DESIGN AND IMPLEMENTATION OF A SPELLING CHECKER FOR TURKISH

Ayşın Solak and Kemal Oflazer
Department of Computer Engineering and Information Science
Bilkent University
Bilkent, Ankara 06533 TÜRKİYE
E–mail: ko@trbilun.bitnet, Fax: (90-4)266-4126

(To Appear in Literary and Linguistic Computing, Oxford Univ. Press, 1993)

Abstract: This paper presents the design and implementation of a spelling checker for Turkish. Turkish is an agglutinative language in which words are formed by affixing a sequence of morphemes to a root word. Parsing agglutinative word structures has attracted relatively little attention except for applications areas for general purpose morphological processors. Parsing words in such languages even for spelling checking purposes requires substantial morphological and morphophonemic analysis techniques, and spelling correction (not addressed in this paper) is significantly more complicated. In this paper, we present the design and implementation of a morphological root-driven parser for Turkish word structures which has been incorporated into a spelling checking kernel for on-line Turkish text. The agglutinative nature of the language complex word formations, various phonetic harmony rules, and subtle exceptions present certain difficulties not usually encountered in the spelling checking of languages like English and make this a very challenging problem.

1 INTRODUCTION

Morphological classification of natural languages according to their word structures places languages like Turkish, Finnish, Hungarian, Quechua, and Swahili to a class called "agglutinative languages." In such languages, words are formed by combinining root words and morphemes. There is a root and several suffixes are combined to this root in order to modify and/or extend its meaning. What characterizes agglutinative languages is that stem formation by affixation to previously derived stems is extremely productive [6]. A given stem, even though itself may be quite complex, can generally serve as basis for even more complex words. Consequently, agglutinative languages contain words of considerable complexity, and parsing such word structures for correctness and structural analysis necessitates a thorough morphological and morphophonemic analysis.

Morphological parsing has attracted relatively little attention in computational linguistics. The reason is that nearly all parsing research has been concerned with English, or with languages morphologically similar to English. Since in such languages words contain only a small number of affixes, or none at all, almost all of the parsing models for them consider recognizing those affixes as being trivial, and thus do not make morphological analyses. In agglutinative languages, words contain no direct indication of where the morpheme boundaries are, and furthermore morphemes take a shape dependent on the morphological and phonological context. A morphological parser requires [6]:

1. A morphophonological component which mediates between the surface form of a morpheme as encountered in the input text and the lexical form in which the morpheme is stored in the morpheme inventory, i.e., a means of recognizing variant forms of morphemes as the same, and
2. A morphotactic component which specifies which combinations of morphemes are permitted.

Morphological parsing algorithms may be divided into two classes as *affix stripping* and *root-driven*

analysis methods. Both approaches have been used from very early on in the history of morphological parsing as we learn from Hankamer [6]:

> Packard's parser [15] for ancient Greek proceeds by stripping affixes off the word, and then attempting to look up the remainder in a lexicon. Only if there is an entry in the lexicon matching the remainder and compatible with the stripped-off affixes is the parse deemed a success.
>
> Brodda and Karlsson [3] apply a similar method to the analysis of Finnish, an agglutinative language, but without any lexicon of roots. Suffixes are stripped off from the end of the word until no more can be removed, and what is left is assumed to be a root.
>
> Sagvall [18], on the other hand, devised a morphological analyzer for Russian which first looks in a lexicon for a root matching an initial substring of the word. It then uses grammatical information stored in the lexical entry to determine what possible suffixes may follow.
>
> In the early 1980's, three different approaches to morphological parsing of agglutinative languages were developed independently: for Quechua [8, 9], for Finnish [10], and for Turkish [5]. These three approaches are identical in the way they treat morphotactics. They all proceed from left to right, in the fashion of Sagvall's parser. Roots are sought in the lexicon that match initial substrings of the word, and the grammatical category of the root determines what class of suffixes may follow. When a suffix in the permitted class is found to match a further substring of the word, grammatical information in the lexical entry for that suffix determines once again what class of suffixes may follow. If the end of the word can be reached by iteration of this process, and if the last suffix analyzed is one which may end a word, the parse is successful.

A left-to-right parsing algorithm for automatic analysis of Turkish words was proposed and implemented by Köksal in his Ph.D. thesis [11]. This algorithm, called "Identified Maximum Match (IMM) Algorithm", tries to find the maximum length substring which is present in a root dictionary. If a match is found, i.e., the root morpheme is identified, the remaining part of the word is considered as the search element for suffixes. This part is searched in a suffix morpheme forms dictionary and the morphemes are identified one by one. The process stops when nothing else remains. However in some cases, although a solution is obtained further consistency analysis proves that this solution is not the correct one. In such cases the previous pseudo-solution is reduced by one character and the search procedure is repeated.

These approaches on morphological parsing of Turkish words have the following shortcoming: They do not consider the fact that in an agglutinative language such as Turkish, words contain semantic information that has to be taken into account. In these parsers, it is only the grammatical category of the stem that determine the suffixes that may follow. However, most of the suffixes in Turkish, especially the derivational ones, can be attached only to a limited number of roots or stems and it is the semantics that determines whether a given derivation is a legal word in the current usage of the language, and furthermore such things may evolve over time. For example, in Turkish, the suffix –ALA is a suffix which can be attached to verbal roots and adds the meaning "continuity" to the verb it is applied, e.g., İT-ELE-MEK (to keep on pushing)[1], ŞAŞ-ALA-MAK (to stay confused), etc. Since this suffix is included in both parsers above as a suffix which derives a verb from a verb, verbs like KOŞALAMAK, SEVELEMEK, KONUŞALAMAK, etc. are also parsed correctly although those are not meaningful (or at least not used) verbs in Turkish.

Another shortcoming of the previous parsers for Turkish is that they allow the iterative usage of derivational suffixes. Although, Köksal [11], prevents the consecutive usage of the same morpheme twice, he still parses the word GÖZLÜKÇÜLÜKÇÜLÜK correctly, so does Hankamer [6], though such a word is not used in the language. It is true that some Turkish suffixes can form an iterative loop, but usually the number of iterations is not high. The word above can be parsed correctly up to the point GÖZLÜKÇÜLÜK (the occupation of oculists), but the words GÖZLÜKÇÜLÜKÇÜ and GÖZLÜKÇÜLÜKÇÜLÜK are rather synthetic and never used in the language. Therefore some semantic control mechanisms should be included within the parser to avoid parsing such meaningless words.

[1] From now on, we will indicate the English meaning of a word in Turkish in parentheses following it.

One of the important application areas of parsing words in natural languages (and especially in agglutinative languages) is spelling checking. Although many spelling checkers for English and some other languages have been developed, so far no such tool has been developed for Turkish. As will be discussed in the following sections Turkish poses a number of interesting problems. Wrong ordering of morphemes and errors in vowel or consonant harmonies may cause the wrong spelling of Turkish words. Contrary to other languages like English, in order to check the spelling of a Turkish word, it is necessary to make significant phonological and morphological analyses.

This paper describes a morphological root-driven parser developed for Turkish word structures and its application to spelling checking. A major portion of this work depends on a detailed and careful research on some features of Turkish that make the parsing problem for this language especially hard and interesting. The following section presents an overview of certain morphophonemic and morphological aspects of the Turkish language which are especially relevant to the problem under consideration.(Appendix A gives more detailed discussion of those aspects together with many examples.) Section 3 presents our approach to the problem along with a description of the parser developed. Finally we describe the spelling checker developed along with an evaluation.

2 THE TURKISH LANGUAGE

Turkish is an agglutinative language that belongs to a group of languages known as Altaic languages. In an agglutinative language, the concept of word is much larger than the set of vocabulary items [6]. Word structures can become relatively long by addition of suffixes and sometimes contain an amount of semantic information equivalent to a complete sentence in another language. A popular example of complex Turkish word formation is ÇEKOSLOVAKYALILAŞTIRAMADIKLARIMIZDANMIŞSINIZ whose equivalent in English is "(it is speculated that) you had been one of those whom we could not convert to a Czechoslovakian," where one word in Turkish corresponds to a full sentence in English. Each suffix has a certain function and modifies the semantic information in the stem preceding it. In our example, the root morpheme ÇEKOSLOVAKYA is the name of the (now abolished) state of *Czechoslovakia* and the suffix –LI converts the meaning into *Czechoslovakian*, while the following suffix –LAŞ makes a verb from the previous stem meaning *to become a Czechoslovakian*, and so on.

2.1 Morphophonemics

Being phonetic, the Turkish language can be adapted to a number of different alphabets. In the past, various alphabets have been used to transcribe Turkish, e.g., Arabic. Since 1928, Latin characters have been used. The Turkish alphabet consists of 29 letters of which 8 (A, E, I, İ, O, Ö, U, Ü) are vowels, and 21 (B, C, Ç, D, F, G, Ğ, H, J, K, L, M, N, P, R, S, Ş, T, V, Y, Z) are consonants.

Turkish word formation uses a number of phonetic harmony rules. Vowels and consonants change in certain ways when a suffix is appended to a root, so that such harmony constraints are not violated.

2.1.1 Vowel Change in Suffixes

Almost all suffixes in Turkish use one of two basic vowels and their allophones. We have denoted these sets of allophones with braces around the main vowels A and I, as {A} and {I}. The allophones of {A} are A and E, and {I} represents I, İ, U, or Ü. The vowels O and Ö are only used in root morphemes (especially in the first syllable) of Turkish words.[2]

The vowel harmony rules require that vowels in a suffix change according to certain rules when they are affixed to a stem. The first vowel in the suffix changes according to the last vowel of the stem. Succeeding vowels in the suffix change according to the vowel preceding it. If we denote the preceding vowel (be it in the stem or in the suffix) by v then {A} is resolved as A if v is A, I, O,or U, otherwise it is resolved as E. On the other hand, {I} is resolved as I if v is A or I, as İ if v is E or İ, as U if v is O or U, and as Ü if v is Ö or Ü. For example, the word "YAPMAYACAKTINIZ" can be broken into suffixes as:

[2]The progressive tense suffix –{I}YOR is an exception.

YAP/M{A}/[Y][3]{A}C{A}{K}[4]/{D}[5]{I}/N{I}Z

It can be seen that the vowels in the correct spelling of the word obey the rules above, while a spelling like "YAPMAYACEKTİNİZ" violates the harmony rules because an {A} in the suffix can not resolve to an E as the preceding vowel is an A. It should be mentioned in passing that there are also some suffixes, such as –KEN, whose vowels never change.

2.1.2 Consonant Harmony

Another basic aspect of Turkish phonology is consonant harmony. It is based on the classification of Turkish consonants into two main groups, *voiceless* and *voiced*. The voiceless consonants are Ç, F, T, H, S, K, P, Ş. The remaining consonants are voiced. Interested readers can find the complete list of consonant harmony rules in Appendix A. As an example, one of the rules says that if a suffix begins with one of the consonants D, C, G, this consonant changes into T, Ç, K respectively, if a voiceless consonant is present as the final phoneme of the previous morpheme, e.g., YOLDA (on the road), but UÇAKTA (on the plane).

Some morphemes are affixed with the insertion of either N, S, Ş, Y when two vowels happen to follow each other (e.g. BAHÇESİ (his[6] garden), BAHÇEYİ (accusative of garden), İKİŞER (two each)), or when there is another morpheme following (e.g. BAHÇESİNDE (in his garden), or in context of some pronouns (e.g., BUNA (to this), KENDİNDEN (from yourself)) and the pronomial suffix –Kİ (e.g. SENİNKİNİ (yours/accusative)). In our example above, the future tense suffix –[Y]{A}C{A}{K} comes after the stem YAPMA and since the last phoneme is a vowel Y is inserted.

2.1.3 Root Deformations

Normally Turkish roots are not flexed. However, there are some cases where some phonemes are changed by assimilation or various other deformations [11]. An exceptional case related to the flexion of roots is observed in personal pronouns BEN (I) and SEN (you) having datives BANA (to me) and SANA (to you) respectively. These are individual cases and can be treated as exceptions.

A more systematic ellipsis occurs when the suffix –{I}YOR comes after the verbal roots and stems ending with the phoneme {A}. In such cases, the wide vowel at the end of the stem is narrowed, e.g., YAP → YAPIYOR (he is doing [it]), but ARA → ARIYOR (he is searching).

Another root deformation occurs as a vowel ellipsis. When a suffix beginning with a vowel comes after some nouns, generally designating parts of the human body, which has a vowel {I} in its last syllable, this vowel drops, e.g. BURUN (nose) → BURNUM (my nose). Similarly, when the passive suffix –{I}L is affixed to some verbs, whose last vowel is {I}, this vowel also drops, e.g. ÇAĞIRMAK (to call) → ÇAĞRILMAK (to be called). Other root deformations and their exceptions can be found in Appendix A.

2.2 Morphology

Turkish roots can be classified into two main classes: *nominal* and *verbal*. The verbal class comprises the verbs, while nominal class comprises nouns, pronouns, adjectives, etc. The suffixes that can be received by either of these groups are different, i.e., a suffix which can be affixed to a nominal root can not be affixed to a verbal root with the same semantic function.

[3] [ ] indicates an optional phoneme that must be inserted before a suffix to satisfy certain harmony rules. In this case, [Y] indicates that the consonant Y must be inserted if the last letter of the stem is a vowel, otherwise it is dropped: e.g., OKU (read) → OKUYACAK (s/he will read), but SOR (ask) → SORACAK (s/he will ask).

[4] The two allophones of {K} are K and Ğ.

[5] The two allophones of {D} are D and T.

[6] In Turkish, there is no distinction of gender (masculine, feminine, neuter), and there are no distinct personal pronouns or corresponding possessive suffixes for different genders. So, while giving the English translations, we will use the male correspondings (*he* and *his*) instead of listing all the three possibilities, i.e., *he/she/it* or *his/her/its*.

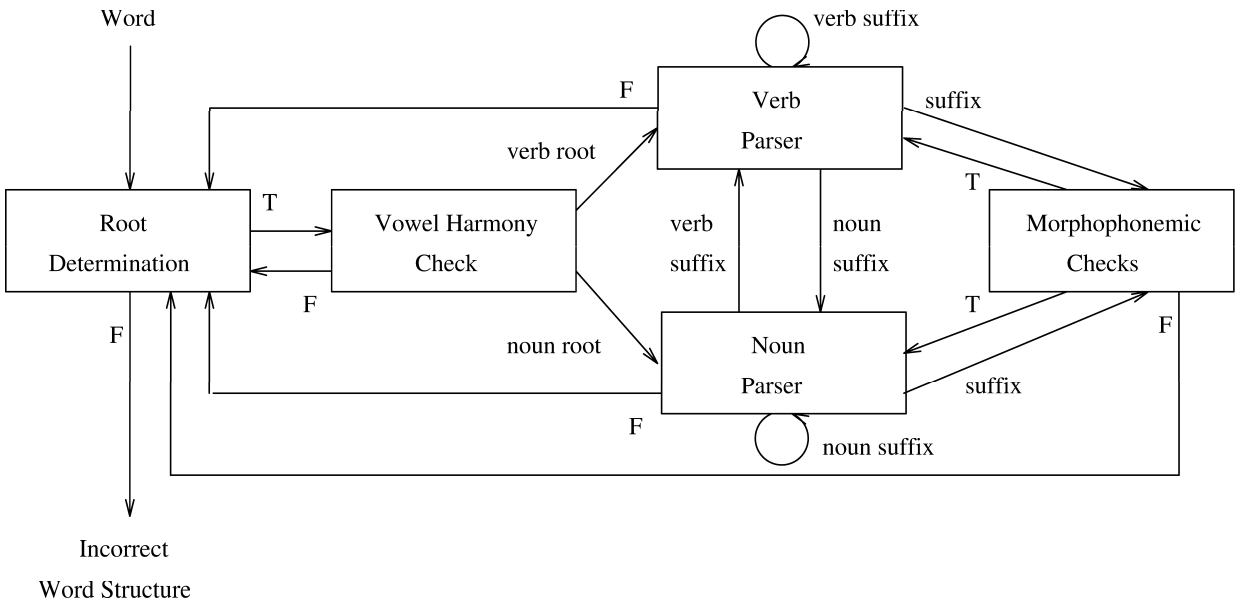

Figure 1: Morphological analysis

Turkish suffixes can be classified as *derivational* and *inflexional*. Derivational suffixes change the meaning and sometimes the class of the stems they are affixed, while a conjugated verb or noun remains as such after the affixation. Inflexional suffixes can be affixed to all of the roots in the class that they belong. On the other hand, the number of roots that each derivational suffix can be affixed changes.

The simplified models for nominal and verbal grammars can be given as follows:[7]

The nominal model:

nominal root + plural suffix + possessive suffix + case suffix + relative suffix

The verbal model:

verbal root + voice suffixes + negation suffix + compound verb suffix + main tense suffix + question suffix + second tense suffix + person suffix

3 A PARSER FOR TURKISH WORD STRUCTURES

Morphological analysis of a Turkish word is handled in three steps:

1. Root determination,
2. Morphophonemic checks, and
3. Morphological parsing.

During these steps a dictionary of Turkish root words, and a set of rules for Turkish morphophonemics, and morphotactics are used concurrently as shown in Figure 1. All these steps are explained in detail in the following sections.

3.1 Root Determination

Before parsing the morphological structures of a Turkish word, the root has to be determined. All parsers use an external list of correctly spelled words in a data structure that serves the function of a dictionary. It is obvious that for an agglutinative language such as Turkish, to provide a dictionary of all possible words is neither an efficient nor a practical approach. So, only root morphemes and some irregular stems are to be held in the dictionary. Our dictionary, of about 23,000 words, has been based on the Turkish

[7] See Appendix A for detailed information on each of the suffixes in these models and the exceptional cases about them.

Writing Guide (Türkçe Yazım Kılavuzu) [24, 25] as the source. The words are placed in a sorted order in an ordered sequential array so that fast searches can be done. Each entry of the dictionary contains a root word in Turkish and a series of flags showing certain properties of that word. We currently have reserved space for 64 different flags for a single word. If the bit corresponding to a certain flag is set for an entry then it means that the word which this entry belongs to has the property represented by that flag. Only 41 flags have been used in the current implementations, but later implementations may use the remaining ones. The list of *some* ofthese flags together with some examples of root words for which those flags are applicable, is given in Table 1. (See [22] for a comprehensive set of these flags.)

The root of a word is searched in the dictionary using a maximal match algorithm. In this algorithm, first the whole word is searched in the dictionary. If it is found then the word has no suffixes and therefore it does not need to be parsed. Otherwise, we remove a letter from the right and search the resulting substring. We continue this by removing letters from the right until we find a root. If no root can be found although the first letter of the word is reached, the word's structure is incorrect.

The maximum length substring of the word that is present in the dictionary is not always its root. If the word can not be parsed correctly using that root, a new root is searched in the dictionary, this time removing letters from the end of the previous root. If a new root can be found the same operations are repeated, otherwise the word is reported as incorrect. For instance, the root of the word YAPILDIN (you were made) is first determined as the noun YAPI (structure). However, the rest of the word does not form a valid sequence of suffixes for a nominal root. Instead of reporting the word as erroneous, a new root is searched, and the verbal root YAP (make, do) is found. Since this one is the real root, the word can be parsed correctly.

As another example consider the word KOYUNLARMI? (are the(y) sheep?) which has an incorrect spelling since the question suffix –M{I} has to be written separate (see page 25). The maximal match algorithm first determines the root as the nominal root KOYUN (sheep), which is the real root, but since the rest of the word can not be parsed correctly, it assumes that the root has been determined wrongly. Hence, a new root is searched and the nominal root KOYU (dark) is found. However, the rest of the word can not be parsed correctly with this root either. Next root determined is the root KOY. This root may either be the nominal root KOY (small bay) or the verbal root KOY (put). Both alternatives are tried but the results are unsuccessful. Since no other root can be found, the word is reported as incorrect.

Root determination presents some difficulties when the root of the word is deformed. For the root words which have to be deformed during certain agglutinations (see Section 2.1.3), a flag indicating that property is set in the dictionary entry. The individual cases such as the dative and plural forms of personal pronouns are inserted into the dictionary and treated as exceptions. For the other root deformations, the root of the word is found by making some checks and some necessary changes. In the following paragraphs, some examples are given to show how the real value of a deformed root is determined.

As the first example, let us consider the vowel ellipsis for nominal roots. In the word OĞLUMUZ (our son) the nominal root OĞUL (son) has taken the shape OĞL when it received the first person plural possessive suffix –[{I}]M{I}Z. In order to determine this root correctly, when the substring OĞL is not found in the dictionary, since it is followed by a vowel, its last two letters are consonants, and the third phoneme from its right end is a vowel, the possibility that it may be a deformed root by vowel ellipsis is considered. The new candidate for the root is obtained by inserting the proper vowel {I}, i.e., U, between the last two consonants of the current candidate, i.e., between Ğ and L, and the word OĞUL is searched in the dictionary. When it is found, the flag corresponding to vowel ellipsis for nominal roots, i.e., IS_UD, is checked. Since it is set for this word, the root of the word OĞLUMUZ is determined as OĞUL, and remaining analyses are performed. If that word were written as OĞULUMUZ, it should be reported as incorrect. In order to handle this case, when the root OĞUL is found in the dictionary, since it is followed by a vowel, the flag IS_UD is checked to see whether it is a root whose last vowel must drop when it is followed by a vowel. Since it is set for this word, but the last vowel of the word has not dropped, the algorithm decides that the root of the word OĞULUMUZ is not the word OĞUL. Later, a new root is searched and since no root can be found, the word OĞULUMUZ has an incorrect structure. As another interesting case, both the words OĞULUM (I am a son) and OĞLUM (my son) are correct, because in the first one, the root OĞUL has received the first singular person suffix –[Y]{I}M (see page 21), while in the second one it received the first person singular suffix –[{I}]M. In order not to report the word OĞULUM as erroneous, when it is recognized that the root OĞUL is a root that has to deform when

Flag	Property of the word for which this flag is set	Examples
CL_NONE	belongs to none of the two main root classes	RAĞMEN, VE
CL_ISIM	is a nominal root	BEYAZ, OKUL
CL_FIIL	is a verbal root	SEV, GEZ
IS_OA	is a proper noun	AYŞE, TÜRK
IS_OC	is a proper noun which has a homonym that is not a proper noun	MISIR, SEVGİ
IS_SAYI	is a numeral	BİR, KIRK
IS_LAS	is a nominal root which can take the suffix –L{A}Ş	KENT, UYGAR
IS_LAT	is a nominal root which can take the suffix –L{A}T	AYDIN, KİR
IS_CI	is a nominal root which can take the suffix –{C}{I}	DAVA, KAVGA
IS_CILIK	is a nominal root which can take the suffix –{C}{I}L{I}{K}	KAR, ÜMMET
IS_CA	is a plural noun	BAKLAGİLLER
IS_KI	is a nominal root which can directly take the relative suffix –Kİ	BERİ, ŞİMDİ
IS_KU	is a nominal root which can directly take the relative suffix –KÜ	BUGÜN, ÖBÜR
IS_UU	is a nominal root which does not obey the vowel harmony rules during agglutination	SAAT, NORMAL
IS_UUU	is a nominal root which has a homonym that does not obey the vowel harmony rules during agglutination	SOL, YAR
IS_SD	is a nominal root ending with a consonant which is softened when a suffix beginning with a vowel is attached	AMAÇ, PARMAK, PSİKOLOG
IS_SDD	is a nominal root ending with a consonant which has a homonym whose final consonant is softened when a suffix beginning with a vowel is attached	ADET, KALP
IS_B_SI IS_SU	is a compound word ending with the third person is a nominal root which shows the irregularities that the root SU shows	ALINYAZISI, AKARSU
F_UD	is a verbal root which has a vowel {I} in its last syllable that drops when the passiveness suffix –{I}L is affixed	AYIR, SAVUR

Table 1: A partial list of flags for dictionary entries

it is followed by a suffix beginning with a vowel, the algorithm checks whether that suffix is one of the suffixes –[Y]{I}M or –[Y]{I}Z.

Another root deformation is the change of the last consonant in some roots. For example, in the word TABAĞIM (my dish), final consonant of the nominal root TABAK (dish), i.e., K, has changed into Ğ, when the first person singular possessive suffix is affixed. In this case, when the substring TABAĞ is not found in the dictionary, since it is followed by a vowel, and its last phoneme is one of the consonants B, C, D, G, and Ğ, the possibility that it may be a deformed root whose last phoneme has changed is considered. Since it does not end with the substring LOĞ,[8] and the final phoneme is not preceded by the consonant N,[9] the final phoneme Ğ is replaced with the consonant K, and the word TABAK is searched in the dictionary. When it is found, the flag corresponding to the change of the final consonant, i.e., IS_SD, is checked. Since it is set for this word, the root of the word TABAĞIM is determined as TABAK. If that word were written as TABAKIM, it would be reported as incorrect.

As another example, let us consider the duplication of the final consonant for some nominal roots. In the word HAKKINIZ (your right), the consonant K at the end of the root HAK (right) is duplicated when it receives the second person plural possessive suffix. When the substring HAKK can not be found in the dictionary, since it is followed by a vowel, its last two phonemes are the same consonants, and the third phoneme from its right is a vowel, the possibility that its last phoneme may have been duplicated is considered. Its last phoneme is deleted and the word HAK is searched in the dictionary. When it is found, the flag corresponding to the duplication of the final consonant, i.e., IS_ST, is checked. Since it is set for this word, the root of the word HAKKINIZ is determined as HAK. If that word were written as HAKINIZ it would be reported as incorrect. As another interesting example, the root of the word TIBBIN (medicine/possessive) is the word TIP (medicine) where its last phoneme is duplicated after changing into a B. In this case, as in the previous one, one of the B's is removed from the end of the word TIBB and the word TIB is searched in the dictionary. When it is not found, since its last consonant is B, it is changed into a P, and the word TIP is searched in the dictionary. When it is found, both the flags IS_ST and IS_SD are checked. Since both are set for this word, the root is determined as TIP. If that word were written as TIPIN, TIBIN, or TIPPIN, it would be erroneous.

For all the other deformations such as vowel ellipsis in the verbal roots, narrowing of the final wide vowel in the verbal roots, midfixing of the plural suffix to the compound words, etc., and their combinations, both the correct and incorrect usage of the roots are determined by using similar methods to the ones above.

For some roots both of the deformed and undeformed forms are valid. For example, both METNİ (text/accusative) and METİNİ (strong/accusative) are correct although the root of both words is METİN (text/strong). Such cases are handled again by the help of certain flags, IS_UDD, IS_SDD, and IS_STT. For instance, to determine the root of the word METNİ as METİN, checking only the flag IS_UD is enough. On the other side, in order not to report the word METİNİ as incorrect, when the root METİN is found, the flag IS_UDD is checked. Since it is set for this word, the root is determined as METİN. Similarly, none of the words ADEDİ (ADET: amount), ADETİ (ADET: custom), ŞIKKI (ŞIK: option), or ŞIKI (ŞIK: chic) is reported as erroneous.

The algorithm for root determination sometimes requires a lot of searches in the dictionary. To determine the root of the word OKULA (to the school), two searches (one for OKULA and the other for OKUL) are enough, but to determine the root of the word ALDIĞIMIZ (that we took), the dictionary is searched 13 times for the words ALDIĞIMIZ, ALDIĞIMI, ALDIĞIM, ALDIĞI, ALDIĞISI, ALDIĞ, ALDIK, ALDI, ALD, ALID, ALIT, ALT, and AL, respectively. Our analyses and tests indicate that on the average 5 to 6 root word look-ups in the dictionary are performed to parse a word.

3.2 Morphophonemic Checks

After the root of the word is found, the rest of the word is considered as the suffixes. Vowels and consonants within suffixes should obey certain rules during agglutination (see Section 2.1). Therefore, the suffixes part of a word must be checked to see whether any of the morphophonemic rules are violated.

[8] If the word were PSİKOLOĞA (to the psychologist), this condition would hold and Ğ would be replaced not with a K but with a G.

[9] If the word were RENĞE, this condition would hold and no replacements would be done.

The vowel harmony check may be done just after the root determination, but other morphophonemic checks should be done during morphological parsing.

3.2.1 Vowel Harmony Check

According to the vowel harmony rules of Turkish (see Section 2.1.1), the first vowel in a suffix must be in harmony with the last vowel of the root, while the succeeding vowels must be in harmony with the vowel preceding them. For example, the word YAPMEK can not pass the vowel harmony check because the vowel E can not follow the vowel A. On the other hand, special checks must be done for the suffixes, such as –KEN, whose vowels do not change. So, when a disharmony is found, we check whether it is the result of such a suffix. For example, after the root of the word YANARKEN (while it is burning) is found as YAN (side, burn), the suffixes part, i.e., ARKEN, is checked to determine whether the word obeys vowel harmony rules. The first vowel A is in harmony with the last vowel of the root, but the next vowel E is not in harmony with the vowel preceding it. At this point, instead of deciding that the word does not obey vowel harmony rules, the phonemes preceding and following the current vowel are checked to determine whether that vowel belongs to one of the suffixes which do not obey vowel harmony rules, i.e., to –[Y]KEN, –[Y]{I}VER, or –[Y]{A}GEL. Since it does, the word passes the vowel harmony check. If this word was written as YANARKAN, it would pass the vowel harmony check, but it would not be parsed correctly during morphological analysis.

Before the vowel harmony check is done, some flags of the root must be checked. For example, if the word is a word of foreign origin that does not obey vowel harmony rules during agglutination (e.g., KONTROL (control)), a vowel disharmony check must be performed. The first vowel in the suffixes part must be in disharmony with the last vowel of the root (e.g., KONTROLLER (controls)). The flag IS_UU is checked to realize such cases. Some roots that are polysemious present another interesting case. They obey vowel harmony rules when they are used with a certain meaning, but disobey them when they are used in the other meaning. For example, both SOLA (to the left) and SOLE (to the note sol) pass the vowel harmony check since their root SOL has two meanings as "left" and "a note in musics."[10] Such cases are handled by the help of the IS_UUU flag.

Another special case occurs when a root which does not obey vowel harmony rules within itself deforms by vowel ellipsis. For example, the root of the word NAKLİ (its transfer) is the noun NAKİL (transfer). If the vowel harmony check is done accepting the root as NAKL it fails because the vowel İ can not follow the vowel A. In such cases, not the deformed root but the real root appearing in the dictionary must be considered, and the suffixes part must be in harmony with the real root, i.e., in our example with the word NAKİL. The wrong form, i.e., NAKLI would also be realized, but not during the vowel harmony check, instead during root determination, because the proper vowel to be inserted between the consonants K and L would be determined as I, and the word NAKIL would then not be found in the dictionary.

A more interesting case is caused by some roots which may deform or not depending on the meaning that they carry. Such roots obey vowel harmony rules when they are not deformed, but not when they are deformed (e.g., AD, KALP). For such roots, the flags to be checked are IS_UUU, IS_STT, and IS_SDD. Therefore, while all the words ADI (AD: name), ADDİ (AD: count), KALPI (KALP: unreliable), and KALBİ (KALP: heart) are correct, the words ADDI,[11] KALPİ, and KALBI can not pass the vowel harmony check.

3.2.2 Other Checks

To perform the other morphophonemic checks, the suffixes must be determined. Because of this, these checks are done during morphological parsing, after each suffix is isolated. During the lexical analysis, if any of the allomorphs of a suffix can be matched, it is sent to the parser without checking whether the correct form of it is used. These checks are done within the parser. Since the vowel harmony check is done beforehand, only the remaining morphophonemic checks must be done at that point. The consonant harmony checks are among these checks (see Section 2.1.2).

[10] The word SOL is pronounced slightly different in the latter case.

[11] The word ADİ passes the check because such a word is present in the dictionary.

Consider the words YAPDIKÇA, YAPTIĞÇA, YAPTIKCA, YAPTIĞCA, and YAPTIKÇA. For all of them, the root will be determined as the verbal root YAP (do). Additionally, all will pass the vowel harmony check. Furthermore, for all of them the suffixes will be isolated as the participial suffix –{D}{I}{K} and the external case suffix –{C}{A}, respectively, and they form a valid sequence of suffixes for a verbal root. However, it is obvious that only one of them (YAPTIKÇA) is correct. In order to recognize the incorrect ones consonant harmony checks must be done. When the suffix –{D}{I}{K} is isolated, since it is a suffix whose initial phoneme changes depending on the phoneme preceding it, the last phoneme of the root YAP is checked. Since it is a voiceless consonant, the suffix must begin with the consonant T. Therefore, the word YAP**D**IKÇA can not pass this check. In addition, the last phoneme of that suffix changes depending on the phoneme it precedes. Since it is followed by a consonant, it must end with the voiceless consonant K. Hence the word YAPTI**Ğ**ÇA is also wrong. Later comes the suffix –{C}{A} whose first phoneme depends on the last phoneme of the stem it is affixed to. The word YAPTIK**C**A can not pass this check because although the suffix –{C}{A} comes after the voiceless consonant K, it does not begin with the voiceless consonant Ç.

Usage of passing vowels or consonants are also checked during morphological analysis. For example, during the morphological analysis of the word GELİYORKEN (while [3^{rd} person singular] is coming), when the first suffix is determined as the progressive tense suffix –[{I}]YOR, since the passing vowel {I} is used, the last phoneme of the root is checked to see whether it ends with a consonant. Later, the participial suffix –[Y]KEN is isolated. Since the passing consonant Y is not used, the phoneme preceding it is checked to see if it is a consonant. If this word were written as GELYORKEN, GELİYORYKEN, or GELYORYKEN, it could not pass the morphophonemic checks, although it obeys to vowel harmony rules and the order of the morphemes are correct.

If a word can not pass any of the morphophonemic checks, considering the possibility that the root may have been determined wrongly, a new root is searched in the dictionary, and the process is repeated.

3.3 Morphological Parsing

For the morphological parsing of Turkish words two separate sets of rules for the two main root classes have been prepared. When the root of a word is found the class of the root determines which set of rules are to be used for further parsing.

3.3.1 Utilities Used

For the implementation of the lexical analyzers and parsers in which the rules are included, two standard UNIX utilities, *lex* and *yacc*, have been utilized respectively [12, 19]. *Lex* and *yacc* were designed as tools to help programmers writing compilers and interpreters, but they have a wide range of applications.

Lex, so called because it generates a lexical analyzer, reads a stream of bytes and groups them into tokens. The user provides a set of high-level, problem-oriented specifications for regular expression matching, and *lex* produces a program in C programming language which recognizes those regular expressions. We have used it to separate the suffixes of a word from left to right.

Yacc (which stands for Yet Another Compiler–Compiler) is used to codify the grammar of a language, and generates a parser. The parser examines the input tokens and groups them into syntactical units. The value of the tokens may be processed by action routines written in C. We have used *yacc* to parse the suffixes using morphological rules of Turkish grammar.

3.3.2 Lexical Analyzers

Two sets of *lex* specifications, one for each root class, are prepared to generate the lexical analyzers which are to be called by the parsers each time a new token is needed. The specifications contain regular expressions that match suffix tokens. The lexical analyzer corresponding to the category of the current stem, sends, as the next suffix token, the maximum length substring from the left of the remaining suffixes part that matches to any allomorph of a suffix in the permitted class.

The following is a small section from the *lex* specification for verbs:[12]

```
A           [AE]
I           [iIUu]
.
%%
.
M{A}L{I}    return (MALI);
M{A}        return (MA);
.
```

Using this specification, the first suffix token of both the words YAPMALISIN (you must do) and GELMELİYIM (I must come) is isolated as the necessitative suffix –M{A}L{I}. Thus, although the suffix –M{A} is also a substring of those words, since its length is less than the suffix –M{A}L{I}, the longest one is matched. If the wrong allomorph of the suffix were used in one of these words, for instance, if the first one were written as YAPMELİSİN, it would be recognized during vowel harmony check.

The morphotactic structure of some words can be analyzed in more than one form. For example, the word EVİNİN may be analyzed into two morphotactic structures as

EV + [S]{I} + [N]{I}N → EVİNİN (his house's), and
EV + [{I}]N + [N]{I}N → EVİNİN (your house's).

However, if a word can be analyzed correctly in one form, we do not look for other possible structures. For instance, using the following *lex* specification prepared for nouns, the word EVİNİN is analyzed as in the second form.

```
I           [iIUu]
.
%%
.
N{I}N       return (NIN);
{I}N        return (IN);
N           return (N);
.
```

Similarly, the maximum length suffix matched for the word KAPININ (the door's, or your door's) is the genitive suffix –[N]{I}N, although that word may have been formed by combining the suffixes –[{I}]N and –[N]{I}N.

The lists of all the suffixes included into the grammar rules for each root class can be found in Appendix B. Certain combinations of these suffixes are matched as if a single suffix token by the lexical analyzers, so that some rules can be simplified. For example, the combination of the negation suffix with the progressive tense suffix is matched as a single suffix –M{I}YOR, to eliminate the check for the deformation of the negation suffix (see page 24). On the other hand, some suffixes are formed by the combination of more than one tokens sent by a lexical analyzer. For example, instead of matching the third person plural possessive suffix –L{A}R{I} as a single suffix token, when the lexical analyzer for nouns sends the third person singular possesive suffix –[S]{I} after the plural suffix –L{A}R, their combination is treated as the suffix –L{A}R{I}.

3.3.3 Parsers

The grammar rules for morphotactics of Turkish word structures have been described in two *yacc* specifications, one for nominal and one for verbal roots. The lexical analyzers described in the previous section

[12] This specification consists of two parts as *definitions* and *rules* section, which are separated by the symbol %%. The definition part contains some *substitutions* which define regular expressions employed in the rules section. These definitions are then referenced by placing braces ({}) around the desired substitution string. For detailed information on *lex* specifications refer to [12] or [19].

produce the suffix token stream. *Yacc* generates the source files for the parsers.

All the models in Section A.2 have been utilized for generating the rules used in the parsers. Additionally, all of the known exceptional cases have been considered. The correct order of suffixes are coded as grammar rules, and necessary checks are done by the help of action routines associated with the rules. Those routines are executed each time the rule is matched. For example, when the lexical analyzer for the noun parser sends –{C}{I} as the suffix token for the word KİTAPÇI (book seller), first the IS_CI flag of the root KİTAP (book) is checked to see whether that root can really receive the suffix –{C}{I}. This flag is set for this root, but one more check is necessary to determine whether the correct allomorph of the suffix is used. The value of the vowel in the suffix has been proven to be correct by the vowel harmony check, therefore, it is only necessary to prove that the suffix must really begin with the consonant Ç in its this usage. Therefore, the final phoneme of the stem it is affixed to is checked, and when it is seen that it is the voiceless consonant P, Ç is proven to be the correct allophone for {C}, i.e., the correct allomorph of the suffix is used. If the word were written as KİTAPCI it would not have passed this check. On the other hand, the word SEVİNÇÇİ will not be parsed correctly because the nominal root SEVİNÇ (happiness) is not marked in the dictionary as a root which can receive the suffix –{C}{I}.

To check whether the correct allomorph of a suffix is used is relatively simple if only the phonetic conditions are to be considered. For the suffixes whose allomorphs change depending on certain rules, such as the causative verb suffix, passive voice verb suffix, and aorist suffix, extra checks must be done. As an example, let's consider the aorist suffix. When the lexical analyzer for the verb parser sends the aorist suffix as the current suffix token, the parser controls whether the correct allomorph of the suffix is used depending on the stem it is affixed to. If the –R allomorph of the suffix is used, the final phoneme of the stem it follows must be a vowel (e.g., OYNAR (he plays)). If the –{I}R allomorph is used, the stem it is affixed to must end with a consonant, and must contain more than one syllables but must not be a compound verb formed with the verb ETMEK, i.e., the flag IS_GER must not be set for that root (e.g., KAYBOLUR (he disappears)), or must be a mono-syllabic root for which the IS_GIR flag is set (e.g., VERİR (he gives)). Otherwise, if the –{A}R allomorph is matched, the stem must again end with a consonant, but this time must be mono-syllabic and the IS_GIR flag must not be set (e.g., YAPAR (he does)), or it must be a compound verb formed with the verb ETMEK (e.g., HİSSEDER (he feels)). As a result of this check the incorrect words such as KAYBOLAR, VERER, YAPIR, HİSSEDİR will be detected.

As an example for difficulties faced during such checks, consider the passive voice suffix –{I}N, and the second person plural suffix for the imperative form of verbs, i.e., –[Y]{I}N. These two suffixes may sometimes take the same form as in the word BULUN. In this word, the suffix –UN may be either of the suffixes –{I}N or –[Y]{I}N. Since the passive voice suffix takes different forms depending on the stem it follows, some checks must be done when any of those forms are matched. If the suffix –UN is considered as the passive voice suffix, the check will be successful since the root BUL ends with the consonant L. If the other possibility is considered, the word will again be parsed correctly since the person suffix must be the last suffix. On the other hand, while the word KAPATIN is being parsed, if the suffix –IN is considered to be the passive voice suffix, it can not pass the check, where it will be parsed correctly if it is considered as the person suffix. To solve this problem, when the suffix –{I}N is matched as the last suffix of a word, it is decided to be the person suffix, and therefore, no check for the passive voice suffix is done. Otherwise, if there exists any suffix following that suffix, it is considered to be the passive voice suffix and the check is done.

The two parsers are alternatively used. First parser to be used is determined according to the class of the root, but as the parsing continues it may be necessary to switch from one parser to another and continue there, or again pass back to the previous one, since the class of a stem can change when it receives certain suffixes. For example, while parsing continues in the noun parser, if the derivational suffix –L{A}Ş, which makes a verb from a noun, is matched, a jump to the verb parser must be done. Such jumps are not possible using the C code generated by *yacc* as it is, so some modifications are done in that code automatically after each time it is generated.

The switches between parsers can sometimes be very complicated. Some suffixes can have two different usages. For instance, the suffix –M{A} can either make a verb a noun or negate it. In such cases both possibilities have to be considered. For example, after the root of the word YAPMADIM (I didn't do) is determined as the verbal root YAP (do), the first suffix will be isolated as –M{A} in the verb parser. First

Input Word:	ÇEKOSLOVAKYALILAŞTIRMADIKLARIMIZDANMIŞSINIZ
Root:	ÇEKOSLOVAKYALI

Input for Noun Parser	Input for Verb Parser
LAŞTIRMADIKLARIMIZDANMIŞSINIZ	TIRMADIKLARIMIZDANMIŞSINIZ
DIKLARIMIZDANMIŞSINIZ	DIKLARIMIZDANMIŞSINIZ
LARIMIZDANMIŞSINIZ	

Table 2: An example to parsing process and switch between parsers

considering the possibility that this suffix is used as a derivational suffix, the noun parser will be invoked. The remaining part of the word can not be parsed by this parser. So accepting –M{A} as the negation suffix, the verb parser will be returned to and parsing will be continued there. On the other hand, since the same suffix is used as a derivational suffix in the word YAPMANIZ (your doing), this word will be parsed successfully in the noun parser, thus returning to the verb parser will not be necessary.

If a word has received more than one derivational suffixes then many switches between parsers will be necessary. In Table 2 an example to such switches is given. In that example, the root of the word ÇEKOSLOVAKYALILAŞTIRMADIKLARIMIZDANMIŞSINIZ (you had been one of those whom we did not convert to a Czechoslovakian) is found as the noun ÇEKOSLOVAKYALI (Czechoslovakian) in our dictionary. Then comes the suffix –L{A}Ş, therefore, a switch to verb parser has to be made. Parsing continues there until the suffix –M{A} is matched. Supposing that this suffix has changed the class of the stem, the noun parser will be returned back. Since the remaining part can not be parsed in the noun parser, the verb parser is activated, and parsing will continue there considering –M{A} as the negation suffix. Then comes the suffix –{D}{I}{K}, which is also a suffix that makes a noun from a verb, therefore, again a switch to the noun parser will be made. Continuing in this parser, the word will be parsed correctly.

For the roots that can take all the suffixes belonging to both nominal or verbal classes, if parsing is unsuccessful in the first parser chosen, the other one must also be tried. For example, the root of the word AÇLAR (hungry people) is AÇ. This root may either be used as a verb (open) or as a noun (hungry). Parsing is first attempted with the verb parser, but this fails. So we backtrack and use the other parser. With the noun parser the word can be parsed successfully.

In Figure 2 an example *yacc* specification[13] is given. These rules appear within the grammar rules for the nominal roots. They are used to parse a word whose root is a numeral. The terminal SAYI indicates that a numeral root has been matched. The rules for the suffixes that a numeral root can receive are represented by the non-terminal sayi_ek. The rules for the non-terminal sayi_isim says that a numeral root stays as a noun if it receives the suffixes –[{I}]NC{I} (the token INCI), –L{I}{K} (the token LIK), or a combination of them: e.g., BİRİNCİ (first), BEŞLİK (set of five), ÜÇÜNCÜLÜK (third place). The suffix –[{I}]NC{I} must take the form –NC{I} when it follows a root ending with a vowel (e.g., İKİNCİ (second)). Because of this, the usage of the passing vowel {I} is checked by the routine Check_I. The non-terminal sayi_fiil shows that by affixing the suffix –L{A} or –L{A}T (the tokens LA and LAT respectively) to a numeral root, a verb can be derived: e.g., KIRKLAMAK, DÖRTLETMEK. The suffix –[Ş]{A}R (the token SAR) may be affixed to a numeral root either alone or after combining with one of the suffixes –L{I}{K} or –L{I} (the tokens LIK and LI respectively): e.g., ALTIŞAR (six each), YEDİŞERLİ (with seven each), YÜZERLİK (able to contain hundred each). Since the consonant Ş is only used in this suffix

[13] This specification consists of two parts as *declarations* and *rules* section, which are seperated by the symbol %%. Token definitions in the declarations section describe all possible tokens that the lexical analyzer will return to the parser, thus the *terminals*. The concatenation and/or union of these tokens form *nonterminals*, which may themselves be used as tokens in other rules. Actions can be associated with a rule. An action consists of C code that will be executed each time the rule is matched. For detailed information on *yacc* specifications refer to [12] or [19].

```
.
.
% token SAYI SAR INCI LIK LAT LA LI
.
%%
.
ad          :   SAYI sayi_ek
            ;

sayi_ek     :   sayi_isim       { call_isim; }
            :   sayi_fiil       { call_fiil; }
            :   sar sayi_oth
            :   LI              { if (Next_YOR) call_fiil; else call_isim; }
            ;

sar         :   SAR             { Check_SAR; }
            ;

sayi_isim   :   INCI            { Check_I; }
            :   INCI LIK        { Check_I; }
            :   LIK
            ;

sayi_fiil   :   LAT
            :   LA
            ;

sayi_oth    :   LIK
            :   LI
            ;
```

Figure 2: *Yacc* specification for numerals

when it is affixed to a root ending with a vowel, its usage is checked by the routine Check_SAR. If the suffix –L{I} comes immediately after a numeral root, if it is followed by the substring YOR it may be the deformed form of the suffix –L{A} (e.g., KIRKLIYORLAR), therefore, a call to the verb parser is done, otherwise the class of the stem remains as a noun.

In the current implementation, the grammar for verb parser consists of 230 rules in which 80 terminals and 81 nonterminals are used, and the grammar for noun parser, consists of 263 rules in which 68 terminals and 94 nonterminals are used.

4 A SPELLING CHECKER FOR TURKISH

Spelling checking is one of the major application areas of parsers for agglutinative languages. We used the morphological parser developed in the implementation of a spelling checker for Turkish. Our approach to spelling error detection is based on checking individual words in the text file by parsing them with no attention to the context. Thus, if a word can be parsed correctly but is the wrong word in the context, we have no intention for and way of flagging it as erroneous. Thus, as in all other spelling programs, the text is examined with respect to words, not with respect to sentences. In addition, we do not yet give any suggestion about the most likely correct words after detecting a misspelled word, i.e., spelling correction is not done.

Syllabification check is used as a heuristic in the spelling checker. The heuristic is *if a word does NOT have*

the proper syllable structure of Turkish, it is misspelled. Analyzing all the words in Turkish Writing Guide [24, 25] and all the suffixes in Turkish [1], we have constructed a regular expression and a corresponding finite state automaton for validating if a word matches the syllable structure rules of Turkish[20]. The word whose spelling is to be checked is first processed with the regular expression. It is reported as misspelled if its syllable structure can not be matched with this expression, i.e., the phonemes of the word do not form valid sequences according to Turkish syllable structures. On the other hand, if it can be matched, its morphological structure is analyzed as it may still be a non-Turkish or a misspelled word. If the morphological structure of the word is found to be incorrect during any step of the analysis, the word is reported as misspelled.

The current lexicon of the spelling checker is based on a list of about 23,000 root words, which covers almost all the root words in the language as listed in various sources. We have also included a large number of technical words for various disciplines like computer science, but clearly our topic specific coverage is limited. The checking kernel can be integrated to different word processing applications or it can be used as a separate application. We have integrated it to GNU-EMACS text editor for use on LaTeX documents. In this form, the program is available for use within the university and around a number of sites on Internet. In our computing environment we monitor usage of our system and let the system send mail to the maintainers about our lexicon coverage baseed on user feedback.

This spelling checker has been implemented using the C programming language in a UNIX environment, on SUN SparcStations workstations at Bilkent University. Extensive test results (see [21]) indicate that it can process at 1000-3000 words (roughly 2-6 pages) per second on these platforms. This is about 1000 times faster than a morphological analysis system based on PC-KIMMO – a general purpose two-level morphological analysis system – for processing the same structure of words [13].

5 CONCLUSIONS

In this paper, we have presented a morphological parser for an agglutinative language, Turkish, and its application to spelling checking of this language.

Parsing agglutinative word structures necessitates some phonological and morphological analyses, presenting special difficulties in the development of parsers for such languages, not encountered in parsers for other languages. As a result, the number of parsers developed for agglutinative languages, and particularly for Turkish, is quite limited, and they all have certain shortcomings. We have solved most of the problems encountered in the previous parsers by making a detailed and careful research on Turkish word formation rules and their exceptions. The results of our research are given in Appendix A. These results may hopefully be helpful for future researchers on Turkish linguistics. We see that even though it is claimed that Turkish word formation rules are well-defined and that Turkish is a very regular language, as used today it shows many irregularities that cause the problem of parsing this language to become a hard and very interesting problem.

Many grammar books have been referred to compile Turkish word formation rules. In those books, after each rule is defined, usually it is reminded that there may occur some exceptions to that rule in some conditions, but mostly those conditions are not well-defined. For example, in all Turkish grammar books, it is stated that "When a Turkish word ending with one of the consonants P, Ç, T, K receives a suffix beginning with a consonant, that final consonant is softened, but there are some such words whose final consonant does not change." However, none of the books says what the common property of those words which do not obey to that rule is, because most probably it is not known yet. In order to include that rule correctly in the parser, all words having the indicated property have been examined, the list of the irregular ones have been obtained, and special checks have been done to catch those irregularities.

Some of the irregularities encountered in the Turkish language are even not mentioned in any of the grammar books. For example, although in some (but not all) of the grammar books we can see the rule "The verbal roots DE (say) and YE (eat) changes as Dİ and Yİ respectively when they receive a suffix beginning with the consonant Y", it is mentioned nowhere that the root DE does not always obey to this rule. For instance, it does not change when it receives the suffix –[Y]{I}P, i.e., the resulting word is not DİYİP, as said in the rule, but DEYİP. In order to include that rule correctly, all the suffixes beginning with Y have been examined, those which do not cause DE to change have been somehow decided, and

they have been handled specially.

In order to obtain reliable results from the spelling checker, all of the known rules and their exceptions have been implemented, but we have missed some rules. For example, it intuitively seems as if that the interrogative form of a verb in optative mood is not valid for some persons (e.g., GELESİN Mİ?), but that rule is not included in our rules since we have not been able to see it stated in any of the grammar books. Hence, later it may be necessary to make minor modifications in our grammar rules.

Some misspellings caused by affixing certain suffixes to some roots, which in fact can not receive them, can not be detected by the spelling checker yet. The reason is that, in the current implementation, all of the roots outside the verbal ones are marked as nominal roots, and they are treated as if they can receive all the inflexional suffixes which can be affixed to nominal roots. However, this is not always true because some of those roots can not receive all of those suffixes. For example, the root HEP (all) does not take the first person singular suffix –[{I}]M although it takes the plural one,[14] i.e., HEPİMİZ (all of us) is correct but HEPİM is not, will the checker can not detect it. To solve this problem, the lexicon used must be refined very carefully and the root classes must be determined based on usage and linguitics information. Obviously, this is a very difficult and time consuming job which requires a good knowledge on Turkish linguistics.

References

[1] Adalı, O., "Türkiye Türkçesinde biçimbirimler (Morphemes in Turkish used in Turkey)", TDK, Ankara, 1979.

[2] Banguoğlu, T., "Türkçenin grameri (Grammar of Turkish)", TDK, Ankara, 1986.

[3] Brodda, B., Karlsson, F., "An experiment with morphological analysis of Finnish", Papers from the Institute of Linguistics, University of Stockholm, Publication 40, Stockholm, 1980.

[4] Can, K., "Yabancılar için Türkçe-İngilizce açıklamalı Türkçe dersleri (Turkish Lessons for Foreigners with Turkish and English Explanations)", METU, Ankara, 1987.

[5] Hankamer, J., "Turkish generative morphology and morphological parsing", paper presented at Second International Conference on Turkish Linguistics, İstanbul, 1984.

[6] Hankamer, J., "Morphological parsing and the lexicon", in Lexical Representation and Process, edited by William Marslen-Wilson, MIT Press.

[7] Hatiboglu, V., "Türkçenin ekleri (Suffixes in Turkish)", TDK, Ankara, 1981.

[8] Kasper, R., Weber, D., "User's reference manual for the C's Quechua adaptation program", Occasional Publications in Academic Computing, Number 8, Summer Institude of Linguistic, Inc., 1982.

[9] Kasper, R., Weber, D., "Programmer's reference manual for the C's Quechua adaptation program", Occasional Publications in Academic Computing, Number 9, Summer Institude of Linguistic, Inc., 1982.

[10] Koskenniemi, K., "Two-level morphology", University of Helsinki, Department of General Linguistics, Publication No. 11, Helsinki, Finland, 1983.

[11] Köksal, A., "Automatic morphological analysis of Turkish", Ph.D. Thesis, Hacettepe University, Ankara, 1975.

[12] Mason, T., Brown, D., "lex & yacc", edited by Dale Dougherty, O'Reilly & Associates, Inc., USA, May 1990.

[13] Oflazer, K., "Two-level Description of Turkish Morphology", In Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics, 1993.

[14] This rule is not published anywhere.

[14] Özel, S., "Türkiye Türkçesinde sözcük türetme ve bileştirme (Word derivation in Turkish used in Turkey)", TDK, Ankara, 1977.

[15] Packard, D., "Computer-assisted morphological analysis of Ancient Greek", Computational and Mathematical Linguistics: Proceedings of the International Conference on Computational Linguistics, Pisa Leo S. Olschki, Firenze, 343 – 355, 1973.

[16] Sagay, Z., "Sözcük çekimi (Word Generation)", Proceedings of Bilişim'78, Ankara, 1978.

[17] Sagay, Z., "A computer translation of English to Turkish", M.S. Thesis, METU, Ankara, 1981.

[18] Sagvall, A., "A system for automatic inflectional analysis implemented for Russian, Data Linguistica 8, Almquist and Wiksell, Stockholm, 1973.

[19] Schreiner, A. T., Friedman, Jr., H. J., "Introduction to compiler construction with UNIX", Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1985.

[20] Solak, A., Oflazer, K., "A finite state machine for Turkish syllable structure analysis", Proceedings of the Fifth International Symposium on Computer and Information Sciences, Vol. 2, Nevşehir, 1195 – 1202, 1990.

[21] Solak, A., Oflazer, K., "Implementation details and performance results of a spelling checker for Turkish", in Proceedings of the Sixth International Symposium on Computer and Information Sciences, October 1991, Side, Turkey.

[22] Solak, A., "Design and implementation of a spelling checker for Turkish", M.S. Thesis, Bilkent University, Ankara, 1991.

[23] Underhill, R., "Turkish", Studies in Turkish Linguistics, edited by Dan Isaac Slobin and Karl Zimmer, 7 – 21, 1986.

[24] "Yeni yazım kılavuzu (New Writing Guide)", Ninth Edition, TDK, Ankara, 1977.

[25] "Yeni yazım kılavuzu (New Writing Guide)", Eleventh Edition, TDK, Ankara, 1981.

A THE TURKISH LANGUAGE

Turkish is an agglutinative language that belongs to a group of languages known as Altaic languages. For an agglutinative language such as Turkish, the concept of word is much larger than the set of vocabulary items. Word structures can grow to be relatively long by addition of suffixes and sometimes contain an amount of semantic information equivalent to a complete sentence in another language. A popular example of a complex Turkish word formation is

ÇEKOSLOVAKYALILAŞTIRAMADIKLARIMIZDANMIŞSINIZ

whose equivalent in English is "(it is speculated that) You have been one of those whom we could not convert to a Czechoslovakian." In this example, one word in Turkish corresponds to a full sentence in English. The word above has the following decomposition into suffixes:

ÇEKOSLOVAKYA/LI/LAŞ/TIR/AMA/DIK/LAR/IMIZ/DAN/MIŞ/SINIZ

Each suffix has a certain function and modifies the semantic information in the stem preceding it. In the previous example, the root morpheme ÇEKOSLOVAKYA is the name of the country *Czechoslovakia* and the suffix –LI converts the meaning into *person from [Czechoslovakia]*, while the following suffix –LAŞ makes a verb from the previous stem meaning *to become one of [the persons from [Czechoslovakia]]*.

Turkish spoken in different regions of Turkey also shows some differences. Spoken Turkish is divided into some *dialect*s each of which is spoken in a certain region of Turkey. One of these dialects, namely *İstanbul Türkçesi*, which is the Turkish spoken in İstanbul area, is chosen as the written language for Turkish. Written Turkish has a certain set of standard rules.

A.1 Morphophonemics

Turkish word formation uses a number of phonetic harmony rules. Vowels and consonants change in certain ways when a suffix is appended to a stem, so that such harmony constraints are not violated.

A.1.1 Vowel Harmony

The best known morphophonemic process in Turkish is the *vowel harmony*. Turkish has an eight-vowel system (A, E, I, İ, O, Ö, U, Ü), made up of all possible combinations of the distinctive features front/back, narrow/wide, and rounded/unrounded. Vowel harmony is a process by which the vowels in all syllables of a word except the first assimilate to the preceding vowel with respect to certain phonetic features. Vowel harmony in Turkish is a left-to-right process operating sequentially from syllable to syllable. The rules are [23]:

1. A non-initial vowel assimilates to the preceding vowel in frontness.
2. A non-initial narrow vowel assimilates to the preceding vowel in rounding.
3. A non-initial wide vowel must be unrounded; that is, **O** and **Ö** do not occur except in first syllables of the words.

Thus, while any of the eight vowels may occur in the first syllable of a word, the vowel of the following syllable is restricted to a choice of two. The features front/back and rounded/unrounded are entirely predictable, and only narrow/wide remains distinctive. Since most of the loanwords do not obey to the vowel harmony rules, there are some stems that are not subject to vowel harmony internally. However, nearly all suffixes are in harmony with the vowel on their left.

Except the progressive tense suffix (–iyor), there are no suffixes in which the wide vowels **O** and **Ö** appear. Therefore, in citing suffixes, if we use the cover symbol {A} for a wide vowel and {I} for a narrow vowel, their allophones.

{A} = A | E
{I} = I | İ | U | Ü.

Thus, the negation suffix can be shown as –M{A}, and the narrative past tense suffix as –M{I}Ş.

When a suffix is affixed to a stem, the first vowel in the suffix changes according to the last vowel of

the stem. Succeeding vowels in the suffix change according to the vowel preceding it. If we denote the preceding vowel (be it in the stem or in the suffix) by **V** then the two classes of vowels are resolved as follows:

{A}	=	A,	if **V** is	A	|	I	|	O	|	U
	=	E,	if **V** is	E	|	İ	|	Ö	|	Ü.

{I}	=	I,	if **V** is	A	|	I
	=	İ,	if **V** is	E	|	İ
	=	U,	if **V** is	O	|	U
	=	Ü,	if **V** is	Ö	|	Ü.

An allomorph is any of the variant forms of a morpheme. For example, the negation suffix –M{A} has two allomorphs, where narrative past tense suffix –M{I}Ş has four:

–M{A}	=	–MA	|	–ME				
–M{I}Ş	=	–MIŞ	|	–MİŞ	|	–MUŞ	|	–MÜŞ.

The allomorph of a suffix that is to be used is determined according to the phonemes of the stem it is affixed. For example, when the suffix –M{I}Ş is affixed to the root GÖR(MEK) ((to) see), the allomorph –MÜŞ is used, because as the vowel preceding the vowel {I} is **Ö** (V = Ö), {I} must resolve to an **Ü** ({I} = Ü):

GÖR + M{I}Ş → GÖRMÜŞ (he had seen).

There are also some non-harmonic suffixes, such as –KEN and –{I}YOR, which are exceptions to harmonic conditioning from the vowel on their left: OKURKEN (while reading), GELİYOR (he is coming).

Because of their different phonetic structures, some loanwords do not obey the vowel harmony rules during agglutination. For example:

ALKOL (alcohol) + L{I} → not ALKOLLU but ALKOLLÜ (containing alcohol).

When certain suffixes beginning with a consonant are affixed to the stems ending with a consonant, a narrow vowel is inserted between them We will show such vowels as [{I}].) This vowel is also determined similarly as explained before. For example the first person plural possessive suffix –[{I}]M{I}Z has eight different allomorphs:

–[{I}]M{I}Z	=	–IMIZ	|	–İMİZ	|	–UMUZ	|	–ÜMÜZ		
	=	–MIZ	|	–MİZ	|	–MUZ	|	–MÜZ.		

When this suffix is affixed to the root KAPI (door), it takes the form –MIZ. But when it is affixed to the root OKUL (school), the allomorph –UMUZ is used.

A.1.2 Consonant Harmony

Another basic aspect of Turkish phonology is *consonant harmony*. In one respect, consonants in Turkish may be divided into two groups as *voiceless* (Ç, F, T, H, S, K, P, Ş) and *voiced* consonants (B, C, D, G, Ğ, J, L, M, N, R, V, Y, Z). Most of the consonant harmony rules listed below are based on this classification [4, 11]:

1. Turkish words mostly end with a voiceless consonant; especially, the voiced consonants B, C, D, or G are rarely found as the final phonemes of the originally Turkish words. If there is one of these consonants at the end of a loanword, it changes to a corresponding voiceless sound of P, Ç, T, or K respectively: e.g., KİTAB → KİTAP (book), İLAC → İLAÇ (medicine).

2. In multi-syllabic words and in certain mono-syllabic roots, the final voiceless consonants P, Ç, T, K are mostly (not always) softened (i.e., it changes to B, C, D, or Ğ respectively) when a suffix beginning with a vowel is attached: e.g., AKORT (tune) → AKORDU (its tune) but AORT (aorta) → AORTU (his aorta).

3. In some suffixes beginning with one of the consonants C, D, or G, this initial consonant might change according to the last phoneme of the stem it follows. If we show these consonants as {C},

{D}, and {G}, their allophones will be:

{C}	=	C	|	Ç
{D}	=	D	|	T
{G}	=	G	|	K.

If the last phoneme of the stem to which one of such suffixes is attached is a voiceless consonant, the initial consonant of the suffix becomes voiceless (Ç, T, or K respectively), otherwise it remains as C, D, or G. Thus, the allomorphs of the definite past tense suffix –{D}{I} can be listed as:

–{D}{I}	=	–DI	|	–Dİ	|	–DU	|	–DÜ
	=	–TI	|	–Tİ	|	–TU	|	–TÜ.

When this suffix is affixed to the root GEL(MEK) ((to) come), i.e., GELDİ (he came), it takes the form –Dİ, and when it is affixed to the root KOŞ(MAK) ((to) run), the allomorph –TU is used, i.e., KOŞTU (he ran).

Furthermore some morphemes beginning with a vowel are affixed to the stems ending with a vowel with the insertion of one of the consonants N, S, Ş, or Y.[15] For example, the genitive suffix can be shown as –[N]{I}N, the third person singular possessive suffix as –[S]{I}, distributive numerical suffix as –[Ş]{A}R, and the acceleration suffix as –[Y]{I}VER. As an example, the suffix –[S]{I} takes the form –İ when it is affixed to the root EV (house), i.e., EVİ (his house), but the allomorph –SI is used when it is affixed to the root KAPI (door), i.e., KAPISI (his door).

There may be some exceptions to such morphophonemic rules. For instance, because of the former existence of an Arabic consonant not pronounced in Turkish, the consonant S is not inserted between some words ending with a vowel and the third person singular possessive suffix [11]:

SANAYİ (industry) + [S]{I} → not SANAYİSİ but SANAYİİ (industry of ...).

For some such words both forms are valid:

CAMİ (mosque) + [S]{I} → either CAMİSİ (mosque of) or CAMİİ.

A similar case happens when a case suffix comes immediately after some pronouns such as BU (this), ŞU (that), O (it), KENDİ (self), after the pronomial suffix –Kİ, or after the third person possessive suffixes –[S]{I} or –L{A}R{I}. In such cases an N is inserted in between:

BU	+	[Y]{I}	→	not	BUYU	but	BUNU
SENİNKİ	+	[Y]{A}	→	not	SENİNKİYE	but	SENİNKİNE

When all the rules above are considered, we reach the result that Turkish suffixes tend to have a highly protean nature. As an extreme example, the participial suffix –{D}{I}{K} has 16 allomorphs.

In the word SATTIĞIN ([the thing] that you sell) that suffix takes the form –TIĞ, because it follows the root SAT(MAK) ((to) sell) which ends with the voiceless consonant T (i.e., {D} = T) and whose last vowel is A (V = A → {I} = I), and it is followed by a suffix beginning with a vowel (i.e., {K} = Ğ).

A.1.3 Root Deformations

Normally Turkish roots are not flexed. However, there are some cases where some phonemes are changed by assimilation or various other deformations [11]. An exceptional case related to the flexion of roots is observed in personal pronouns. When the first and second singular personal pronouns BEN (I) and SEN (you) take the dative suffix, they change as:

BEN	+	[Y]{A}	→	not	BENE	but	BANA (to me)
SEN	+	[Y]{A}	→	not	SENE	but	SANA (to you).

When these two roots take the plural suffix, their structures completely change:

BEN	+	L{A}R	→	not	BENLER	but	BİZ (we)
SEN	+	L{A}R	→	not	SENLER	but	SİZ (you).

[15] We will show such consonants as [N], [S], [Ş], and [Y] respectively.

These are individual cases and can be treated as exceptions.

A more systematic change occurs when the suffix –[{I}]YOR comes after the verbs ending with the wide vowel {A}. In such cases, the wide vowel at the end of the stem is narrowed:

KAPA + [{I}]YOR → not KAPAYOR but KAPIYOR.
As an exceptional case, when not only the suffix –[{I}]YOR but also any of the suffixes beginning with the consonant Y is affixed to the roots DE(MEK) ((to) say) or YE(MEK) ((to) eat), they change as Dİ and Yİ respectively:

DE + [{I}]YOR → not DEYOR but DİYOR
DE + [Y]{A}N → not DEYEN but DİYEN [16]
YE + [Y]{I}P → not YEYİP but YİYİP.

One of the most important deformations in roots and stems occur as the result of the second consonant harmony rule. This rule says that when some words ending with one of the voiceless consonants P, Ç, T, K take a suffix beginning with a vowel, that consonant changes into B, C, D, or Ğ respectively:

DÖRT + {I}N{I}Z → not DÖRTÜNÜZ but DÖRDÜNÜZ
TABAK + [{I}]M → not TABAKIM but TABAĞIM.

If an N precedes a final K, the consonant K either stays as it is or it changes into a G:

TANK + [Y]{A} → TANKA
RENK + [Y]{A} → not RENKE but RENGE.

A similar change occurs when a suffix beginning with a vowel is affixed to a word ending with –LOG. In such a case, the final G changes into Ğ:

PSİKOLOG + [Y]{A} → not PSİKOLOGA but PSİKOLOĞA.

Another root deformation occurs as a vowel ellipsis. When a suffix beginning with a vowel comes after some nouns, generally designating parts of the human body, which has a vowel {I} in its last syllable, this vowel drops:

AĞIZ + [{I}]M{I}Z → not AĞIZIMIZ but AĞZIMIZ.
Similarly, when the passiveness suffix –{I}L is affixed to some verbs, whose last vowel is {I}, this vowel also drops:

AYIR + {I}L → not AYIRIL but AYRIL.
When a noun which has to face with vowel ellipsis receives the first person singular or plural suffixes, i.e., –[Y]{I}M or –[Y]{I}Z, although these suffixes begin with vowel, the last vowel of the root does not drop:

OĞUL + [Y]{I}Z → not OĞLUZ but OĞULUZ.

When a suffix beginning with a vowel is affixed to some originally Arabic roots ending with a consonant, or when such a root is combined with another word beginning with a vowel, the final consonant of the root is duplicated:

HAK + [{I}]M → not HAKIM but HAKKIM
ZAN + ETMEK → not ZANETMEK but ZANNETMEK.

When the plural suffix –L{A}R is affixed to certain compound words a deformation occurs. This suffix, coming before the possessive suffix at the end of the stem, forms a 'mid'fixing:

GÖZYAŞI + L{A}R → not GÖZYAŞILAR but GÖZYAŞLARI.

A.2 Morphology

Turkish roots can be classified into two main classes: *nominal* and *verbal*. The suffixes that can be received by either of these groups are different, i.e., a suffix which can be affixed to a nominal root can not be affixed to a verbal root with the same semantic function. There are also some roots which can

[16] The verb DEMEK sometimes shows exception to this exception either. For example:
DE + [Y]{I}P →not DİYİP but DEYİP.

nominal root	plural suffix	possessive suffix	case suffix	relative suffix

plural suffix	–L{A}R	
possessive suffixes	–[{I}]M	–[{I}]M{I}Z
	–[{I}]N	–[{I}]N{I}Z
	–[S]{I}	–L{A}R{I}
case suffixes	**internal**	**external**
	–[Y]{I}	–[Y]L{A}
	–[Y]{A}	–{C}{A}
	–{D}{A}	–L{I}
	–{D}{A}N	–S{I}Z
	–[N]{I}N	
relative suffix	–Kİ	

Figure 3: The nominal model

take all the suffixes either the nouns or the verbs can take (e.g., TAT (taste)), as well as others which never take suffixes (e.g., VE (and)).

Turkish suffixes can be classified as *derivational* and *inflexional*. Derivational suffixes change the meaning and sometimes the class of the stems they are affixed to, while a conjugated verb or noun remains as such after the affixation. Inflexional suffixes can be affixed to all of the roots in the class that they belong to. On the other hand, the number of roots each derivational suffix can be affixed to differs.

Inflexional suffixes may be divided into two groups according to the root class that they can be affixed to, i.e., a *noun* paradigm and a *verb* paradigm.

A.2.1 Noun Paradigm

The elements of the noun paradigm, in order, can be shown as in Figure 3 [1, 16, 17, 23]. All of these elements (except the root) are optional.

The plural suffix –L{A}R is added directly to the nominal root before any other suffix or ending. In the plural forms of the pronouns BU, ŞU, O an N is inserted between the word and the suffix.

Possesive pronouns (in English: my, your, his/her/its, our, your, their) are represented by suffixes in Turkish: e.g., EVİM (my house), ARABAN (your car). If the possessed noun is plural, possessive suffixes come after the plural suffix: e.g., EVLERİM (my houses), ARABALARIN (your cars). When the third person plural possessive suffix –L{A}R{I} comes after a plural noun, two L{A}R's combine and one of them drops:

EV + L{A}R + L{A}R{I} → not EVLERLERİ but EVLERİ.

Certain compound nouns have the third person singular possessive suffix already in their structure: e.g., ATEŞBÖCEĞİ (fire-fly), SAFRAKESESİ (gall bladder). Such words receive the possessive suffixes after removing the possessive suffix which is already in their structure:

ATEŞBÖCEĞİ + [S]{I} → not ATEŞBÖCEĞİSİ but ATEŞBÖCEĞİ

The nominal roots SU (water) and NE[17] (what) create some irregular cases when they receive possessive suffixes [1]:

[17] The regular forms for the root NE are also valid: NEM, NEN, NESİ, NEMİZ, NENİZ, NELERİ.

SUYUM (not SUM) NEYİM

Case suffixes can be grouped in two classes as *internal* and *external* case suffixes. Internal case suffixes are more frequently used than the external ones. They are named as follows:

–[Y]{I}	:	accusative	–{D}{A}	:	locative	–[N]{I}N	:	genitive
–[Y]{A}	:	dative	–{D}{A}N	:	ablative			

Declensions of pronouns have some irregular forms. In the dative cases of BEN and SEN, the front vowels become back (see page 20). In the genitive cases of BEN and BİZ, –İM is used instead of the regular form –İN:

BEN + [N]{I}N → not BENİN but BENİM (my)

Additionally, as mentioned on page 20, when a case suffix is attached to certain nouns an N is put in before the case suffix. Among such nouns we should add the portmanteau words having the characteristics mentioned above:

ATEŞBÖCEĞİ + [Y]{A} → not ATEŞBÖCEĞİYE but ATEŞBÖCEĞİNE

The relative suffix –Kİ may be added only to genitive or locative suffixes: e.g., KAPININKİ (the door's), BİZDEKİ ([the one] which is in our [hand, home]). It is possible to affix the relative suffix directly to a nominal root which indicates a a time or a place e.g., DEMINKİ (of a while ago), YARINKİ (tomorrow's)), or KARŞIKİ ([the one] on the opposite side), AŞAĞIKİ (the lower one). The number of such roots are quite limited. A noun stem that received the relative suffix may take the plural suffix and any case ending: e.g., BURADAKİLER (those who are here), BURADAKİLERLE (with those who are here). In its singular form an N is put between –Kİ and the case-ending: e.g. BURADAKİNDEN (from the one who is here).

A.2.2 Verb Paradigm

The verb paradigm is more complex than the noun paradigm. Its elements, in order, are shown in Figure 4. Among these elements, the obligatory ones are the root, the main tense suffix, and the person suffix.

There are four voices of verbs in Turkish: reflexive, reciprocal, causative, and passive. Combination of these suffixes are possible, but they must appear in the indicated order, and the reflexive and reciprocal are mutually exclusive: e.g. GÖRMEK (to see) → GÖRÜŞMEK (to see each other) → GÖRÜŞTÜRMEK (to cause to see each other) → GÖRÜŞTÜRÜLMEK (to be caused to see each other).

Neither the reflexive nor the reciprocal are productive roots; thus, they can be considered as derivational suffixes: DÖVMEK (to beat) → DÖVÜNMEK (to beat oneself), but KOŞMAK → KOŞUNMAK is invalid. ANLAMAK (to understand) → ANLAŞMAK (to understand one another), but not OKUMAK (to read) → OKUŞMAK.

The causative voice of verbs takes various forms. The set of rules applied to determine which allomorph is to be chosen for a given verb can be found in Solak [22]. The causative verb suffixes can be used repeatedly:

KAPA (close (it)) → KAPAT (cause it to become closed) →
KAPATTIR (have someone close it) → KAPATTIRT (have someone have someone close it).

The passive voice verb suffix also takes different forms. The allomorph to be chosen for a given verb is determined by the set of rules [11, 22]. The passive and reflexive forms of some verbs have the same structure, but they differ in their meanings. For example, the verb YIKANMAK is in passive voice in the sentence *Bulaşık yıkandı.* (The dishes were washed.), where it is in reflexive voice in the sentence *Ali yıkandı.* (Ali washed himself.).

There are two suffixes which give a verb negative sense: –M{A} (not) and –[Y]{A}M{A} (can/may not). The suffix –[Y]{A}M{A} is used to express impossibility: SÖYLEMEM (I don't say), SÖYLEYEMEM (I can't say).

Compound verb suffixes can be affixed to verbs to add them certain additional semantics. Among

verbal root	voice suffixes	negation suffix	compound verb s.	main tense s.	question suffix	second tense s.	person suffix

voice suffixes	**reflexive**	**reciprocal**	**causative**	**passive**
	–[{I}]N	–[{I}]Ş	–{D}{I}R	–{I}L
			–{I}T	–{I}N
			–Ɪ	–N
			–{I}R	
			–{A}R	
negation suffixes	–M{A}	–[Y]{A}M{A}		
compound verb suffixes	–[Y]{A}BİL	–[Y]{A}YAZ		
	–[Y]{A}DUR	–[Y]{A}KAL		
	–[Y]{I}VER	–[Y]{A}KOY		
	–[Y]{A}GEL	–[Y]{A}GÖR		
main tense suffixes	–{D}{I}	–S{A}		
	–M{I}Ş	–[Y]{A}		
	–[Y]{A}C{A}{K}	–M{A}L{I}		
	–[{I}]R	–Φ		
	–{A}R			
	–[{I}]YOR			
	–M{A}KT{A}			
question suffix	–M{I}			
second tense suffixes	–[Y]{D}{I}	–[Y]S{A}		
	–[Y]M{I}Ş			
person suffixes	–M	–[Y]{I}M	–[Y]{I}N	
	–N	–S{I}N	–[Y]{I}N{I}Z	
	–Φ	–[Y]{I}Z	–S{I}NL{A}R	
	–K	–S{I}N{I}Z		
	–N{I}Z	–L{I}M		
	–L{A}R			

Figure 4: The verbal model

them the potentiality and possibility suffix –[Y]{A}BİL, and the acceleration suffix –[Y]{I}VER are the most frequently used ones. More than one compound verb suffix may be added to a verb: e.g., SÖYLEYİVEREBİLİR MİSİN? (Could you please say?).

Main tense suffix is one of the obligatory suffixes for the verbs. There are nine tenses: definite past (–{D}{I}), narrative past (–M{I}Ş), future (–[Y]{A}CA{K}), aorist (–{I}R, –{A}R, –R), progressive (–{I}YOR, –M{A}KT{A}), conditional (–S{A}), optative (–[Y]{A}), necessitative (–M{A}L{I}), and imperative (–Φ). The last four are not tenses in the strict sense of the term, but their place in the verb model is the same as main tense suffixes.

As causative and passive voice suffixes, the aorist suffix also changes according to some specific rules [22]. In the negative form of a verb which is in present tense the aorist suffix is not used. The first singular and plural person suffixes are directly affixed to the negation suffix, while the other person suffixes are affixed with the insertion of a Z in between: e.g., VERMEM (I don't give), but VERMEZSİN (you don't give).

The progressive tense suffix –[{I}]YOR causes a deformation on some stems it is affixed to (see page 21). The same deformation occurs in the negation suffix when it is followed by the suffix –[{I}]YOR:

SEV + M{A} + –[{I}]YOR → not SEVMEYOR but SEVMİYOR.

The suffix –M{A}KT{A} can also be considered as a progressive tense suffix since it is used to indicate that an action continues in the present time.

There is no special suffix for imperative in Turkish. Whether a verb is in imperative form is understood via the person suffix. Every verb stem can be considered as in the second person singular imperative form (for positive orders positive stems, for negative orders negative ones): e.g., GEL! (Come!), KAPATMA! (Don't close!).

The question suffix –M{I} is written separate from the word it follows; but it is subject to vowel harmony. Its place within the verb is not consistent; it may appear after the main tense suffix, or after the person suffix, depending on the tense of the verb. It comes after the person suffix if the tense suffix is definite past, conditional, or optative: e.g., GELDİN Mİ? (Did you come?), GELSEM Mİ? (Should I come?), GELSİN Mİ? (Do you want him to come?). For the remaining tenses, the place of the question suffix is between the main tense suffix and the person suffix: e.g., GELİR MİYİZ? (Do we come?), GELECEK MİSİN? (Will you come?). No matter in which tense the verb is, the question suffix comes after the third person plural suffix: GELMELİLER Mİ? (Must they come?), GELİYORLAR MI? (Are they coming?).

A second tense information be added to a verb through the second tense suffixes. These suffixes are formed by removing the İ from the definite past, narrative past, and conditional forms of the verb İMEK, i.e., İDİ, İMİŞ, İSE: e.g., GELİYORDUM (I was coming), GELİRMİŞSİN ([I am told that] you come), GELECEKSEK (if we will come). When these forms are used as independent words, without being subject to the vowel harmony, they play the same role as the second tense suffixes: i.e., GELİYOR İDİM, GELİR İMİŞSİN, GELECEK İSEK. The second tense suffixes are affixed to verb stems ending with a vowel with the insertion of a Y in between: e.g., GELSEYDİ (if he came), GELEYMİŞ (I wish he had come), GELMELİYSE (if he must come).

None of the second tense suffixes can be used with the imperative suffix. Additionally, the narrative second tense suffix can not be used with definite past tense suffix, and the conditional second tense suffix can not come after the optative and the conditional tense suffixes: i.e., OKUYDU, OKU**DU**YMUŞ, OKU**SA**YSA are not valid.

The last obligatory suffix for verbs is the person suffix. Different suffixes are used to represent the first, second, and third singular, and plural persons. They also show differences depending on the main or second tense suffix they are affixed to. For example, in the following conjugations of the verb GEL(MEK) all the underlined suffixes are the first person plural suffix: GELDİK GELMELİYİZ GELELİM See Solak [22] for detailed information on conjugation of person suffixes.

A.2.3 Verbal Nouns

In Turkish, sentences can be classified as *verb sentences* and *noun sentences.* In verb sentences, there is an action, and this action is represented by a verb within the sentence: e.g., *Okula gittim.* (I went to the school.). On the other hand, in a noun sentence there is no explicit verb: e.g., *Öğrenciyim.* (I am a student). The noun sentences of Turkish correspond to the sentences formed by the verb *to be* in English. In Turkish, instead of using an extra verb in such sentences, some suffixes which play the role of the verb *to be* in English are added to the subject of the sentence. These suffixes can be shown as in Figure 5. The only obligatory suffix in this paradigm is the person suffix.

Negation concept shows differences in noun and verb sentences. In a verb sentence, it is obtained by adding a negation suffix to the verb of the sentence (see page 23): e.g., *Okula gitmedim.* (I didn't go to the school.). There is no such a suffix for the verbal noun of a noun sentence. Instead, the word DEĞİL is used for this purpose: *Öğrenci değilim.* (I am not a student).

As for the verb sentences, interrogative noun sentences are formed by adding the question suffix: e.g., *Okula gittim mi?* (Did I go to the school?), *Öğrenci miyim?* (Am I a student?).

Nominal sentences can be given tense information by using tense suffixes. As seen in Figure 5, there are three tense suffixes that can be added to a noun stem. They correspond to the second tense suffixes in the verb model. Thus, they are the definite past, narrative past, and conditional forms of the verb İMEK

nominal stem	question suffix	tense suffix	person suffix	probability suffix

question suffix –M{I}

tense suffixes –[Y]{D}{I}
–[Y]M{I}Ş
–[Y]S{A}

person suffixes

–M	–[Y]{I}M
–N	–S{I}N
–Φ	
–K	–[Y]{I}Z
–N{I}Z	–S{I}N{I}Z
–L{A}R	

probability suffix –{D}{I}R

Figure 5: The verbal noun model

(see page 25), and they may also be used as independent words, i.e., İDİ, İMİŞ, İSE: i.e., *Öğrenciydim.* and *Öğrenci idim.* can both be used. To express remaining tenses and modes apart from these three tenses in noun sentences, the infinitive OLMAK (to be) is used: e.g., *Öğrenci olacağım.* (I will be a student.), *Öğrenci olmalıyım.* (I must be a student.).

As in the verb model, here also, the third person plural suffix may come either before or after the tense suffix: e.g., both ZAYIFLARMIŞ ([I heard that] they were thin) and ZAYIFMIŞLAR are valid. When this suffix should come after a plural noun, one of the –L{A}R's drops: e.g., ÖĞRENCİLER (students) → ÖĞRENCİLERDİ (they were students), not ÖĞRENCİLERLERDİ, or ÖĞRENCİLERDİLER.

The suffix –{D}{I}R is not an obligatory suffix. It is usually not used in spoken language. In fact, it changes the meaning of the sentence a bit; it adds a probability, or sometimes a definiteness concept. For example the sentence *Arkadaşınız burada.* (Your friend is here.) means "I am sure that he is here", but *Arkadaşınız buradadır.* means "he must be here (perhaps, I think, probably)". However, it is certainly used in statements which express permanent validities: e.g., *Kedi bir hayvandır.* (Cat is an animal.). –{D}{I}R can also be used after the verbs in narrative past, progressive, or future tense, in necessitative mode, or in narrative form of one of these tenses: e.g., GELİYORDUR ([I am sure] he is coming), GELMELİYMİŞTİR (probably he must have come).

A.2.4 Participles

In Turkish, verb sentences can be transformed into a noun, adjective, or adverb clauses by adding certain suffixes to the verb of the sentence. (See [22] for a comprehensive list of these suffixes.(

During the transformation, the obligatory suffixes of the verb, i.e., main tense and person suffixes, are removed, and then the participles are affixed. Among the participles, only –M{A}D{A}N and –M{A}KS{I}Z{I}N can not be used with negation suffix since they include negation in themselves: OKUYACAĞINIZ (that you will read), GELMEYENLER (those who don't come), VERİLMEDEN (before/without being given).

–M{A}{K} forms the infinitive form of the Turkish verbs. The infinitive can be used as a noun, and may take any of the case endings but genitive. It never takes possessive suffixes: e.g., OKUMAĞA, OKUMAKTAN are valid, but OKUMAĞIN, OKUMAKLARI are not. Similarly, certain participles may be used as a nominal root, i.e., they may take all the suffixes that a nominal root can take: e.g.,

GELİŞİNİZE (to your coming), VERDİKLERİNDENDİ (it was one of those that you gave).

–[Y]KEN has a somewhat different usage than the other participles. Originally it is the –[Y]{A}N relative participle of the verb İMEK [4]. Like the other forms of this verb, it may be used as a suffix or as an independent word, i.e., İKEN. It is an invariable suffix, that is, it is not subject to the vowel harmony. It is affixed to a verb in the necessitative mode, or in any tense, except the definite past: e.g., OKU**R**KEN (while reading), OKU**MALI**YKEN (while he must read). It is not used with person suffixes, but it can follow the third person plural suffix –L{A}R: e.g., GELİR**LER**KEN (while they come). Second tense suffixes are not used with –[Y]KEN. It can also be affixed to a nominal stem causing a noun sentence transform into a noun clause: e.g. ÖĞRENCİYKEN, (when [the person] was a student), EVDELERKEN (when they are/were at home).

–C{A}S{I}N{A} shows some similarities with –[Y]KEN. It is affixed to certain tense bases, namely present, narrative past, and narrative of progressive and future: e.g., UÇ**AR**CASINA (just like flying), UÇ**UYORMUŞ**CASINA (as if (s/he) was sleeping), and it can also be affixed to nouns and adjectives: e.g., ÇOCUKCASINA (just like a child), ÇILGINCASINA (just like a crazy (person)).

A.2.5 Derivational Suffixes

Derivational suffixes are the suffixes which produce a new word having a different meaning than the word they are affixed to. As inflexional ones, derivational suffixes which can be added to nouns and verbs form different sets. Some derivational suffixes change the class of the word they are affixed to. Thus, they make nouns from verbs, or verbs from nouns. Others produce new nouns from nouns, or new verbs from verbs.

Some derivational suffixes may be received by all of the stems in the class that they belong to. The participles can be considered among them; i.e., they may be affixed to all verbs. Another group of the derivational suffixes can be attached to a great number, but not all, of the stems in their class. –{C}{I}, –L{A}Ş, –L{I}{K} are some examples to such suffixes. On the other hand, most of the derivational suffixes can be received by only a small number of stems.

There are hundreds of derivational suffixes in Turkish [1, 2, 7, 14, 11]. Some of them can only be added to some stems after they combine with some others. Such combinations should be examined as a single suffix. For example, the suffixes –L{A}N, –L{A}Ş, –L{A}T are the combinations of the suffix –L{A} with the suffixes –N, –[{I}]Ş, and –T, respectively [7]. Thus, although –L{A} can not be affixed alone to the nouns KUL (slave), YER (place), or KİR (dirt) to form verbs, the verbs KULLANMAK (to use), YERLEŞMEK (to settle down), and KİRLETMEK (to make dirty) are frequently used.

Examining all the derivational suffixes in Turkish necessitates a great effort and too much time. Even if we knew all the derivational suffixes, we should still examine all of the vocabulary of the language to determine which suffix can really be affixed to which roots.