also be used to generate surface forms from their respective output representations.

Figure 2.1 presents a high-level organizational diagram of the Turkish morphological processor. These transducers are also complemented by a series of helper transducers that help with analysis of misspelled forms, proper nouns violating the use of the apostrophe for separating morphemes, normalizing special characters, etc.

2.6 Processing Real Texts

When one wants to process naturally occurring text from sources ranging from professionally written news text to tweets or blog comments, the basic morphological analyzer proves inadequate for many reasons. In a language like Turkish with almost one to one mapping between phonemes and letters and with morphological processes conditioned by phonological constraints, processing such text requires one to deal with mundane and not so mundane issues brought by tokens encountered in various sources.

2.6.1 Acronyms

An acronym like *PTT* has no vowels as written, but being a noun, it can take suffixes that a normal noun takes. So forms such as *PTT'ye* "to the PTT" or *PTT'den* "from the PTT" are perfectly valid in that the phonological processes are correct *based on the explicit pronunciation of the root PTT*. However forms such as *PTT'ya* or *PTT'ten* are ill-formed as they violate the morphophonological

processes. The problem is that the written form is insufficient to condition the morphological processes: there are no vowels as written and we do not know whether the pronunciation of the root ends in a vowel or in a certain type of a consonant. Such cases have to be handled in the root lexicon by carefully adding enough of a set lexical marker symbols in the representation of such roots so that morphophonological constraints can be conditioned: For instance for *PTT*, we would have to indicate that it ends in a vowel and the last vowel in its pronunciation /pe-te-te/ is written as *e*.

2.6.2 Numbers

Numbers when written using numerals can also take suffixes: e.g., *2014'te* "in 2014," *35.si* "the 35th" *1000'den* "from 1000", *2/3'ü* "the two-thirds of" or *2/3'si* "the two-thirds of." Like acronyms, phonological processes proceed based on the *pronunciation* of the number. So for instance the last vowel in the pronunciation of *2014* (iki bin on dört) is /2/ and the pronunciation ends in the unvoiced dental consonant /t/. The vowel harmony and consonant assimilation rules need to access this information which is nowhere in the written form. Thus one needs to really build the equivalent of a text-to-speech system for numbers which can at least infer the relevant properties of its pronunciation and encode them in its lexical representation. This is a rather nontrivial add-on to the basic morphological analyzer.

2.6.3 Foreign Words

Texts contain foreign common words and proper names which when used as constituents in a sentence have to be properly inflected. However, the morphophonological processes again proceed according to the pronunciation of the foreign word in its original language: one sees sequences like *... serverlar ve clientlar ...* (... servers and clients) where the vowel harmony in the inflected words is fine when the pronunciations of *server* and *client* in English are considered, but not when their written forms (on which the analyzer crucially depends) are considered.[12] Another example is the sports page headline from many years ago:

Bordeaux'yu yendik (We beat Bordeaux (in soccer)).

Here the poor reader who would not know the pronunciation of *Bordeaux* in French would be puzzled by the selection of the +*yu* accusative case morpheme.

This is a tough problem to solve in a principled way. One needs again something akin to a full text-to-speech component that would extract enough information from

[12]Users of such words have the bizarre presumption that readers know how to pronounce those words in English!

the foreign root to condition the morphology. However this is beyond the scope of a morphological processor for Turkish. It may be better to incorporate a solution involving lenient morphology (Oflazer 2003) which can ignore the violation of a very small number of morphographemic rules but only across root-morpheme boundaries.

2.6.4 Unknown Words

When large amounts of text are processed there will always be unknown words that cannot be recognized and analyzed by a morphological processor. Such words can be misspelled words, or to a lesser extent words whose roots are missing from the root lexicons of the morphological analyzer. In a morphological processing pipeline, words found to be misspelled can be corrected with a spelling corrector and an analysis can be reattempted. However for words with roots missing from the lexicon, one can probably do better by analyzing any suffix sequence (assuming the affixes are correct) and then based on those, infer what the roots can be. For instance, when a word sequence such as ... *showları zapladım* ... (I zapped through the shows) is encountered, assuming both are unknown by the morphological analyzer, one can either posit that both words are singular nouns with no possessive agreement and with nominative case or attempt to processes with a version the morphological analyzer whose noun and verb root lexicons have been replaced by a single entry that always matches the regular expression, $?^+$, which matches one or more occurrence of any symbol in the morphological analyzer's surface alphabet. Such an analyzer will skip over any prefix of the unknown words positing the skipped portion as the root, provided the rest of the unknown token can be parsed properly as a valid sequence of morphemes in Turkish. For example, the first word above could be segmented as *showlar+ı* or *show+lar+ı* positing roots *showlar* and *show* respectively and then emitting features corresponding to recognized morphemes. Similarly, the second word can be segmented as *zapla+dı+m* and *zapla* can be posited as a verb root since the remaining morphemes can only attach to verbs.

2.7 Multiword Processing

Multiword expression recognition is an important component in lexical processing that aims to identify segments of input text where the syntactic structure and the semantics of a sequence of words (possibly not contiguous) are usually not compositional. Idiomatic forms, support verbs, verbs with specific particle or pre/postposition uses, morphological derivations via partial or full word duplications are some examples of multiword expressions. Further, constructs such as time-date expressions which can be described with simple (usually finite state) grammars or named-entities and whose internal structure is of no real importance to the

overall analysis of the sentence can also be considered under this heading. Marking multiword expressions in text usually reduces (though not significantly) the number of actual tokens that further processing modules use as input, although this reduction may depend on the domain the text comes from.

Turkish presents some interesting issues for multiword expression processing as it makes substantial use of support verbs with lexicalized direct or oblique objects, subject to various morphological constraints. It also uses partial and full reduplication of forms of various parts-of-speech, across their whole domain to form what we call *non-lexicalized* collocations, where it is the duplication and contrast of certain morphological patterns that signal a collocation rather than the specific root words used. This can be considered another example of morphological derivation involving a sequence of words.

Turkish employs multi-word expressions in essentially four different forms:

1. *Lexicalized Collocations* where all components of the collocations are fixed,
2. *Semi-lexicalized Collocations* where some components of the collocation are fixed and some can vary via inflectional and derivational morphology processes and the (lexical) semantics of the collocation is not compositional,
3. *Non-lexicalized Collocations* where the collocation is mediated by a morphosyntactic pattern of duplicated and/or contrasting components—hence the name *non-lexicalized*, and
4. *Multi-word named-entities* which are multi-word proper names for persons, organizations, places, etc.

2.7.1 Lexicalized Collocations

Under the notion of lexicalized collocations, we consider the usual fixed multiword expressions whose resulting syntactic function and semantics are not readily predictable from the structure and the morphological properties of the constituents.

Here are some examples of the multi-word expressions that we consider under this grouping[13]:

- *hiç olmazsa*
  - `hiç+Adverb`
    `ol+Verb+Neg+Aor+Cond+A3sg`
  - `hiç_olmazsa+Adverb`
    "at least" (literally "if it never is")

[13]In every group we first list the morphological features of all the tokens, one on every line and then provide the morphological features of the multiword construct followed by a gloss and a literal meaning.

- *ipe sapa gelmez*
  - `ip+Noun+A3sg+Pnon+Dat`
    `sap+Noun+A3sg+Pnon+Dat`
    `gel+Verb+Neg+Aor+A3sg`
  - `ipe_sapa_gelmez+Adj`
    "worthless" (literally "(he) does not come to rope and handle")

2.7.2 Semi-lexicalized Collocations

Multiword expressions that are considered under this heading are compound and support verb formations where there are two or more lexical items the last of which is a verb or is a derivation involving a verb. These are formed by a lexically adjacent, direct or oblique object, and a verb, which for the purposes of syntactic analysis, may be considered as single lexical item: e.g., *devam et-* (literally *to make continuation*—to continue), *kafayı ye-* (literally *to eat the head*—to get mentally deranged), etc.[14] Even though the other components can themselves be inflected, they can be assumed to be fixed for the purposes of the collocation, and the collocation assumes its morphosyntactic features from the last verb which itself may undergo any morphological derivation or inflection process. For instance in

- *kafayı ye-* "get mentally deranged" (literally "eat the head")
  - `kafa+Noun+A3sg+Pnon+Acc ye+Verb...`

the first part of the collocation, the accusative marked singular noun *kafayı*, is the fixed part and the part starting with the verb *ye-* is the variable part which may be inflected and/or derived in myriads of ways. With multiword processing, these can be combined into one form

- `kafayı_ye+Verb...`

  For example, the following are some possible forms of the collocation:

- kafayı ye*dim* "I got mentally deranged"
- kafayı yi*yeceklerdi* "they would have become mentally deranged"
- kafayı yi*yenler* "those who got mentally deranged"
- kafayı ye*diği* "the fact that (s/he) got mentally deranged"
- kafayı ye*dirdi* "(he) caused (us) to get mentally deranged"

[14] Here we just show the roots of the verb with - denoting the rest of the suffixes for any inflectional and derivational markers.

Under certain circumstances, the "fixed" part may actually vary in a rather controlled manner subject to certain morphosyntactic constraints, as in the idiomatic verb:

- *kafa(yı) çek-* "get drunk" (but literally "to pull the head")
  - `kafa+Noun+A3sg+Pnon+Acc çek+Verb...`

  which can be replaced by
  - `kafa_çek+Verb...`

- *kafaları çek-*
  - `kafa+Noun+`**`A3pl`**`+Pnon+Acc çek+Verb...`
  - `kafa_çek+Verb...`

  which can also be replaced by
  - `kafa_çek+Verb...`

In these examples, the fixed part has to have the root *kafa* but can be in the nominative or the accusative case, and if it is in the accusative case, it may be marked plural, in which case the verb has to have some kind of plural agreement (i.e., first, second, or third person plural), *but no possessive agreement markers are allowed.*

In their simplest forms, it is sufficient to recognize a sequence of tokens one of whose morphological analyses matches the corresponding pattern, and then coalesce these into a single multiword expression representation. However, some or all variants of these and similar semi-lexicalized collocations present further complications brought about by the relative freeness of the constituent order in Turkish, and by the interaction of various clitics with such collocations.[15]

When such multiword expressions are coalesced into a single morphological entity, the ambiguity in morphological interpretation could be reduced as we see in the following example:

- *devam etti* "(he) continued" (literally "made a continuation")
  - devam
    - `devam+Noun+A3sg+Pnon+Nom` "continuation"
    - `deva+Noun+A3sg+P1sg+Nom` "my therapy"

[15]The question and the emphasis clitics which are written as separate tokens can occasionally intervene between the components of a semi-lexicalized collocation. We omit the details of these.

– etti
  · `et+Verb+Pos+Past+A3sg` "made"
  · `et+Noun+A3sg+Pnon+Nom^DB+Verb+Past+A3sg` "it was meat"
– `devam_et+Verb+Pos+Past+A3sg`
  "(he) continued" (literally "made a continuation")

Here, when this semi-lexicalized collocation is recognized, other morphological interpretations of the components (the second in each group) above) can safely be removed, contributing to overall morphological ambiguity reduction.

2.7.3 Non-lexicalized Collocations

Turkish employs quite a number of non-lexicalized collocations where the sentential role of the collocation has (almost) nothing to do with the parts-of-speech and the morphological features of the individual forms involved. Almost all of these collocations involve partial or full duplications of the forms involved and can actually be viewed as morphological derivational processes mediated by reduplication across multiple tokens.

The morphological feature representations of such multiword expressions follow one of the patterns:

- $\omega\ \omega$
- $\omega\ Z\ \omega$
- $\omega + X\ \omega + Y$
- $\omega_1 + X\ \omega_2 + X$

where ω is the duplicated string comprising the root, its part-of-speech and possibly some additional morphological features encoded by any suffixes. X and Y are further duplicated or contrasted morphological patterns and Z is a certain clitic token. In the last pattern, it is possible that ω_1 is different from ω_2.

Below we present list of the more interesting non-lexicalized expressions along with some examples and issues.

- When a noun appears in duplicate following the first pattern above, the collocation behaves like a manner adverb, modifying a verb usually to the right. Although this pattern does not necessarily occur with every possible noun, it may occur with many (countable) nouns without much of a further semantic restriction. Such a sequence has to be coalesced into a representation indicating this derivational process as we see below.
  – *ev ev* $(\omega\ \omega)$ "house by house" (literally "house house")
    · `ev+Noun+A3sg+Pnon+Nom`
      `ev+Noun+A3sg+Pnon+Nom`

are combined into

· `ev+Noun+A3sg+Pnon+Nom^DB+Adverb+By`

- When an adjective appears in duplicate, the collocation behaves like a manner adverb (with the semantics of *-ly* adverbs in English), modifying a verb usually to the right. Thus such a sequence has to be coalesced into a representation indicating this derivational process.
  - *yavaş yavaş* ($\omega\ \omega$) "slowly" (literally "slow slow")
    · `yavaş+Adj`
    `yavaş+Adj`

    are combined into

    · `yavaş+Adj^DB+Adverb+Ly`

- This kind of duplication can also occur when the adjective is a derived adjective as in
  - *hızlı hızlı* ($\omega\ \omega$) "rapidly" (literally "with-speed with-speed")
    · `hız+Noun+A3sg+Pnon+Nom^DB+Adj+With`
    `hız+Noun+A3sg+Pnon+Nom^DB+Adj+With`

    being replaced by

    · `hız+Noun+A3sg+Pnon+Nom`
    `^DB+Adj+With^DB+Adverb+Ly`

- Turkish has a fairly large set of onomatopoeic words which always appear in duplicate as a sequence and function as manner adverbs. The words by themselves have no other use and literal meaning, and mildly resemble sounds produced by natural or artificial objects. In these cases, the first word can be duplicated but need not be, but both words should be of the part-of-speech category `+Dup` that we use to mark such roots.
  - *harıl hurul* ($\omega_1 + X\ \omega_2 + X$) "making rough noises" (no literal meaning)
    · `harıl+Dup`
    `hurul+Dup`

    gets combined into
  - `harıl_hurul+Adverb+Resemble`

- Duplicated verbs with optative mood and third person singular agreement function, as manner adverbs, indicating that another verb is executed in a manner indicated by the duplicated verb:
  - *koşa koşa* (ω ω)
    - koş+Verb+Pos+Opt+A3sg
      koş+Verb+Pos+Opt+A3sg

    gets combined into
  - koş+Verb+Pos^DB+Adverb+ByDoingSo
    "by running" (literally "let him run let him run")
- Duplicated verbs in aorist mood with third person agreement and first with positive then negative polarity, function as temporal adverbs with the semantics "as soon as one has *verb*ed"
  - *uyur uyumaz* ($\omega + X\ \omega + Y$)
  - · uyu+Verb+**Pos**+Aor+A3sg
    uyu+Verb+**Neg**+Aor+A3sg
    gets combined into
    · uyu+Verb+Pos^DB+Adverb+AsSoonAs

    "as soon as (he) sleeps" (literally "(he) sleeps (he) does not sleep")

  It should be noted that for most of the non-lexicalized collocations involving verbs (like the last two above), the verbal stem before the inflectional marker for mood can have additional derivational markers and all such markers have to duplicate. For example:
  - *sağlamlaştırır sağlamlaştırmaz* "as soon as (he) fortifies (causes to become strong)" ($\omega + X\ \omega + Y$)
    - sağlam+Adj^DB+Verb+Become
      ^DB+Verb+Caus^DB+Verb+**Pos**+Aor+A3sg
      sağlam+Adj^DB+Verb+Become
      ^DB+Verb+Caus^DB+Verb+**Neg**+Aor+A3sg

    which gets combined into
    - sağlam+Adj^DB+Verb+Become+
      ^DB+Verb+Caus+Pos
      ^DB+Adverb+AsSoonAs

An interesting point is that non-lexicalized collocations can interact with semi-lexicalized collocations since they both usually involve verbs. For instance,

below we have an example of the verb of a semi-lexicalized collocation being repeated in a non-lexicalized collocation:

– *kafaları çeker çekmez*

In this case, first the non-lexicalized collocation has to be combined into

– *kafaları* `çek+Verb+Pos^DB+Adverb+AsSoonAs`

and then the semi-lexicalized collocation is handled, to give

– `kafa_çek+Verb+Pos^DB+Adverb+AsSoonAs`

to get an idiomatic case combined with duplication meaning "as soon as (we/you/they) get drunk."

- Finally, the following non-lexicalized collocation involving adjectival forms involving duplication and a question clitic is an example of the last type of non-lexicalized collocation.

  – *güzel mi güzel* (ω *Z* ω) "very beautiful" (literally "beautiful (is it?) beautiful")

  · `güzel+Adj`
  `mi+Ques`
  `güzel+Adj`

  which gets combined into

  – `güzel+Adj+Superlative`

Oflazer et al. (2004) describe a post-processing system that implemented the multi-word processing scheme described above for Turkish.

2.8 Conclusions

This chapter has presented an overview of Turkish morphology and the architecture of a state-of-the-art wide coverage morphological analyzer for Turkish implemented to be used in a variety of natural language processing downstream applications. We also touched upon issues that one encounters when processing real text such as numeric tokens, acronyms, foreign words, unknown words, etc. Finally we gave an overview of the multiwords that one needs to deal after morphological processing but before any further additional processing is attempted.

Appendix: Turkish Morphological Features

In this appendix we present an overview of the morphological features that the morphological analyzer produces. The general format of an analysis is as given in Sect. 2.4.1: any derivations are indicated by ^DB. The first symbol following a ^DB is the part-of-speech of the derived form and the next feature symbol is usually a semantic marker that indicates the semantic nature of the derivation. If the second symbol is +Zero that indicates a implied covert derivation without any overt morphemes.

1. **Major Root Parts of Speech**: These mark the part-of-speech category of the root word. This is not necessarily the part-of-speech of the final word if the word involves one or more derivations.

Feature	Indicates	Feature	Indicates
+Noun	Noun	+Adj	Adjective/modifier
+Adverb	Adverb	+Verb	Verb
+Pron	Pronoun	+Postp	Postposition
+Num	Number	+Conj	Conjunction
+Det	Determiner	+Interj	Interjection
+Ques	Question clitic	+Punc	Punctuation
+Dup	Onomatopoeia words		

2. **Minor Parts of Speech**: These follow one of the part-of-speech category symbols above and either denotes a further subdivision that is morphosyntactically relevant or a semantic marker that indicates the nature of the derivation.

   (a) After +Noun

Feature	Indicates	Example
+Prop	Proper noun	Çağla, Mahkemesi'nde

   (b) After +Pron

Feature	Indicates	Example
+Demons	Demonstrative pronoun	bu "this"
+Ques	Interrogative pronoun	kim "who"
+Reflex	Reflexive pronoun	kendim "myself"
+Pers	Personal pronoun	biz "we"
+Quant	Quantifying pronoun	hepimiz "all of us"

(c) After `+Num`

Feature	Indicates	Example
`+Card`	Cardinal number	iki “two”
`+Ord`	Ordinal number	ikinci “second”
`+Dist`	Distributive number	ikişer “two each”

(d) After `^DB+Noun`

Feature	Indicates	Example
`+Inf1`	Infinitive	gitmek “to go”
`+Inf2`	Infinitive	gitme “going” , gitmem “my going”
`+Inf3`	Infinitive	gidiş (going)
`+PastPart`	Past participle	gittiği (the fact that (he) went)
`+FutPart`	Future participle	gideceği “the fact that he will go”
`+FeelLike`	“the state of feeling like”	gidesim ((the state of) me feeling like going)

(e) After `^DB+Adj`: These are markers that indicate the equivalent of subject, object, or adjunct extracted relative clauses.

Feature	Indicates	Example
`+PastPart`	Past participle	gittiğim [yer] “[the place] I am going”
`+FutPart`	Future participle	gideceğim [yer] “[the place] I will be going”
`+PresPart`	Present participle	giden [adam] “[the man] who is going”
`+NarrPart`	Evidential participle	gitmiş [adam] “[the man] who (is rumored) to have gone”
`+AorPart`	Aorist participle	geçer [not] “passing [grade]” , dayanılmaz [sıcak] “unbearable [heat]”

3. Nominal forms (Nouns, Derived Nouns, Derived Nominal and Pronouns) get the following inflectional markers. Not all combinations are valid in all cases:

   (a) Number/Person Agreement

Feature	Indicates	Example
`+A1sg`	1st person singular	ben “I”
`+A2sg`	2nd person singular	sen “you”
`+A3sg`	3rd person singular	o “he/she/it” , all singular nouns
`+A1pl`	1st person plural	biz “we”
`+A2pl`	2nd person plural	siz “you”
`+A3pl`	3rd person plural	onlar “they” , all plural nouns

   (b) Possessive Agreement

Feature	Indicates	Example
`+P1sg`	1st person singular possessive	kalemim “my pencil”
`+P2sg`	2nd person singular possessive	kalemin “your pencil”
`+P3sg`	3rd person singular possessive	kalemi “his/her/its pencil”
`+P1pl`	1st person plural possessive	kalemimiz “our pencil”
`+P2pl`	2nd person plural possessive	kaleminiz “your pencil”
`+P3pl`	3rd person plural possessive	kalemleri “their pencil(s)”
`+Pnon`	No possessive	kalem “pencil”

   (c) Case

Feature	Indicates	Example
`+Nom`	Nominative	çocuk “child”
`+Acc`	Accusative	çocuğu “child as definite object”
`+Dat`	Dative	çocuğa “to the child”
`+Abl`	Ablative	çocuktan “from the child”
`+Loc`	Locative	çocukta “on the child”
`+Gen`	Genitive	çocuğun “of the child”
`+Ins`	Instrumental/	kalemle “with a pencil”
	accompanier	çocukla “with the child”
`+Equ`	Equative (by object)	bizce “by us”

4. Adjectives do not take any inflectional markers. However, the cases `^DB+Adj+PastPart` and `^DB+Adj+FutPart` will have a possessive marker “one of the first six of the seven above” to mark subject agreement with the verb that is derived into the modifier participle. For example, *gittiğim [yer]* “[the place]

(that) **I** went" will have ... ^DB+Adj+PastPart+P1sg , *gittiğimiz [yer]* "[the place] (that) **we** went" will have ... ^DB+Adj+PastPart+P1pl.

5. Verbs will have multiple classes of markers

   (a) Valency changing voice suffixes are treated as derivations. These voice markers follow ^DB+Verb. A verb may have multiple causative markers.

Feature	Indicates	Example
+Pass	Passive	yıkandı "it was washed"
+Caus	Causative	yıkattı "he had it washed"
+Reflex	Reflexive	yıkandı "he washed himself"
+Recip	Reciprocal/	selamlaştık "we greeted each other"
	Collective	gülüştük "we all giggled"

   (b) The following markers marking compounding and/or modality are treated as deriving new verbs with a semantic twist. These markers also follow ^DB+Verb. All except the first have quite limited applicability.

Feature	Indicates	Example
+Able	Able to *verb*	okuyabilir
		"[s/he] can read"
+Repeat	*verb* repeatedly	yapadurdum
		"I kept on doing [it]"
+Hastily	*verb* hastily	siliverdim
		"I quickly wiped [it]"
+EverSince	have been *verb*ing ever since	bilegeldiğimiz
		"that we knew ever since"
+Almost	Almost *verb*ed but did not	düşeyazdım
		"I almost fell"
+Stay	Stayed/frozen while *verb*ing	uyuyakaldılar
		"they fell asleep"
+Start	Start *verb*ing immediately	pişirekoydum
		"I got on cooking [it]"

   (c) Verbal polarity attaches to a verb (or the last verbal derivation (if any), unless last verbal derivation is from a +Noun or +Adj is a zero derivation).

Feature	Indicates	Example
+Pos	Positive polarity	okudum "I read"
+Neg	Negative polarity	okumadım "I did not read"

(d) Verbs may have one or two tense, aspect or mood markers. However not all combinations are possible.

Feature	Indicates	Example
+Past	Past tense	okudum "I read"
+Narr	Evidential past tense	okumuşum "it is rumored that I read"
+Fut	Future tense	okuyacağım "I will read"
+Prog1	Present continuous tense—process	okuyorum "I am reading"
+Prog2	Present continuous tense—state	okumaktayım "I am in a state of reading"
+Aor	Aorist mood	okur "he reads"
+Desr	Desiderative mood	okusam "wish I could read"
+Cond	Conditional aspect	okuyorsam "if I am reading"
+Neces	Necessitative aspect	okumalı "he must read"
+Opt	Optative aspect	okuyalım "let's read"
+Imp	Imperative aspect	oku "read!"

(e) Verbs also have Person/Number Agreement markers. See above. Occasionally finite verbs with have a copula +Cop marker.

6. Semantic markers for derivations

(a) The following markers mark adverbial derivations from a *verb*—hence they appear after ^DB+Adverb.

Feature	Indicates	Example
+AfterDoingSo	After having *verb*ed	okuyup "after having read"
+SinceDoingSo	Since having *verb*ed	okuyalı "since having read"
+As	As . . . *verb*s	okudukça "as he reads"
+When	When . . . is done *verb*ing	okuyunca "when he is done reading"
+ByDoingSo	By *verb*ing	okuyarak "by reading"
+AsIf	As if *verb*ing	okurcasına "as if he is reading"
+WithoutHaving-DoneSo	Without having *verb*ed	okumadan "without having read" okumaksızın "without reading"

(b) +Ly marks manner adverbs derived from an adjective: *yavaş* (slow) derives *yavaşça* "slowly".

(c) +Since marks temporal adverbs derived from a temporal noun: *aylar* "months" derives *aylardır* "since/for months."

(d) `+With` and `+Without` mark modifiers derived from nouns: *renk* "color" derives *renkli* "with color" and *renksiz* "without color."
(e) `+Ness` marks a noun derived from an adjective with semantics akin to *-ness* in English: *kırmızı* "red" derives *kırmızılık* "redness," *uzun* "long" derives *uzunluk* "length."
(f) `+Become` and `+Acquire` mark verbs productively derived from nouns with the semantics of becoming like the noun or acquiring the noun: *taş* "stone" derives the verb stem *taşlaş* "become a stone/petrify"; *para* "money" derives the verb stem *paralan* "acquire money."
(g) `+Dim` marks derives a diminutive form a noun: *kitap* "book" derives *kitapçık* "little book/booklet".
(h) `+Agt` marks a noun derived from another noun involved in someway with the original noun; the actual additional semantics is not predictable in general but depends on the stem noun: *kitap* derives *kitapçı* "bookseller," *gazete* "newspaper" derives *gazeteci* "journalist," *fotoğraf* derives *fotoğrafçı* "photographer."

7. The following will follow a postposition to indicate the case of the preceding nominal it will subcategorize for. This is not morphologically marked but is generated to help with parsing or morphological disambiguation. Their only use is to disambiguate the case of the preceding noun if it has multiple morphological interpretations.

- `+PCAbl`
- `+PCAcc`
- `+PCDat`
- `+PCGen`
- `+PCIn`
- `+PCNom`

References

Beesley KR, Karttunen L (2003) Finite state morphology. CSLI Publications, Stanford University, Stanford, CA
Clements GN, Sezer E (1982) Vowel and consonant disharmony in Turkish. In: van der Hulst H, Smith N (eds) The structure of phonological representations. Foris, Dordrecht, pp 213–255
Karttunen L (1993) Finite-state lexicon compiler. Technical report, Xerox PARC, Palo Alto, CA
Karttunen L, Beesley KR (1992) Two-level rule compiler. Technical report, Xerox PARC, Palo Alto, CA
Karttunen L, Chanod JP, Grefenstette G, Schiller A (1996) Regular expressions for language engineering. Nat Lang Eng 2(4):305–328
Kornfilt J (1997) Turkish. Routledge, London
Koskenniemi K (1983) Two-level morphology: a general computational model for word-form recognition and production. PhD thesis, University of Helsinki, Helsinki
Oflazer K (1994) Two-level description of Turkish morphology. Lit Linguist Comput 9(2):137–148
Oflazer K (2003) Lenient morphological analysis. Nat Lang Eng 9:87–99

Oflazer K, Inkelas S (2006) The architecture and the implementation of a finite state pronunciation lexicon for Turkish. Comput Speech Lang 20:80–106

Oflazer K, Çetinoğlu Ö, Say B (2004) Integrating morphology with multi-word expression processing in Turkish. In: Proceedings of the ACL workshop on multiword expressions: integrating processing, Barcelona, pp 64–71

Sproat RW (1992) Morphology and computation. MIT Press, Cambridge, MA

van der Hulst H, van de Weijer J (1991) Topics in Turkish phonology. In: Boeschoten H, Verhoeven L (eds) Turkish linguistics today. Brill, Leiden