Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 22, Sayı 2, 552-561, 2018

Süleyman Demirel University
Journal of Natural and Applied Sciences
Volume 22, Issue 2, 552-561, 2018

DOI: 10.19113/sdufbed.34562

Contex Free Grammer For Turkish

İlknur DÖNMEZ[*1], **Eşref ADALI**[2]

[1]Bilgi Üniversitesi, Doğa ve Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 34060, İstanbul
[2]İstanbul Teknik Üniversitesi, Bilgisayar ve Bilişim Fakültesi, Bilgisayar Mühendisliği Bölümü, 34467, İstanbul

(Alınış / Received: 13.11.2017, Kabul / Accepted: 24.05.2018, Online Yayınlanma / Published Online: 03.07.2018)

Keywords
Natural language processing,
Formal language theory,
Turkish context free grammar,
CFG rules for Turkish

Abstract: Formal Grammar which is introduced by Chomsky is one of the most important development in Natural Language Processing, a branch of Artificial Intelligence. The mathematical reresentation of languages can be possible using Formal Grammars. Almost all natural languages have word classes such as noun, adjective, verb. In addition to this one sentence consist of noun phrase and verb phrase. Noun phrase may consist of location, destination and source elements. Despite many similarities between the languages, there exist important dissimilarities in grammar rules of the languages belonging to different language families. In our study the most appropriate formal grammar representing Turkish language is investigated. Accuracy of the suggested grammars' rules is evaluated in two different corpus. This study is the enhanced version of "Turkish Context Free Grammar Rules with Case Suffix and Phrase Relation" that was presented on UBMK 2016 International Conference on Computer Science & Engineering [1]. Different from the first study, this study includes all word and sentence types of Turkish. Adjectives and prepositions are considered. The quoted sentences, incomplete sentences and question sentences are included. The genitive phrase structures including verbal word are included. In this study, the noun phrases are also defined in detail.

Türkçe İçin Bağlamdan Bağımsız Dil Temsili

Anahtar Kelimeler
Doğal dil işleme,
Biçimsel dil teorisi,
Türkçe bağlamdan bağımsız dil temsili,
Türkçe için BBD kuralları

Özet: Biçimsel gramerler, yapay zekanın bir dalı olan doğal dil işleme alanındaki en önemli gelişmelerden biridir. Dillerin matematiksel modellerle ifade edildiği bu gramer yapıları 1950'li yıllarda Chomsky tarafından ortaya atılmıştır. Dünyada kullanılan pek çok doğal dilde bulunma bildiren, ayrılma bildiren, kaynak bildiren öbekler ve isim, sıfat, zarf gibi ortak sözcük sınıfları kullanılmaktadır. Pek çok ortak özelliğe rağman özellikle farklı dil ailesinden olan dillerin dil bilgisi kuralları arasında önemli farklılıklar bulunmaktadır. Çalışmamızda Türkçeyi temsil eden en uygun biçimsel dil kuralları incelenmiş ve önerilen gramer kurallarının doğruluğu farklı derlemler üzerinde sınanmıştır. Bu çalışma 2016'da UBMK Bilgisayar Mühendisliği Bölüm Başkanları Toplantısında sunulan "Türkçe Hal Ekleri ve Öbekleri Kapsayan Bağlamdan Bağımsız Dil Temsili" çalışmamızın [1] genişletilmiş ve tüm sözcük türlerini içeren ve tüm cümle yapılarını kapsayan halidir. İlk çalışmadan farklı olarak edatlar ve sıfatlar sözcük tipleri ve içinde isim ve fiil barındıran tamlama yapıları dikkate alınmıştır. Bu çalışmada isim öbeği kavramı daha detaylı bir şekilde incelenip tanımlanmıştır. Alıntılanmış ve eksiltili cümle tipleri ve soru cümleleri çalışmaya dahil edilmiştir.

1. Introduction

The scientific studies about languages start at 1900's. The answer of what is natural language [2], [3], what are the features of natural language [4], [5], [6], can natural language be represented mathematically [15], [8], can we create a universal language [9] questions are searched. Formal language theory is suggested by Noam Chomsky in the 1950's [10], [11] and many other scientist studied grammatical representation of languages [12],[13].

English is represented by Context free grammar (CFG) which is a type of formal grammar. Different CFG grammar rules are determined in different studies for English [14], [15]. The main source for "CFG for English" is "An Introduction to Natural Language Processing" book of Daniel Jurafsky and James H. Mart. In this book "CFG for English" is a chapter of the book [16]. There are also studies related with CFG for English [17], [18]. English CFG does not need so much rules for suffixes relative to

* *Corresponding author: buyukkuscu@itu.edu.tr*

Turkish which is agglutinative language. Turkish can be represented by CFG. In this study, the most appropriate Context Free Grammar and rules are searched for Turkish.

Z. Güngördü and C. Demir are studied Turkish syntactic structure (Parsing Turkish using the lexical functional grammar(LFG) formalism) in 1993 [19]. In this study LFG grammar does not consider some of the verbal phrases (VP). T.Güngör ve S. Kuru used ATN for extracting Turkish suffixes [20]. This is extensive study that includes different type of phrases with verbal items. In this study some standard NP generation and phrase generation networks are defined. When we investigate in detail this networks are not enough for generating all kind of sentences. Because the study does not consider the recursion in the sentence and type transformation.

R.Çakıcı studied automatic induction of a CCG grammar for Turkish [21], [22]. This study uses machine learning techniques and supervised learning method so this study strongly related with used data. Our literature research shows that it is the only study using Combinatory Categorial Grammar (CCG) which is a kind of CFG. In this study the word type transformation and Turkish specific phrase types are not included.

In 2006, Ö. İstek studied on a link grammar for Turkish [23]. This study does not include multi-word expressions and punctuation symbols.

In 2007 E. İ. Ünkar parsed the Turkish sentences for text watermarking [24]. In this study word types are not detailed. In this study word "olası" is called modifier and it is not called modifier as a adverbial verb.

The importance of our study is consideration of general representation of Turkish sentences. Using all Turkish phrase structure, word transformations between word types with suffixes, and the recursive sentence structure inside phrases make our study different from other studies.

In this paper, section 1 includes introduction and previous work related to the concept. In section 2 Turkish specific features are defined. In section 3, Turkish specific CFG rules are introduced. Section 4 and section 5 are related with data, evaluation and results. CFG representation is done step by step. First the rules for simple sentence and noun phrases in simple sentence are defined. Then the rules for complex sentence and noun phrase in complex sentence are defined. At the end CFG rules for compound sentence, quoted sentence and incomplete sentences are defined.

2. Turkish Specific Features

Turkish has specific features:

1. Phrase structure with case suffixes
2. Free phrase order in the sentence
3. Compatibility between predicate and the other phrases in sentence
4. The sentence recursion using participles (verbal adjectives) and con-verbs (verbal adverbs) and gerund (verbal nouns)
5. Transformation between word types using suffixes.

2.1. Phrase structure with suffixes

Turkish sentences can be separated to their phrases via using specific case suffixes. Each phrase type has a role in the sentence. There are lots of studies related with Turkish phrase structure [25], [26], [27]. The phrases that are seen on Table 1 are used in this study.

Table 1. Phrases that are used in Turkish CFG

Phrases	Suffixes
P1: Subject phrase	-
P2: Nominative object phrase	-
P3: Accusative object phrase	-(y)ı, -(y)i, -(y)u, -(y)ü
P4: Destination Phrase	-(y)a, -(y)e
P5: Location Phrase	-da, -de, -ta, -te
P6: Source Phrase	-dan, -den, -tan, -ten
P7: Instrument Phrase	-la, -le
P8: Adverb Phrase	-
P9: Preposition Phrase	-
$V_{predicate}$: Predicate	-

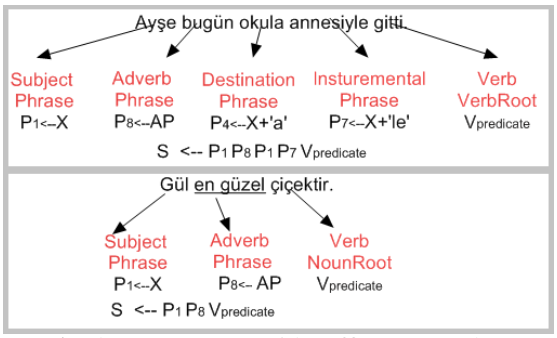

Figure 1. Phrase structure with suffixes example

In Figure 1, "Ayşe bugün okula annesiyle gitti." sentence is devided to its phrases. This sentence has a verb root in its predicate. "Gül en güzel çiçektir." sentence is separated to its phrases. This sentence has a noun root in its predicate.

2.2. Free phrase order

In Turkish, the phrase order is so flexible that the sentence S can be formed with all of the permutations of $P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_9$ phrases and $V_{predicate}$. $V_{predicate}$ is used at the end of the sentence in a regular sentence. The computational analysis of the syntax and interpretation of "free" word order in Turkish are studied by Hoffman in 1995 [28].
In Figure 2, all of the permutation of "Ayşe", "okula" and "annesiyle" words are used. We assumed this sentence regular so predicate is at the end. But even we change the order of predicate, the meaning does not change.

Ayşe okula annesiyle gitti. S <-- P_1 P_4 P_7 $V_{predicate}$
Ayşe annesiyle okula gitti. S <-- P_1 P_7 P_4 $V_{predicate}$
Okula annesiyle Ayşe gitti. S <-- P_4 P_7 P_1 $V_{predicate}$
Okula Ayşe annesiyle gitti. S <-- P_4 P_1 P_7 $V_{predicate}$
Annesiyle Ayşe okula gitti. S <-- P_7 P_1 P_4 $V_{predicate}$
Annesiyle okula Ayşe gitti.S <-- P_7 P_4 P_1 $V_{predicate}$
S <-- p $V_{predicate}$
p is permutation of all phrases (P_i)

Figure 2. Free phrase order example

2.3. Compatibility between predicate and the other phrases

In Turkish the predicate has time and model suffixes which should be compatible with adverb phrase. It has subject suffix which should be compatible with subject phrase.

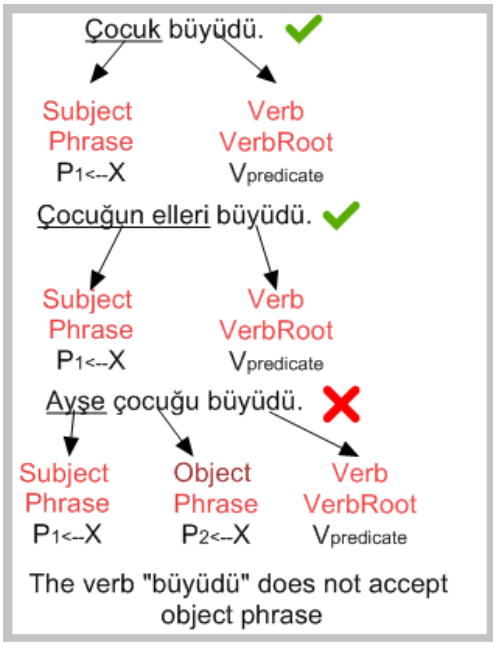

Figure 3. Predicate and other phrases example

As seen in Figure 3, predicate is also determinative for the phrases that the sentence may include. For example if the sentence predicate is "ol (be)", the sentence do not involve an object phrase and if the predicate is "oturmak (reside)", the sentence do not involve a source phrase.

2.4. Recursion in sentences

One of the common features of the languages is the recursive structures in sentence [29], [30]. Syntax of a phrase or a sentence is constructed from repeated rules. In this paper, we consider the recursion of sentence. A sentence should have judgment or should convey a statement. Sentences some times has inner statement and judgment inside. In Turkish recursion in the sentence is done with verbal forms.

When we assume p as one of the permutation of $P_1P_2P_3P_4P_5P_6P_7P_8P_9$ phrases, a simple sentence can be done by "p + V+suffix" rule. In "Ali okula geldi." sentence; "Ali" is subject, "okula" is dative phrase, "geldi" is constructed from a verb and past suffix. The recursion of "p+V+suffix" can be seen also in complex sentence. As seen in Figure 4 "Okula gelen Ayşe dün üzgündü." sentence has "p+V+suffix+X+p+V+suffix" form.

Okul --> Noun (X)
Okula --> Destination Phrase (P4<--X+'a')
Okula gel --> Sentence S<--P4 Vpred (S<--pV)
Okula gelen --> Adjective (X veya (pV+suffix))
Okula gelen Ayşe --> Subject Phrase (P1<--X veya (pV+suffix X))
Okula gelen Ayşe dün --> Subject & Adverb Phrase (P1P8 veya pV+suffix X P8)
Okula gelen Ayşe dün üzgündü --> S<--P1P8 Vpred (pV+suffix veya pV+suffix X pV+suffix)

Figure 4. Recursion example

In compound sentence, the sentence is generated by conjunction of complex and/or basic sentences. The compound sentence also has recursive structure $S \leftarrow S+C+S$ or $S \leftarrow$ p+V+suffix+C+p+V+suffix . Here, S is the sentence and C is the conjunction. "+" is concatenation operator.

2.5. Transformation structure with suffixes

Turkish is a agglutinative language so suffixes are deterministic features for phrase types; subject type; singularity or plurality; time and model type. There are lots of studies related with morphological structure of Turkish [31], [32]. Suffixes are also used to transform basic word types (noun, adjective and verb).

Güzel-->Noun
Güzelleşseydin--> Verb
Güzelleşen --> Adjective;
Güzelleştirdiğimiz -->Noun;
Güzelleştirmek-->Noun
Güzelleştirmekse --> Verb

Kaz-->Verb
Kazma--> Noun;
Kazmak--> Noun;
Kazan--> Adjective
Kazmalaş--> Verb;
Kazmalan-->Verb
Kazmalaşma-->Noun;
Kazmalanmak-->Noun
Kazmalanan-->Adjective
Kazmalanmaktır-->Verb

Figure 5. Transformation between noun and verb example

As seen in Figure 5 it is possible to make transformation between noun and adjective; noun and verb and verb and

adjective with suffixes. In Turkish rules related with original type can be applied to transformed type.
Different from the English, in Turkish the time phrases, location phrases, adverb phrase and destination phrases can be easily separated from Verb Phrase because of the case suffixes and phrases free order.

3. CFG For Turkish

In this study the aim is creating a context free grammar with its rules to handle all Turkish text and to allow deriving all possible text.

3.1. CFG for simple sentence

In formal grammar representation a language is represented by {P, N, T, S} so that P is generation rule, N is non terminal, T is terminal and S is starting symbol [11].

Table 2. CFG for simple sentence

Non Terminal
S ← p $V_{predicate}$ | p $V_{predicate}$ mi? (S: Simple sentence)
P_1 ← X | λ (P_1: Subject phrase)
P_2 ← X | λ (P_2: Nominative Object Phrase)
P_3 ← X+i | λ (P_3: Accusative Object Phrase)
P_4 ← X+e | X+a | λ (P_4: Destination Phrase)
P_5 ← X+de | X+da | λ (P_5: Location Phrase)
P_6 ← X+den | X+dan | λ (P_6: Source Phrase)
P_7 ← X+le | X+la | λ (P_7: Instrument Phrase)
P_8 ← akşama doğru | sabah | bugün| λ (Adverb Phrase)
P_9 ← ancak | değin etc. | λ (P_9:Prepositional Phrase)
X ← Ali | Gecenin yarısı|okul etc. (X: Noun Phrase)
$V_{predicate}$ ← koştu | geldi | yaptı (Predicate)

Turkish case markers and phrase relations are represented with formal grammar for simple sentence on Table 2 with determined rules.
As seen on the Table 2, the suffixes -i, -e, -de, -den, -le, noun phrase, noun, adverb phrase, adverb and predicate are terminals. S, $P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_9$, X and $V_{predicate}$ are non-terminals. λ denotes the empty string.
Each rule has a name part and an expansion of the name part. $P = \{P_i : 1 <= i <= 9\}$ is $\prod$, for the generation rule $p \in \prod$, S is denoted as $S \leftarrow pV_{predicate}$ or $S \leftarrow pV_{predicate}$ mi?.
As seen in Figure 6 "Ali bugün sabah okula geldi mi?" simple sentence can be represented with "S ← p $V_{predicate}$ mi?" rule and p includes P1, P8, P4.

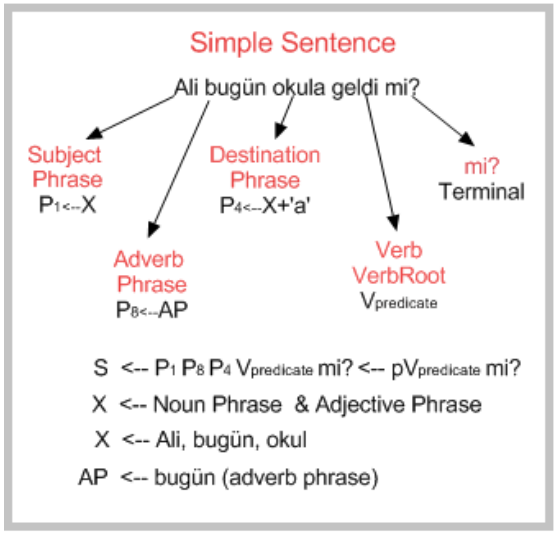

Figure 6. Simple sentence parsing example

3.2. CFG rules for noun phrase in simple sentence

All kinds of Turkish noun phrases for simple sentence can be generated using the rules on Table 3. Noun phrase in simple sentence will not include verbal words. The noun phrases in simple sentence are simple nouns, possessive constructions, adjectives and noun modifiers with nouns etc. They are constructed nouns, adjectives and pronouns, conjunctions and suffixes. X can be formed by combining more than one noun, conjunctions between noun phrases, using possessive suffixes.

Table 3. CFG for noun phrase (NP) in simple sentence

Non terminal
X ← X X (NP without suffix)
X ← X+in X+i | X+i (NP with suffix)
X ← X+in X+si | X+si (NP with suffix)
X ← X C X (Conjoining NP's)
C ← ve | veya | , | etc. (Conjunctions)
X ← benim X 'im' | X 'im' (NP with poss. suffix sg1)
X ← senin X 'in' | X 'in' (NP with poss. suffix sg2)
X ← onun X 'i' | X 'i' (NP with poss. suffix sg3)
X ← bizim X 'imiz' | X 'imiz' (NP with poss. suffix pl1)
X ← sizin X 'iniz' | X 'iniz' (NP with poss. suffix pl2)
X ← onların X 'ileri' | X 'ileri' (NP with poss. suffix pl3)
X ← X ki (Using ki in NP)
X ← Ayşe | ben | ak etc. (Nouns, adverbs and adjectives)

As seen in Figure 7, "senin robotun, demir kapı kolu ve camın pervazı" is a noun phrase. It may be used in "Senin robotun, demir kapı kolu ve camın pervazı bozuldu." sentence. This noun phrase is constructed from possessive constructions, adjectives and noun modifiers with nouns and conjunctions. In this table sg1 is used for singular 1st person, pl1 i used for plural 1st person and so on.

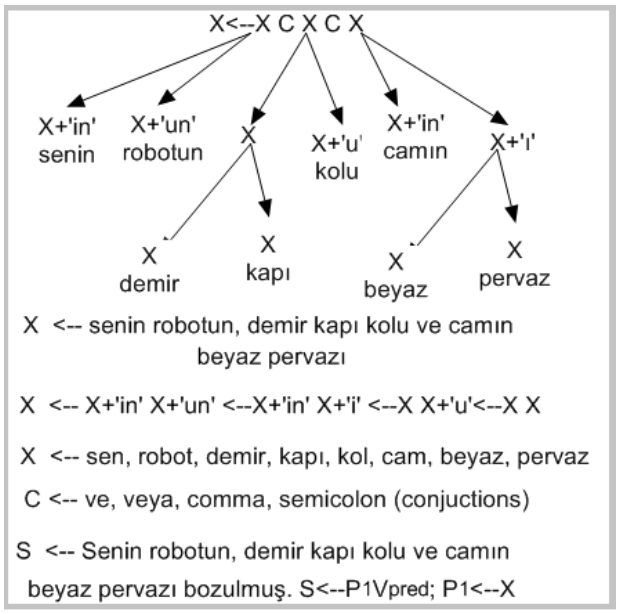

Figure 7. Noun phrase parsing example

3.3. CFG for complex sentence

Complex sentence includes gerunds (verbal noun), participles (verbal adjective) or/and con-verbs (verbal adverb) inside its phrases. CFG representation rules of simple and complex sentences are seen on Table 4.

Table 4. CFG rules for simple and complex sentence

Non-terminal
S ← p $V_{predicate}$ | p $V_{predicate}$ mi? (Simp/Comp sentence)
P1 ← X | λ (Subject phrase)
P2 ← X | λ (Nominative Object Phrase)
P3 ← X+i | λ (Accusative Object Phrase)
P4 ← X+e | X+a | λ (Destination Phrase)
P5 ← X+de | X+da | λ (Location Phrase)
P6 ← X+den | X+dan | λ (Source Phrase)
P7 ← X+le | X+la | λ (Instrument Phrase)
P_8 ← p$V_{converb}$ | hızlıca | sabah etc. | λ (Adverb Phrase)
P_9 ← X gibi|X'e göre| X için | λ (Prepositional Phrase)
$X_{withGerund}$ ← pV_{gerund} (X with verbal noun)
$X_{withParticiple}$ ← p$V_{participle}$X | p$V_{participle}$ (X with Partic)
V_{gerund} ← gelmek | gidiş | örme etc (Verbal nouns)
$V_{participle}$ ← gelen | öpülesi | yapacak etc (Verbal adj)
$V_{converb}$ ← koşarak | kayıp etc (Verbal adverb)
X ← $X_{withGerund}$ | $X_{withParticiple}$ |at |kız |ay etc. (X: NP)
$V_{predicate}$ ← gelecektin | Ayşedir etc. (Predicate)

Different from the simple sentence Noun phrases may include verbal gerunds (verbal nouns), participles (verbal adjectives) or/and con-verbs (verbal adverbs). For example "Okula gelen Ayşe bugün çok üzgündü. (Ayşe who came to school was unhappy today.)" is complex sentence. The subject phrase "Okula gelen Ayşe (Ayşe who came to school)" includes verbal adjective. λ denotes the empty string.

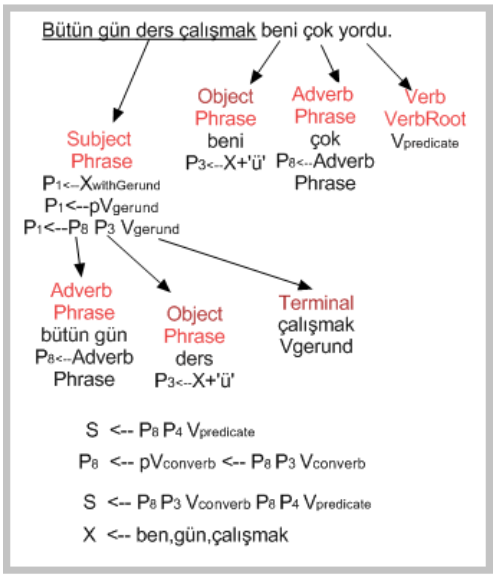

Figure 8. Parsing complex sentence including gerund

"Bütün gün ders çalışmak beni çok yordu." is complex sentence with gerund. This sentence can be parsed as seen in Figure 8.

Table 5. Generating NP's in complex sentences

Additional Rules to Table 3 and Table 4	Example
NP with Gerunds	
$X_{withGerund}$ ← V_{gerund}	almak, gidiş
$X_{withGerund}$ ← pV_{gerund} ;p isn't empty	eve gelmek
$X_{withGerund}$ ← V_{gerund}+'in' X+'i'	atışın sesi
$X_{withGerund}$ ← X+'in' V_{gerund}+'i'	dersin bitişi
$X_{withGerund}$ ← V_{gerund} +V_{gerund}	alış veriş
NP with Participle	
$X_{withParticiple}$ ← $V_{participle}$X	biten iş
$X_{withParticiple}$ ← $V_{participle}$	biten, olan
$X_{withParticiple}$ ← p$V_{participle}$X;p isn't empty	at süren kız
$X_{withParticiple}$ ← p$V_{participle}$;p isn't empty	bu işi yapan
$X_{withParticiple}$ ← $V_{participle}$+'in' +X+'i'	yapanın sonu
$X_{withParticiple}$ ← X+'in'+ $V_{participle}$+'i'	atın koşanı
$X_{withParticiple}$ ← $V_{participle}$ + $V_{participle}$	alan veren
$X_{withParticiple}$ ← $V_{participle}$ + V_{gerund}	süren oluş
X ← $X_{withGerund}$ | $X_{withParticiple}$	

As seen on the Table 5 $X_{withGerund}$ can be generated in different ways. Due to the adjective can be used instead of nouns in Turkish, X ← $X_{withGerund}$ and X ← $X_{withParticiple}$ rules are all valid. To have a general rule set, this rules should be added to the CFG rules for complex sentences and to the noun phrases rules for simple sentences.
"Acele kararlarla yönetilen şirket" is a noun phrase. It will be subject phrase if the sentence is "Acele kararlarla yönetilen şirket hata yapmaya mahkumdur.". As

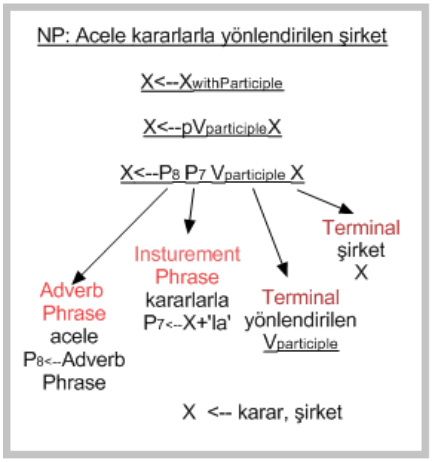

Figure 9. Noun phrase in complex sentence

seen in Figure 9 noun phrase X is generated from $X \leftarrow P_8P_7VparticipleX$.

Table 6. Con-verbs in complex sentences

Complex sentence with con-verbs	
$P_8 \leftarrow pV_{converb}$;p is empty	olarak, koşup
$P_8 \leftarrow pV_{converb}$;p is not empty	okula koşup
More than one $V_{converb}$	alıp verip

The verbal item in complex sentence may be in gerund, participle and con-verb (verbal adverbs) form. In Table 6, the example generation rules for complex sentence with converb are seen.

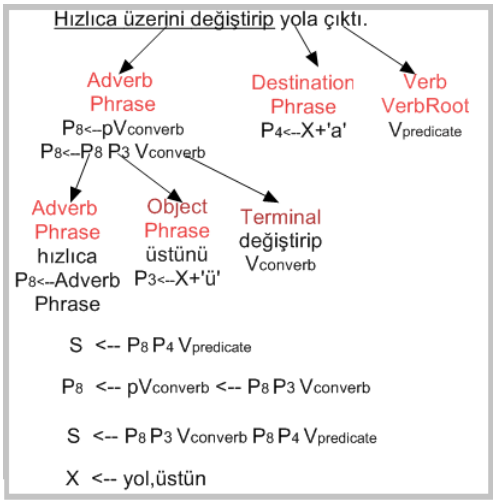

Figure 10. Complex sentence including converb parsing example

"Hızlıca üzerini değiştirip yola çıktı" sentence is a complex sentence with con-verb (verbal adverb). Adverb phrase P_8 include a verbal adverb. As seen in Figure 10 Adverb phrase P_8 can be generated from $P_8 \leftarrow P_8P_3Vconverb$ or $P_8 \leftarrow pVconverb$.

3.4. CFG for compound sentence

In compound sentences, there is equal emphasis on sentences which are connected by conjunctions. The rules on Table 4 contains generation rules for simple and compound sentence. When we add the rules which are seen on the Table 7 to the complex sentence rules on the Table 4, Turkish general context free grammar with phrases and suffixes can be maintained that contains simple, complex and compound sentence.

Table 7. CFG rules for compound sentence

Non-terminal
$S \leftarrow$ S C S | C S | S (Basic Complex and Compound Sent.)
C $\leftarrow$ ama | ve | veya | , | ; etc

If "S" denotes the complex or simple sentence; "$S \leftarrow$ S C S" and "$S \leftarrow$ C S C S" denotes the compound sentence. Here "C" symbol denote the conjunction like "ama, ile, ve, ya, ya da" or some special punctuations like ";" and ",". Generally conjunctions are used between the sentences. "Ali geldi, okula gitti ama hiç birşey söylemedi. (Ali came, went school but did not tell anything.)" is an example of compound sentence. In Turkish some sentence may start with conjunctions. "Ya buradan gidersin ya da ben giderim (either you go or I go)" is an example.

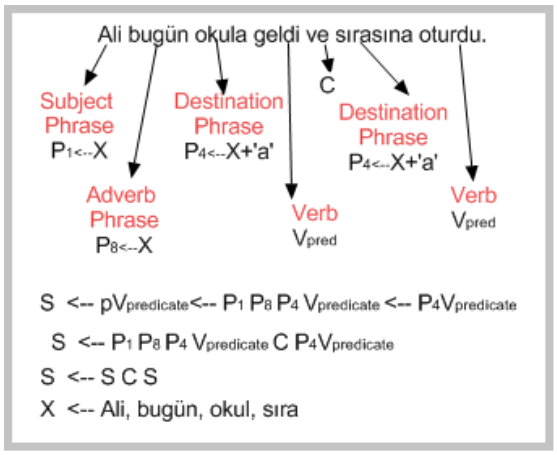

Figure 11. Compound sentence parsing example

"Ali bugün okula geldi ve sırasına oturdu" sentence is a compound sentence. It is formed from two different sentence "Ali bugün okula geldi" and "Sırasına oturdu" sentences with a conjunction "ve". As seen in Figure 11 compound sentence S can be generated from $S \leftarrow P_1P_8P_4VpredicateCP_4Vpredicate$ or $S \leftarrow pVpredicateCpVpredicate$.

3.5. Turkish CFG for sentence that include quotation

In Turkish the quoted speech should be processed like a Noun Phrase for Context Free Grammar. The main evidence for this hypothesis is the use of case suffixes after the quoted speech. We know that case suffixes are used for phrases after the noun phrase.

1. Example for nominative phrase (no suffix): “Ali ‘haydi buraya gel’ dedi. What did Ali said? The answer is "Lets come here". It is the object phrase of the sentence.

2. Example of accusative pharse ("-i" case suffix):"Biz okulda ilk 'Ali gel'i, 'Topu tut'u öğrendik

3. Example of source phrase ("-den" case suffix): 'Para ver'den başka cümle bilmiyor. .

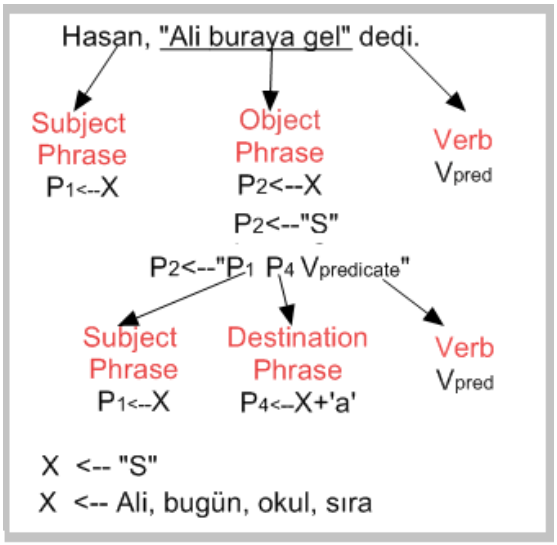

Figure 12. Quoted sentence parsing example

"Hasan, 'Ali buraya gel' dedi" sentence is a quoted sentence. It is formed from two different sentence. Even "Ali buraya gel" is a sentence, it is used instead of object phrase of the main sentence. As seen in Figure 12 quoted sentence S can be generated from $S \leftarrow P_1P_2V\,predicate$ and $P_2 \leftarrow$ "S" $P_2 \leftarrow P_1P_4V\,predicate$.

Table 8. CFG for sentence that include quotation

A rule for Noun Phrase
$X \leftarrow$ "S"

For the CFG representation of sentences that include quotation; the rule $X \leftarrow$ "S" in Table 8 should be added to the Table 5. Here the quotation mark outside the sentence is important.

3.6. Turkish CFG for incomplete sentence

In incomplete sentences the sentence does not have to contain a predicate. In dialogs incomplete sentence may be generated from noun phrase, subject phrase, object phrase, destination phrase, location phrase, source phrase, instrument phrase, adverb phrase or/and prepositional phrases. As seen in Figure 13 there are some incomplete sentence example. This incomplete sentence is generated from noun phrase and case suffix like a phrase.

Table 9. CFG for incomplete sentence

A rule for Sentence
$V\,predicate \leftarrow \lambda$

For the CFG representation of incomplete sentences, the rule "S $\leftarrow$ p" in Table 9 should be added to the Table 4

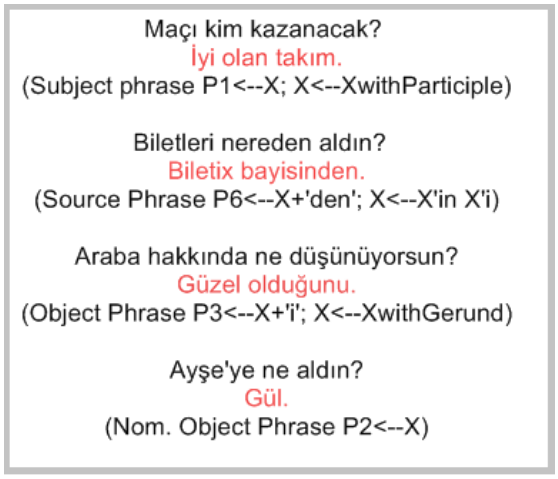

Figure 13. Incomplete sentence parsing example

or predicate may be empty string $V\,predicate \leftarrow \lambda$. If the sentences is accepted without predicate via this CFG, it means incomplete sentence will be accepted via this CFG. Because of CFG Rules with Verb Suffixes and Transitivity Control are mentioned in our first study and because of these issues are not core for Turkish sentence representation, these issues are not included in this paper.

3.7. Turkish CFG related with verb and noun type transformation

In Table 10 the verbs that has a noun root is seen. In Table 11 the nouns that has a verb root is seen. In these tables the transformation rules between noun and verb are listed.

Table 10. CFG rules for verb which has noun root

Non-terminal Definition
$X \leftarrow$ ağaç | kuşl ev | güzell yaz etc. (Noun root)
$V \leftarrow$ X+CTS | X+'len'CTS | X+leCTS | X+leşCTS
$V \leftarrow V$+CTS (Verb case, time and subject suffix)
X$\leftarrow V_{gerund} \leftarrow V + Suffix_{gerund}$ (Verbal nouns)
X$\leftarrow V_{participle} \leftarrow V + Suffix_{participle}$ (Verbal adj)
X$\leftarrow V_{converb} \leftarrow V + Suffix_{converb}$ (Verbal adverbs)
$V_{predicate} \leftarrow V$ | X T S (Predicate)

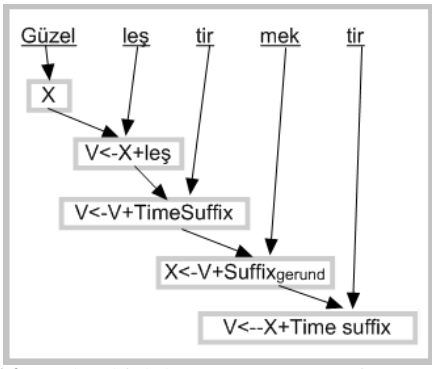

Figure 14. Verb which has noun root parsing example

As seen in Figure 14 "Güzelleştirmektir" verb has noun root "güzel". There are two times noun to verb and one

time verb to noun transformation in this example. The verb that has noun root can be generated directly with "len", "leş", "le" suffix and/or case, time and subject suffixes. "Evdir" verb has TMS suffixes to a noun "ev".

Table 11. CFG rules for noun which has verb root

Non-terminal Definition
V ← al, gel, git, ol, üşü, yap, boya, yarat, sat...
$V_{gerund} \leftarrow V + Suffix_{gerund}$ (Verbal nouns)
$V_{participle} \leftarrow V + Suffix_{participle}$ (Verbal adj)
X ← $X_{withGerund}$ ← pV_{gerund}
X ← $X_{withParticiple}$ ← p$V_{participle}$X | p$V_{participle}$
X ← X X (NP without suffix)
X ← X+in X+i | X+i (NP with suffix)
X ← X+in X+si | X+si (NP with suffix)
X ← X C X (Conjoining NP's)
C ← ve | veya | , | etc. (Conjunctions)
X ← benim X 'im' | X 'im' (NP with poss. suffix sg1)
X ← senin X 'in' | X 'in' (NP with poss. suffix sg2)
X ← onun X 'i' | X 'i' (NP with poss. suffix sg3)
X ← bizim X 'imiz' | X 'imiz' (NP with poss. suffix pl1)
X ← sizin X 'iniz' | X 'iniz' (NP with poss. suffix pl2)
X ← onların X 'ileri' | X 'ileri' (NP with poss. suffix pl3)
X ← X ki (Using ki in NP)

In Table 11, sg1 is used for singular 1st person, pl1 i used for plural 1st person and so on.
After we create a verbal adjective or verbal noun we can use it as a noun. And we can use it to create new noun phrases using the rules related with noun phrase generation. For example "gelmek (to come)" is gerund word and "zamanın gelmesi" can be generated from X ← X+in X+i rule. This noun phrase which include verbal item is shown in Figure 15.

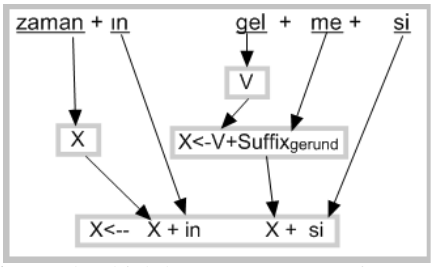

Figure 15. Verb which has noun root parsing example

The rules in Table 10 and the Table 11 are selected from the previous tables CFG rules to show the verb to noun and noun to verb type transformation.

4. Data

The data is taken from İTÜ NLP Machine Learning Corpus for the evaluation[34]. This corpus is generated for English to Turkish machine translation project. One million parallel sentence in English and Turkish is included in this corpus. For our evaluation we generate 3 datasets. For dataset-1 the random 1000 simple sentences, for dataset-2 1000 simple and complex sentences, and for dataset-3 1000 all type of sentences are taken from this corpus. There are 23 same sentence between dataset-1 and dataset-2 and there are 69 same sentence between dataset-2 and dataset-3. Because of the random selection of sentences, the common sentences will not effect the result. First dataset is used for the evaluation of CFG For Simple Sentence; 1000 simple and complex sentences are used for the evaluation of CFG for simple and complex Sentence; 1000 all kind of sentences are used for the other CFG's.
To compare and find the average result, Turkish National Corpus also has been used. This corpus is generated from 9 different area datasets. This dataset has 50.000.000 words [35]. 10.000 sentences is downloaded from this dataset and Datasets are grouped according to their sentence types similar with the İTÜ Corpus datasets. For dataset-1 the random 1000 simple sentences, for dataset-2 1000 simple and complex sentences, and for dataset-3 1000 all type of sentences are taken from this corpus. There are 43 same sentence between dataset-1 and dataset-2 and there are 78 same sentences between dataset-2 and dataset-3.

5. Evaluation and Results

In "the evaluation metric in generative grammar" study of John Goldsmith, he proposes two ways to evaluate the accuracy of a generative language [33]. First method is for understanding how appropriate the formal language is for the given language data and the second one is for understanding the formal language accuracy in dependent from the language data.
In first method to understand how appropriate the formal language is for the given language data we test each sentence if the sentence can be generated via using related rules in related CFG. We evaluate the score "True" for this sentence, if the rules allow to derive the sentence. If the sentence is not generated via using related rules in related CFG, we evaluate the score "False" for this sentence. The CFG accuracy is equal to total true scored sentences number divided by all sentences number.
In Table 12 the evaluated values are compared to our previous "Turkish Context Free Grammar Rules with Case Suffixes and Phrase Relation" study (Version-1) [1].

Table 12. Suggested CFG accuracy values compared with previous version

	Accuracy	
CFG Rules	Version-1	Version-2
Simple Sentence	94,6	96,4
Simp&Complex S.	78,3	85,2
All Sentences	73,7	81,9

When we searched the reason of increasing accuracy, we found that, preposition like "için, e göre, gibi" usage for prepositional phrase and the additional rules related with incomplete, quoted sentences caused this average accuracy difference. We can not compare CFG-IV and CFG-V with the previous version directly because previous version does not have rules for incomplete, quoted sentences. The Accuracy-V1 values are taken from the previous study.
As seen on Table 12, compared with the first study, including all word and sentence types; considering adjectives and

prepositions and involving the inverted sentences, incomplete sentences and the genitive case structures including verbal items cause a 8.2% increase in accuracy value of CFG for all sentence types.
The second method was for understanding the formal language accuracy in dependent from the language data. To approximate the second method and to decrease the error, the evaluation is done for independent language corpus, and the average accuracy value is taken into account.
As seen on the Table 13, the CFG rules are grouped according to their function. For example CFG-I is used for evaluation of simple sentences and CFG V is used for evaluation of all sentences including simple complex, compound, quoted and incomplete sentences.

Table 13. CFG and related rules

CFG Rules	Related Table that include rules
CFG-I	Simple sent. rules Table 2 and 3
CFG II	Simp and Complex sent. Table 3-4-5-6
CFG III	Simp,Complex,Compound Table 3-4-5-6-7
CFG IV	CFG III +Quoted Sent Table 3-4-5-6-7-8
CFG V	All sentence types Table 3-4-5-6-7-8-9

In Table 14 all the accuracy value of different CFG's related with included sentence type is seen. The accuracy values are calculated in İTÜ Machine Learning Corpus and Turkish National Corpus (TNC) and the average accuracy is calculated to decrease the error.

Table 14. Grammars accuracy in different corpus's

CFG Rules	İTÜ Corpus	TNC Corpus	Avarage
CFG-I	97.4	95.4	96.4
CFG II	86.2	84.2	85.2
CFG III	84.2	69	76.6
CFG IV	85.6	72.6	79.1
CFG V	86.5	77.3	81.9

As seen on the Table 14, there is a nearly %10 difference between the accuracy value of CFG-III, CFG-IV and CFG-V. The reason for this may be the more generic content of TNC corpus. It has some speech like quoted, incomplete and not regular sentences and phrases.
Finally the average accuracy values on different fields are calculated. 500 sentences are used for different field types. Here we used both İTÜ and TNC corpus. The result can be seen on Table 15. This grammar representation may have different accuracy value in different fields in other words the average accuracy value of the output will change according to field.
In the sentences related with academy, because of there is not much quoted sentences and incomplete sentences; the accuracy result for CFG II, CFG III and CFG IV become similar. The big difference between CFG II, CFG III and CFG IV is seen on "Story" field.

6. Discussion and Conclusion

As a conclusion Turkish is one of the most regular and rule based language in the world. So when we represent

Table 15. Suggested grammars accuracy in different fields

CFG Rules	Academic	Story	Twit	Noval	News
Simple S.	97.4	96.2	94.4	97.2	96.8
CFG II	86.2	83.3	82.3	87.0	88.1
CFG III	86.2	60.4	75.6	83.8	84.1
CFG IV	86.6	72.4	77.1	84.6	87
CFG V	86.6	82.8	87.6	84.7	87.1

Turkish sentences with CFG rules which have recursive structure and test this representation on two different corpus; the average accuracy value is found as 81.9 %. The suggested context free grammar rule for Turkish covers a large amount of Turkish corpus.
As it is known, in the recent days natural language understanding is one of the important topics. In order to understand semantic meaning of a sentence, we should separate the sentence to its meaningful parts and the functionality. Relationship between these parts should be known. The correctly parsed sentences are necessary in many fields of Natural Language Processing. With our study the sentence may be parsed to its phrases and basic word types, with its suffixes and transformation and relation of words can be provided.
The suggested CFG has general and ruled base method. Using the suggested CFG the sentence can be separated into its phrases. In Turkish understanding the phrases so important to understand the sentence. Who did the action? When did the action? Where did the action etc. Using the CFG rules the phrases can be also decomposed to its root words. We can understand complex sentences with their verbal adjective, gerund or con-verb. so we can understand the inner sentences. This CFG also recognizes question, compound, quoted and incomplete sentences. Noun phrase parsing rules can be use in Name Entity Recognition (NER) application. It is expected that suggested CFG for Turkish become a source for different area in NLP.

References

[1] Dönmez, İ., & Adalı, E. 2017. Türkçe Hal Ekleri ve Öbekleri Kapsayan Bağlamdan Bağımsız Dil Temsili Kuralları. Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, 10(1), 33-40.

[2] Chomsky, N. 1986. Knowledge of language: Its nature, origin, and use. Greenwood Publishing Group.

[3] Chomsky, N. 2002. On nature and language. Cambridge University Press.

[4] Hauser, M. D., Chomsky, N., & Fitch, W. T. 2002. The faculty of language: what is it, who has it, and how did it evolve?. science, 298(5598), 1569-1579.

[5] Pinker, S., & Jackendoff, R. 2005. The faculty of language: what's special about it?. Cognition, 95(2), 201-236.

[6] Chomsky, N. 2006. Language and mind. Cambridge University Press.

[7] Charniak, E. 1996. Statistical language learning. MIT press.

[8] Harris, Z. S. 1968. Mathematical structures of language.

[9] Montague, R. 1970. Universal grammar. Theoria, 36(3), 373-398.

[10] Chomsky, N. 1959. On certain formal properties of grammars. Information and control, 2(2), 137-167.

[11] Chomsky, N. 1963. Formal properties of grammars.

[12] Blackburn, P., & Gardent, C. 1995. A specification language for lexical functional grammars. ss. 39-44. Blackburn, P., & Gardent, C. 1995. In Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics. Morgan Kaufmann Publishers Inc.

[13] Everett, D., et all. 2005. Cultural constraints on grammar and cognition in Piraha: Another look at the design features of human language. Current anthropology, 46(4), 621-646.

[14] Lambek, J. 2014. From Rules of Grammar to Laws of Nature. Novinka.

[15] Charniak, E. 1997. Statistical parsing with a context-free grammar and word statistics. AAAI/IAAI, 2005(598-603), 18.

[16] Jurafsky, D. 2000. Speech and language processing: An introduction to natural language processing. Computational linguistics, and speech recognition.

[17] Murase, T. et all. 2001. Incremental CFG Parsing with Statistical Lexical Dependencies. In NLPRS pp. 351-358.

[18] Oates, T. et all. 2003. Leveraging Lexical Semantics to Infer Context-Free Grammars. In ECML Workshop on Learning Contex-Free Grammars pp. 65-76.

[19] Güngördü, Z., & Oflazer, K. 1995. Parsing Turkish using the lexical functional grammar formalism. Machine Translation, 10(4), 293-319.

[20] Güngör, T., & Kuru, S. 1993. Representation of Turkish morphology in ATN. In Proceedings of Second Symposium on Artificial Intelligence and Artificial Neural Networks (92-104)

[21] Cakici, R. 2005. Automatic induction of a CCG grammar for Turkish. In Proceedings of the ACL student research workshop Association for Computational Linguistics.(73-78).

[22] Cakici, R. 2009. A wide-coverage morphemic CCG lexicon for Turkish. In Parsing with Categorial Grammars Workshop ESSLLI 2009 Bordeaux, France Book of Abstracts.

[23] İstek, Ö. 2006. A link grammar for Turkish. Bilkent Universitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, Ankara.

[24] Meral, H. M., et all. 2007. Syntactic tools for text watermarking. In Security, Steganography, and Watermarking of Multimedia Contents IX.International Society for Optics and Photonics. Vol. 6505, p. 65050X

[25] Durrant, P. 2013. Formulaicity in an agglutinating language: The case of Turkish. Corpus Linguistics and Linguistic Theory, 9(1), 1-38.

[26] İşgüder-Şahin, G. G., & Adali, E. 2014. A pilot study on automatic inference rule discovery from Turkish text. In Application of Information and Communication Technologies (AICT), 2014 IEEE 8th International Conference on IEEE 1-5

[27] Dönmez, İ., & Adali, E. 2015. Extracting phrase-content pairs for Turkish sentences. In Application of Information and Communication Technologies (AICT), 2015 9th International Conference on IEEE 128-132.

[28] Hoffman, B. 1995. The computational analysis of the syntax and interpretation of" free" word order in Turkish. IRCS Technical Reports Series, 130.

[29] Pullum, G. K., & Scholz, B. C. 2010. Recursion and the infinitude claim. Recursion in human language, 104, 113-38.

[30] van der Hulst, H. 2010. Recursion and human language (Vol. 104). Walter de Gruyter.

[31] Arslan, E., & Orhan, U. 2016. Graph-based lemmatization of Turkish words by using morphological similarity. In INnovations in Intelligent SysTems and Applications (INISTA), 2016 International Symposium on IEEE (1-5).

[32] Kunduracı, A., & Göksel, A. 2016. Morphology: the base processor. In Mediterranean Morphology Meetings Vol.10, 88-97

[33] Goldsmith, J. 2011. The evaluation metric in generative grammar. 50th anniversary celebration for the MIT Department of Linguistics.

[34] Tantug, A. C. et all. 2014. The effect of parallel corpus quality vs size in English-to-Turkish SMT. In Sixth International Conference on Web services & Semantic Technology Chennai.

[35] Aksan, J. et all. 2012. Construction of the Turkish National Corpus (TNC). In LREC pp. 3223-3227.