SYNTAX

Prof. Dr. Eşref ADALI

Chapter – VI

E-mail : adali@itu.edu.tr

www.adalı.net or www.xn--adal-oza.net

What is Syntax

Syntax is the study of formal relationships between words in a sentence.

Prescriptive Grammar

In every language, the order of words in sentences is expected to have a rule and order. Sentences written or spoken according to the rules are found grammatically correct. However, it cannot be said that everything written and said is in accordance with the grammar rules. Grammar, which is intended to determine the rules of a language, is called prescriptive grammar. Prescriptive grammar dictates how we should write and speak. Prescriptive grammar requires people to write and speak correctly.

Descriptive Grammar

Everything written and said does not exactly comply with the rules. For this reason, the written and spoken language should also be examined and evaluated in terms of grammar. The grammar that serves this purpose is called descriptive grammar. Descriptive grammar does not look at whether what is written or said is true or false, but says whether it is regular or irregular.

Traditional and Modern Syntax

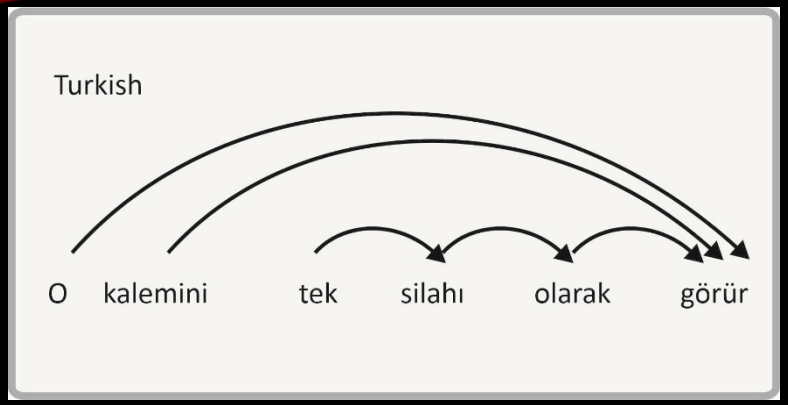
Traditional classes

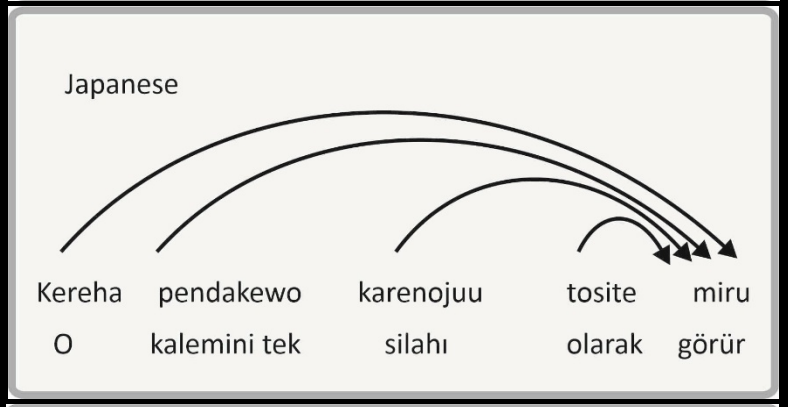
- Noun
- Verb
- Pronoun
- Adjective
- Adverb
- Preposition
- Conjunction

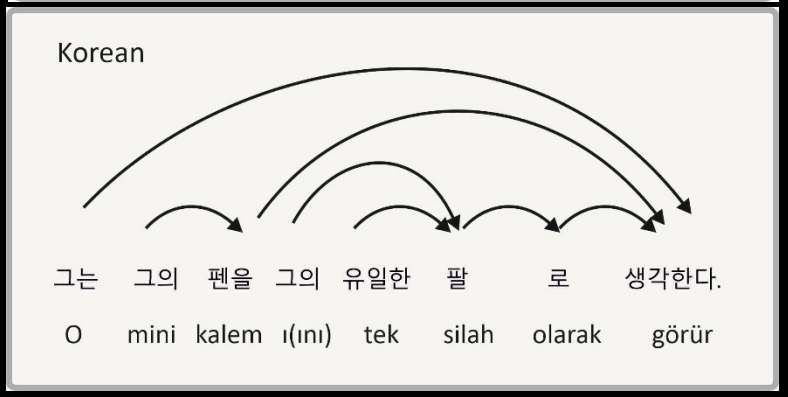
Modern classes

- Noun phrase
- Verb phrase

Structure of Senteces-I

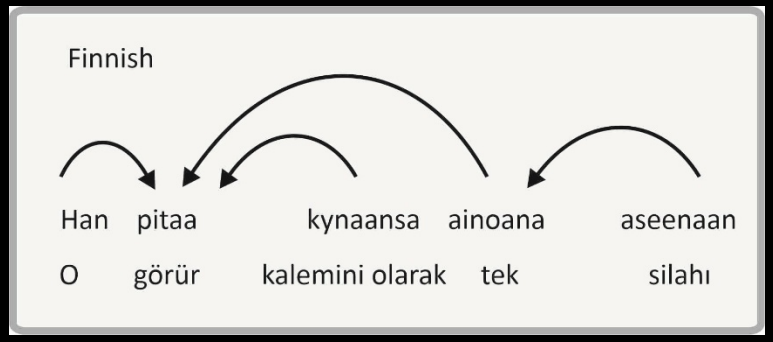


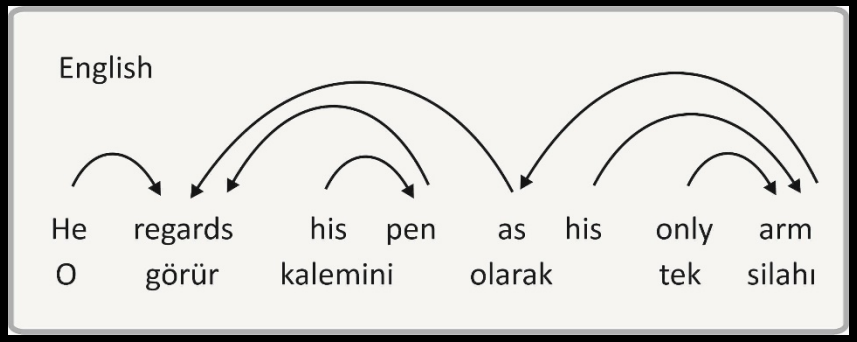


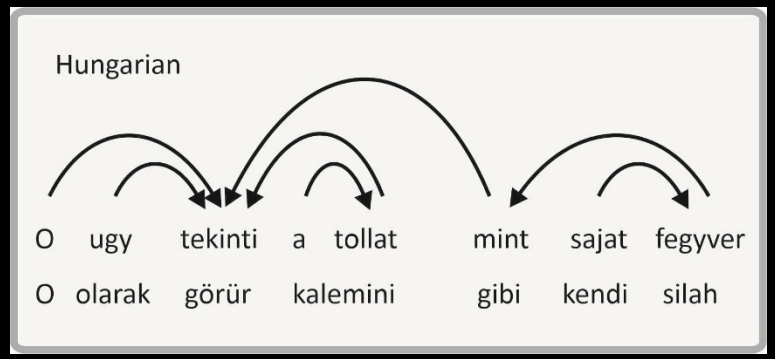


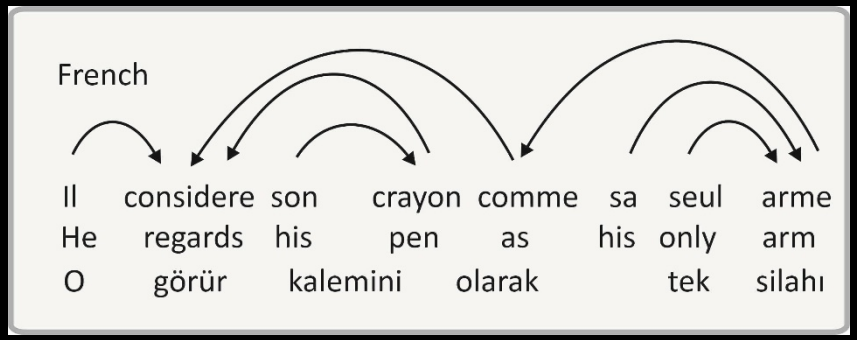
- Turkish : SOV
- Finnish : SVO
- English : SVO
- Chiness : SOV

Structure of Senteces-II









Structure of Senteces-III

- While the order in the sentence cannot be changed in Indo-European languages, it can be changed in Turkish.
- In other words, when the order of the elements in an English sentence is changed, the meaning of the sentence changes.
- When the order of the elements in a Turkish sentence is changed, the basic meaning of the sentence does not change, but the stressed word changes.

  - Arabayı sabunla yıka.
  - Arabayı yıka sabunla.
  - Sabunla arabayı yıka.
  - Sabunla yıka arabayı.
  - Yıka arabayı sabunla.
  - Yıka sabunla arabayı.

araba	: car
sabun	: soap
yıka	: wash
İle	: with

- Because of this feature, Turkish is considered a language with free sequential syntax. The English equivalent of the sample Turkish sentence can be written in one form:
- Wash the car with soap.
- We can say that in a Turkish sentence that can be written in six ways, it is the case suffixes of the words that make the action understood. It is clearly seen that the sentences lose their meaning when we delete the case suffixes.

Parts-of-Speech

Parts-of-speech can be divided into two broad classes

Open Class	Open for new entities (*nouns, verbs, adjectives and adverbs*)
Close Class	Have relatively fixed entities (*preposition, pronoun*) Closed class words are generally functional words like *of, it, you, and, or*

Open Class

Noun : person, place or thing

- Proper noun *(Eren, Beethoven)*
- Common noun *(apple, flowers)* (some languages have masculine and feminine noun like French, Arabic)
  - Count noun *(dog, dogs, two cats)*
  - Mass noun *(information, salt)*

Verb : action and process

- Main verbs *(eat, go, walk)*
- Auxiliary verb *(be, can)*

Adjective : Modifies a noun or a pronoun by describing, identifying, or quantifying words.

An adjective usually precedes the noun or the pronoun which modifies.

(white, black), (good, bad), (young, old)

Adverb : Can modify a verb, an adjective, another adverb, a phrase, or a clause.

An adverb indicates manner, time, place, cause, or degree and answers questions such as "how," "when," "where," "how much".

(home, here, downhill) (very, somewhat, extremely), (slowly, slinky, delicately) (yesterday, Friday)

Closed Class

Prepositions : on, under, over, near, by, from, to, with

Determiners : *a, an, the*
(some languages do not have determiner like “the”)
(some languages have many determiner like “le, la, les in French”)

Pronouns *I, you, he, she, it, we, they, who, other*
(some languages have singular and plural second person like “sen, siz” in Turkish)
(some languages have only one third person like “O” in Turkish”)

Conjunctions : *and, or, but, as, if, when*

Numerals : *one, two, first, second*

Particles : *up, down, on, off, in, out, at, by*

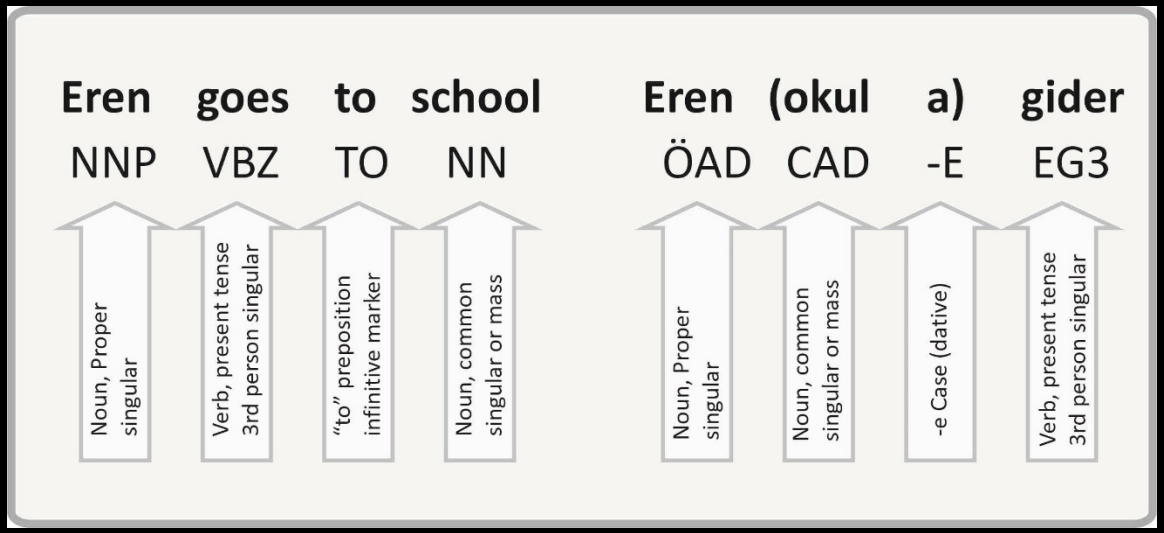
Auxiliary verbs : *can, may, should, are*

Parts-of-Speech (Tagging)

- It is a facilitating approach for language studies to specify the classes of the elements of the sentence as tags.
- The abbreviation labels prepared for the labeling of English sentence elements and used in Penn Treebank and Brown collections. These abbreviations cannot be used to label Turkish items.
- This is because the two languages are members of different language families.

- English has article (the, a) but not Turkish. There is an uncertain.
- There is a preposition (*of, in, by*) in English, there is no preposition in Turkish.
- The way of forming a adjective with comparison and superposition in English is different from Turkish. For example *big, bigger, biggest* – *büyük, daha büyük, en büyük*
- The way of constructing comparison and superposition adverbs in English is different from Turkish. E.g. *fast, faster, fastest* - *hızlı, daha hızlı, çok hızlı.*
- There is a particle (*up, off*) in English, there is no particle in Turkish.
- All verbs are regular in Turkish; The verb includes time and person information. Therefore, it requires special abbreviation for each case. In English, verbs are divided into two classes, regulars and irregulars. In the regular ones, only the third singular mode contains person information and tense information.
- Elements that connect clauses in English, such as *which, that, what, who, whose, how, where*, etc., do not exist in Turkish.
- Turkish is an agglutinative language, especially there are many construction affixes. The number of derivational suffixes in English is relatively few.
- In Turkish, a sentence can consist of only one word. For example, «*sevinçliyim* (I am happy)." It is a sentence consisting of one word, it is clear that the subject is me and the time of verb is the present tense. Turkish sentences can be formed from more than one word or they can be composed of sub-sentences.

Parts-of-Speech (Tagging)



- Ad: Noun (proper noun, common noun, count noun, mass noun)
- Eylem: Verb (main verbs - eat, go, walk, auxiliary verb: be, can, may, should)
- Adıl: Pronoun (I you, he, she, it, we, they, who, other)
- Önad: Adjective (white, black, good bad, young old)
- Belirteç: Adverb (home, here, downhill, somewhat, extremely, slowly)
- Tanımlık: Article determiner: a, an, the)
- İlgeç: Postposition (ago, apart, aside, away, hence, on, short, through)
- Takı (ön takı): Preposition (about, after, among, on, at, beside, over, near, by, from, to, with etc.)
- Bağlaç: Conjunction (and, or, but, as, if, when)

Penn Treebank

Penn Treebank-I

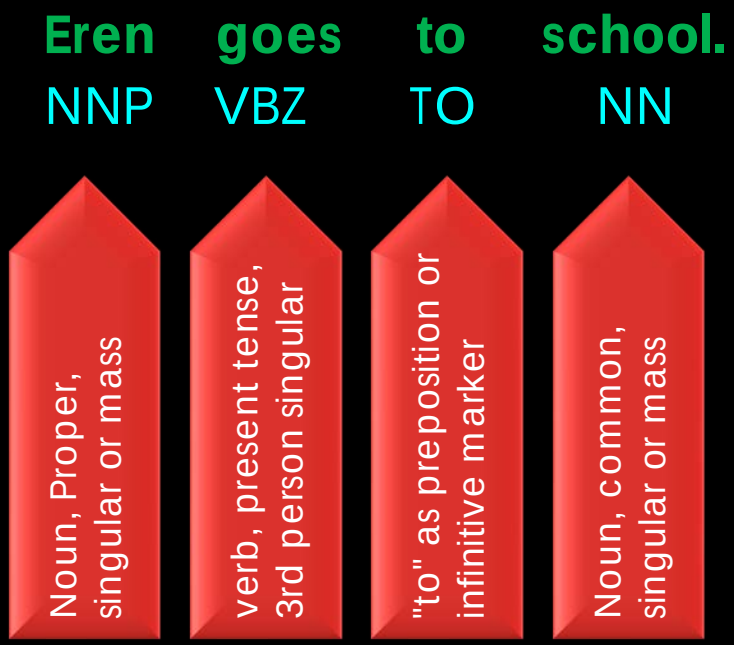
Tag	Description	Examples	Tag	Description	Examples
$	Dollar sign	$ -$ --$ A$ C$ HK$	DT	Determiner	a, the
#	Pound sign	#	EX	Existential there	there
``	Opening quotation mark	` ``	FW	Foreign word	kebap
''	Closing quotation mark	' ''	IN	Preposition or conjunction, subordinating	of, in, by
(	Opening parenthesis	([{	JJ	Adjective or numeral, ordinal	white, yellow
)	Closing parenthesis	)] }	JJR	Adjective, comparative	bigger, smaller
,	Comma	,	JJS	Adjective, superlative	best, wildest
--	Dash	--	LS	List item marker	1,2, one
.	Sentence terminator	. ! ?	MD	Modal auxiliary	can, should
:	Colon or ellipsis	: ; ...	NN	Noun, common, singular or mass	apple
CC	Conjunction, coordinating	and, but, or	NNS	Noun, common, plural	apples
CD	Numeral, cardinal	one, two, three	NNP	Noun, proper, singular	Eren

Penn Treebank-II

Tag	Description	Examples	Tag	Description	Examples
NNPS	Noun, proper, plural	Turks	UH	interjection	ah, oops
PDT	Pre-determiner	all, both	VB	verb, base form	eat
POS	Genitive marker	‘s	VBD	verb, past tense	ate
PRP	pronoun, personal	ı, you, he	VBG	verb, present participle or gerund	eating
PRP$	pronoun, possessive	your, one’s	VBN	verb, past participle	eaten
RB	adverb	quickly, never	VBP	verb, present tense, not 3rd person singular	eat
RBR	adverb, comparative	faster	VBZ	verb, present tense, 3rd person singular	eats
RBS	adverb, superlative	fasters	WDT	WH-determiner	which, that
RP	particle	up, off	WP	WH-pronoun	what, who
SYM	symbol	% & ' '' ''.). * + ,. < = > @	WP$	WH-pronoun, possessive	whose
TO	"to" as preposition or infinitive marker	to	WRB	Wh-adverb	how, where

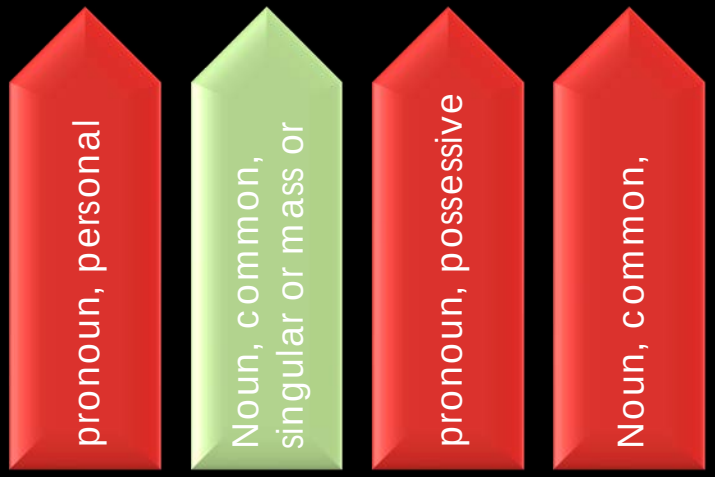
Tagging

The process of marking up the words in a text as corresponding to a particular part-of-speech, based on both its definition and its context.



Tagging Ambiguity

I	book	my	flight.
PP	NN	PP$	NN
PP	VB	PP$	NN



The word "**book**" is ambiguous:

- *to book or the book*

- The important issue of tagging is the solving of ambiguity problems. In some language like Turkish 50% of words have two meanings.
- To solve the ambiguity problems rule base and stochastic methods are used.

Disambiguation

Rule base methods

Since, preceding word is a "*personal pronoun*",
***so, 'book'* should be a verb.**

Stochastic methods

Corpus		
First word	**Second word**	
I	book (tagged as verb)	45%
I	Book (tagged as noun)	2%

***so, 'book'* should be a verb.**

Modern Methods for Lexicon

1957 N. Chomsky argued that context-free linguistics is far from explaining the syntactic features of a language. He said that the conversion of syntactic structures to other syntactic structures can be done with stronger rules.

1965 N. Chomsky proposed transformative grammar based on deep and surface structures. Transformational grammar suggestion was found insufficient and unnecessary by G. Lakoff and J. McCawley. The weakness of Chomsky's proposal was demonstrated in 1970 by S. Peter and R. Ritchie. Thereupon, Chomsky proposed a more restricted form of transformative grammar in 1997.

- Lexical Functional Grammar: Developed by J. Bresnan and R. Kaplan.
- Catagorial Grammar: Developed by R. Montague, B. Partee and E. Bach.
- Generalized Phrase Structure Grammar: Developed by G. Gazdar, I. Sag.
- Head-driven Phrase Structure Grammar: Developed by I. Sag and C. Pollard.

Modern Methods for Lexicon

- Chomsky added the concepts of *deep* and *surface* structure to grammar theory and the transformation in the transition from deep structure to surface structure.
- He called the structure in the mind of man, which includes the semantic interpretation of the syntax of the language and the phonological features of the language, and the surface structure for the form of the deep structures that have undergone transformations.

Phrase: It is a set of words that do not have a predicate and a subject. It is tagged with a single attribute.

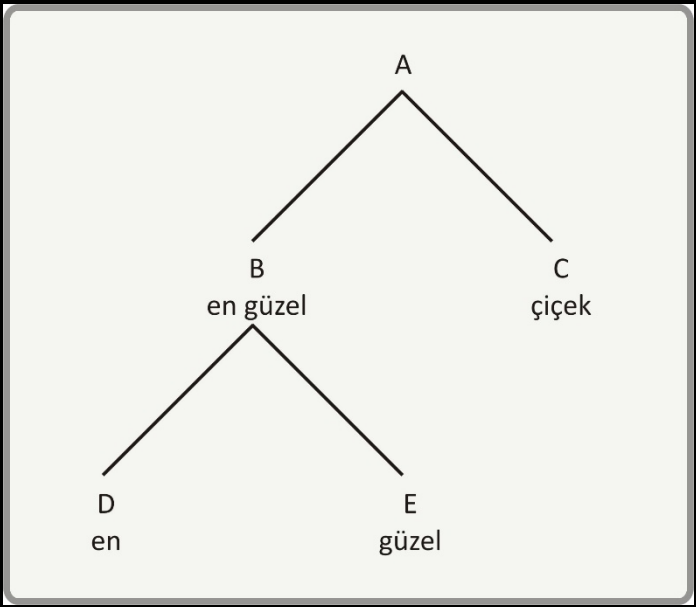
Clause: A set of words that have their own predicate and subject. They are also referred to as phrases.

Constituent: In syntax studies, words or word sets with a single task are defined as constituent. In clause linguistics and subordinate linguistics, sentences are divided into constituent elements. According to the traditional approach, each of the elements that make up the sentence is a founding member.

- Traditional linguists say that a sentence explains thought.
- On the other hand, modern linguists say that sentences are made up of clauses and each has a subject and a predicate. Accordingly, they define clauses as syntax units with subject and predicate. A sentence can consist of one or more clauses. As a result, they say that each clause describes a thought.

Context-free Grammar (CFG)-I

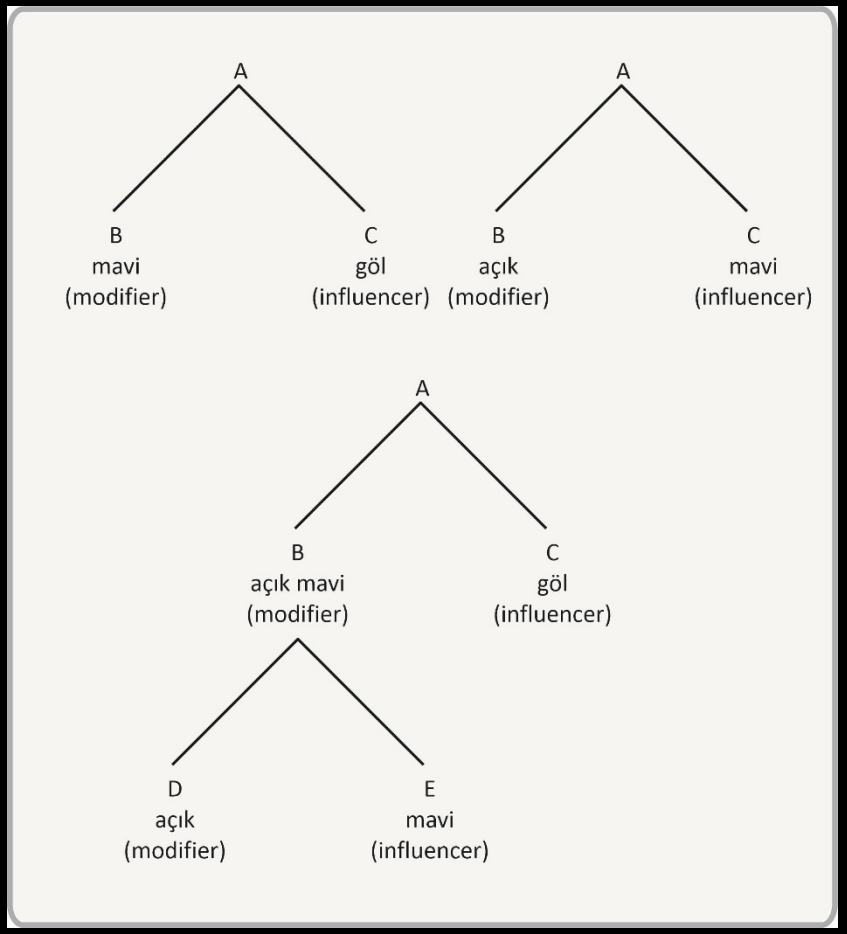
Context-Free Grammar (CFG), also known as Phrase Structure Grammar (PSG), divides the sentence into phrases and shows the relationships between these phrases regardless of meaning. PSG is also called Backus-Naur Form (BNF).



- The nodes of the tree A, B, C, D, E and the lines between the nodes are called the branches of the tree.
- Node A is called the parent node of the tree, and other nodes are called its children (child nodes).
- Children of the same node are siblings of each other.
- It is seen that the sibling nodes are at the same level in the tree structure.
- Structures whose sibling nodes are at the same level are called balanced trees.
- Each node is treated as a constituent and has a function against its sibling. Nodes C, D and E are the last nodes of the tree and cannot be divided into other constituent.

Head Word - Modifier

- **Influencer**: The name given to the word that determines the syntactic class of a phrase. For example, the word that determines the class of *kızarmış ekmek* (toasted bread). Because of this feature, it is defined as the active word of the constituent or simply as the influencer or head word.
- **Modifier**: The word that describes and changes the nature of the influencer is called modifier. The modifier can be a adjective, an adverb, or a relative clause.
  - *Blue lake*: blue: modifier and adjective, lake: influencer and noun.
  - *Light blue*: light: modifier and adjective, blue: adjective
  - *Speaks slowly*: slow: modifier and Adverb, speaks is verb.



Sentence - Clause

Dependency Parsing

Gördüm.
Bu sabah okula erken geldim.
Mavi bereli sarışın kız dün okula geldi.
Bugün okula gelirken yolda gördüğüm adam ünlü bir şarkıcıymış.

Verb: Indicates the action performed by the subject

- Subject do not drop language: The subject must appear explicitly in a sentence or clause. Eg. English and French

  I go downtown today.

- Subject drop languages: The subject may be drop but does not couse a decrease in meaning of the sentence. Eg. Turkish, Russian, Japanese, Italian, Spanish

  Kent merkezine giderim.

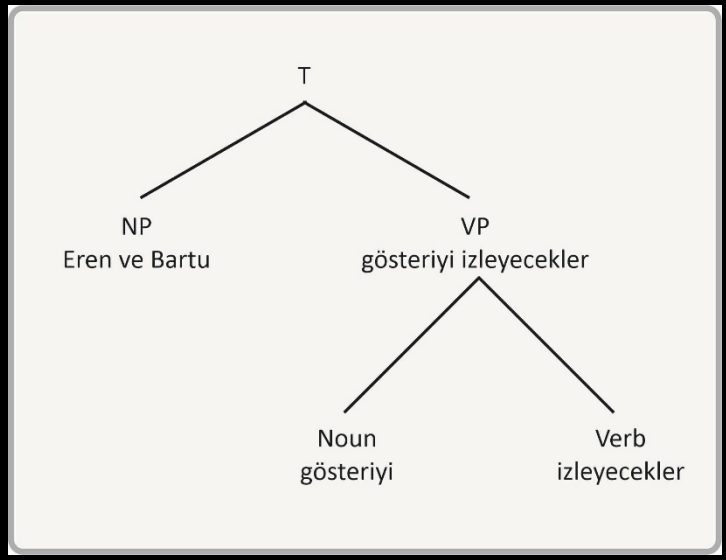
Subject-Verb Compatibility

Ben okumayı severim.
Sen okumayı seversin.
Bartu okumayı sever.
Biz okumayı severiz.
Gazeteler ıslandı.

Phrases-I

Phrase: The constituent that act as subject, object and complement in a sentence are called phrases.

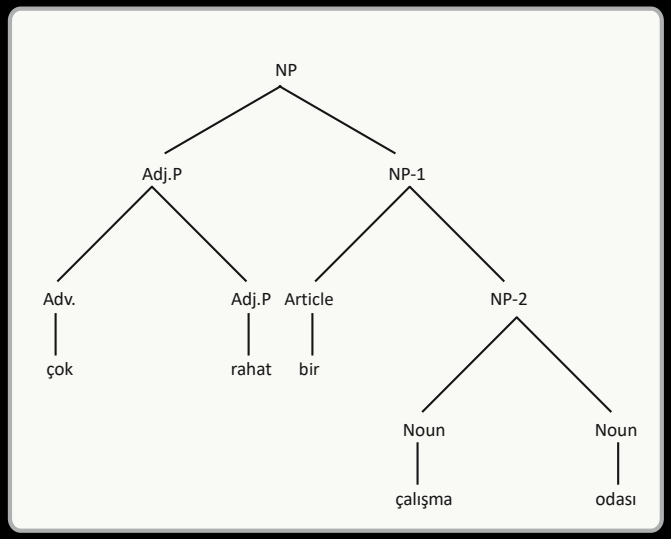
- Noun phrase: NP
- Verb phrase: VP
- Adjective phrase: Adj.P
- Adverb phrase: Adv.P
- Preposition Phrase: PP



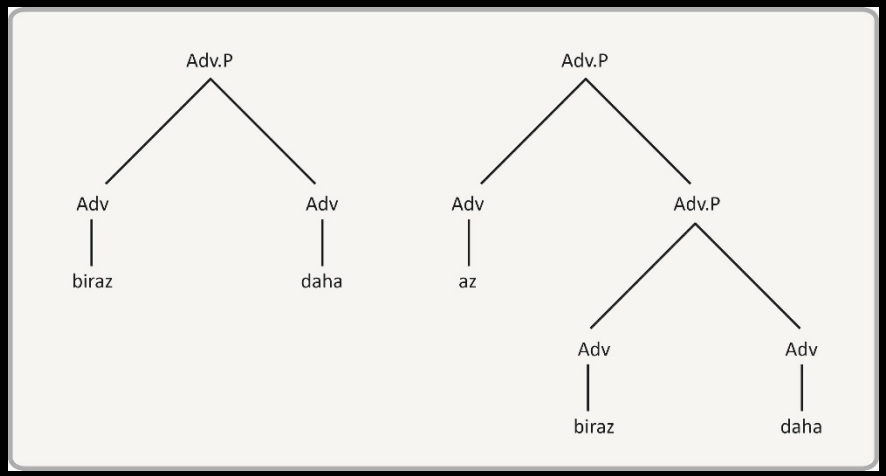
S → NP VP

Phrases-II

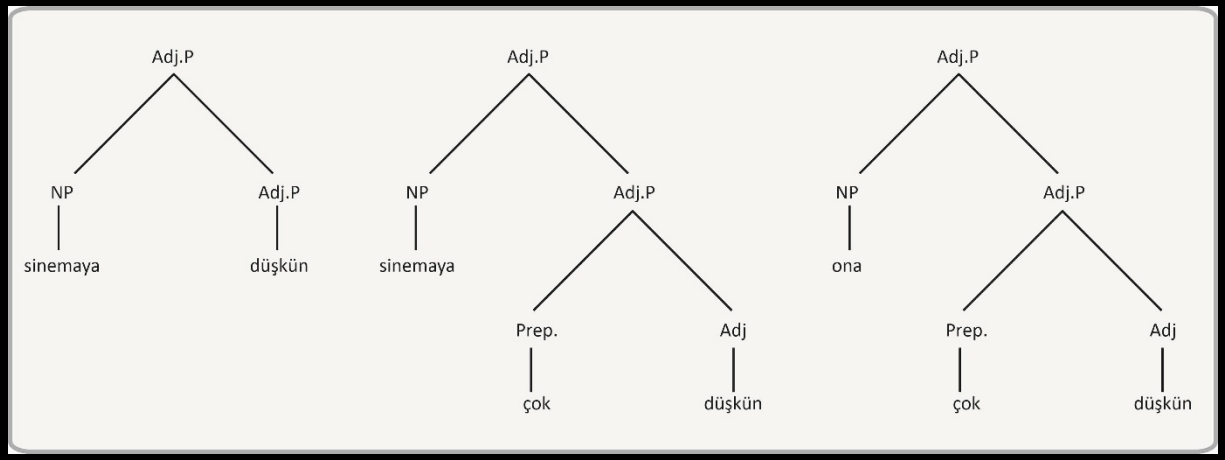
Noun Phrases



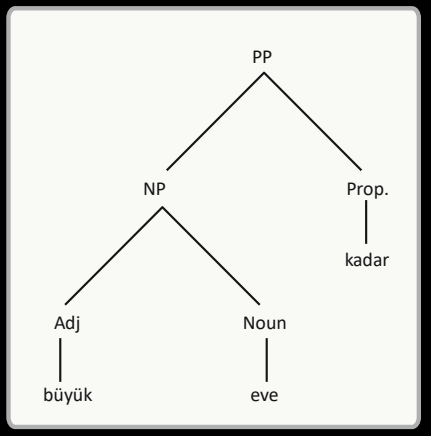
Adverb Phrases



Adjective Phrases

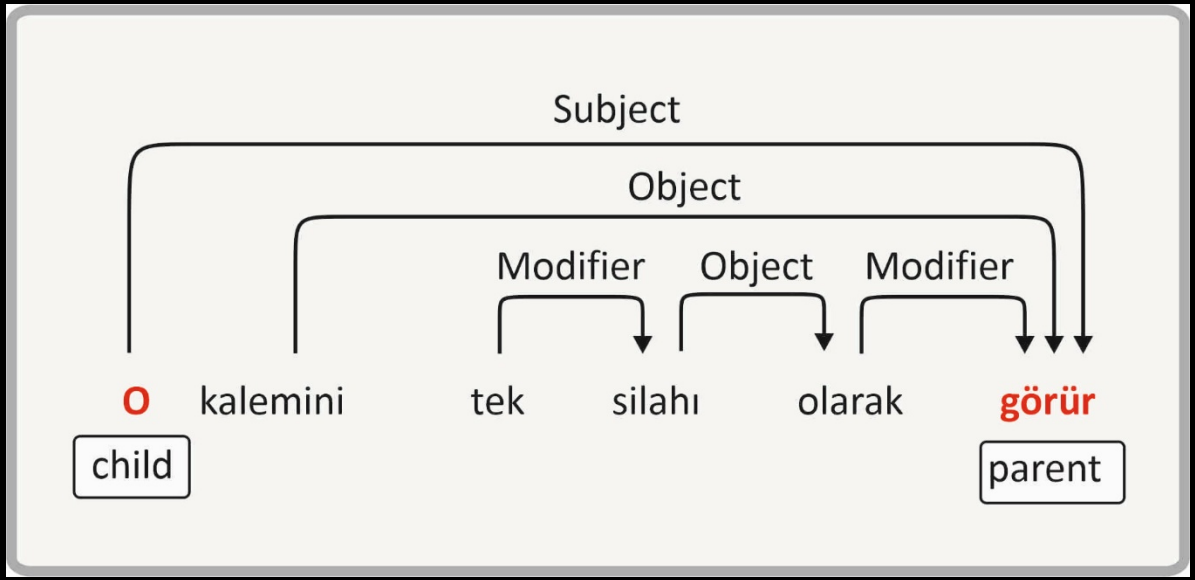


Preposition Phrases



Dependency Grammer

The theory was proposed by Tesnière in 1959. Tesnière defines a sentence as a regular set of words. The human mind knows the relationships between neighboring words and constructs the sentence accordingly. In DG, each child element is linked to a parent element. Today, the DG-item relationship is defined as a satellite (child) - owner (parent) relationship.



Parsing of Sentence

- Phrase Based Parsing (PBP)
- Dependency Parsing (DP)

Phase Based Parsing

- Rule Based Parsing
- Probobality Based Parsing

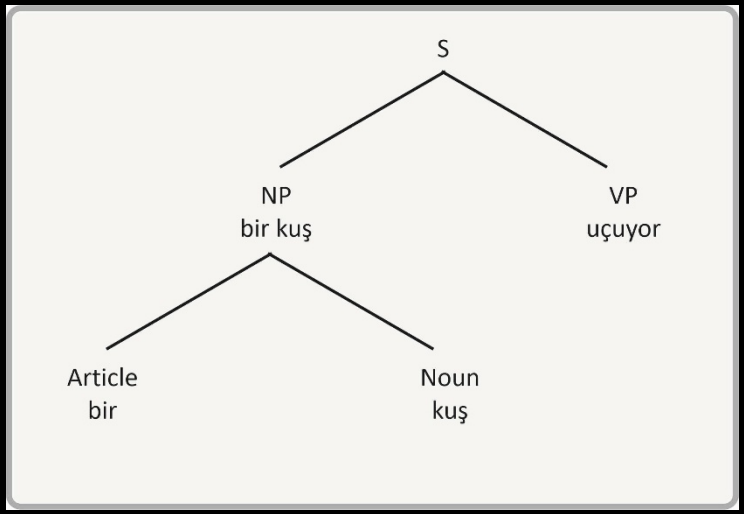
Rule Based Parsing-I

Rule-based solutions treat constituen as essential components of a sentence. Therefore, it is very important to identify the constituen. It is clear that the tree to be found in the sentence will be wrong when the constituen are incorrectly determined. As just explained, it is also important to determine the class of a word.

- Yirmi sekiz ocak pazar günü Ankara'ya gideceğim.
- Ankara'ya yirmi sekiz ocak pazar günü gideceğim.
- Ankara'ya gidişim yirmi sekiz ocak pazar günü .

S → NP, VP

Parsing Tree



Rule Based Parsing-II

Top-Down Parsing

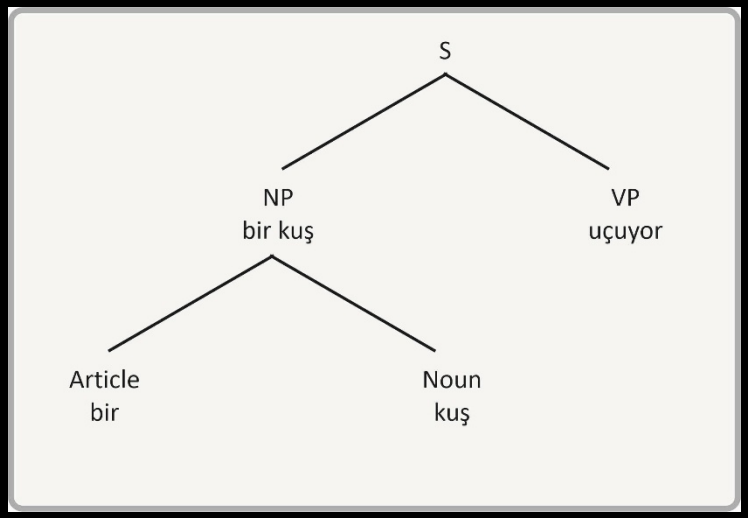
The top-down parsing method starts parsing the sentence from the parent node, continuing to the leaves.

S → (NP) AdvP*NP VP *The Kleene star indicates that there may not be, may exist, and may be more than one.

S → Art Noun

Noun → common Noun

NP → proper Noun



Bottom-Up Parsing

Rule Based Parsing-III

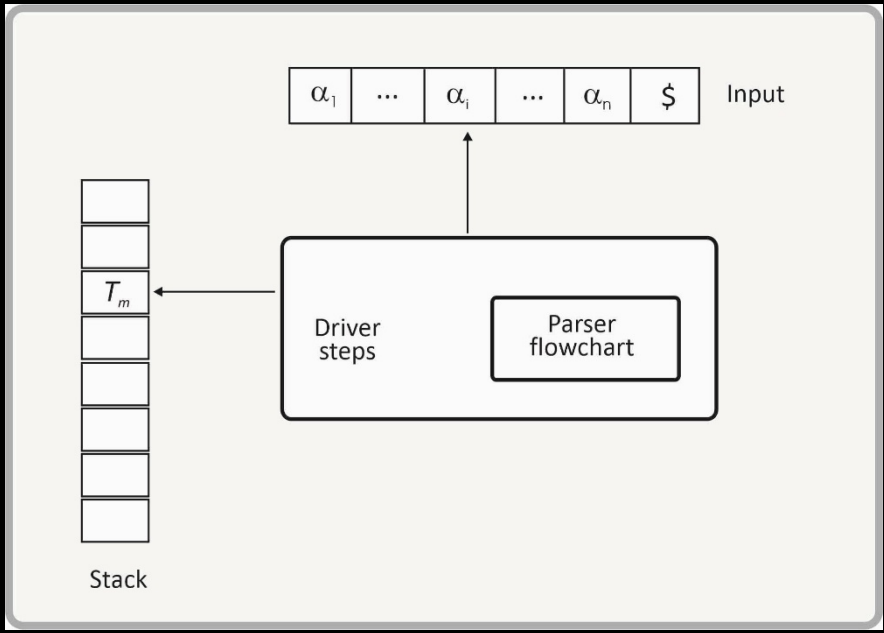
Bottom-Up Parsing (LR Parsing)

This method, which was first proposed by Yngve in 1955, was later used in compilers by Aho and Ullman under the name Translate and Reduce. The input of bottom-up parsing is the words of the sentence. The parser tries to construct the sentence from the words it receives as input. It constantly uses grammar rules while doing these operations. When the sentence is established, the parsing process is completed.

A bird is flying

A: indefinite article
Bird: noun
Flying: verb + 3rd sg + Present tense

L: Scan entry from left to right.
R: Generate the most accurate derivation.
k: The number of forward looks required.



Probobality Based Parsing

In order to overcome the uncertainty problem encountered in rule-based methods, probability-based solutions are preferred. Probability-based methods basically use machine learning methods. For this purpose, Treebank Corpus (corporate consisting of example sentences with tree based given and classes of elements specified) is used. A certain part of Treebank corpus is used to test the developed parser. Penn WSJ Treebank is used in studies on English and ITU-Sabancı-Metu Tree-bank corpus is used in studies on Turkish.

References

[1] L. Karahan, Türkçede Söz Dizimi, Akçağ Yayınları, 15. Baskı, 2010
[2] N. Chomsky, Syntactic Structures, Mounton de Gruyter, 2002
[3] N. Chomsky, Aspects of The Theory of Syntax, The MIT Press, 1965
[4] N. Chomsky, The Minimalist Program, The MIT Press, 1997
[5] E. E. Erguvanlı The Function of Word Order in Turkish Grammar. PhD thesis, Department of Linguistics, University of California, Los Angeles, 1979.
[6] R. Kaplan, J. Bresnan, Lexical-Functional Grammar: A Formal System for Grammatical Representation. In Bresnan 1982b, 173–281. Reprinted in Dalrymple et al. (1995: 29–135).
[7] J. Bresnan, A. Asudeh, I. Toivonen, S. Wechsler, Lexical-Functional Syntax, Wiley Blackwell, 2016
[8] C. Pollard, I Sag, Head-driven Phrase Stucture Grammar, CSLI, Chicago, 1994
[9] H. C. Bozşahin, Ulamsal Dilbilgisi ve Türkçe, Dergipark.ulakbim.gov.tr/dad/article/view/ 5000129417/ 5000118617
[10] M. Collins, Probabilistic Context-Free Grammars (PCFGs), http://www.cs.columbia.edu/~mcollins/ courses/nlp2011/notes/pcfgs.pdf
[11] M. Collins, Lexicalized Probabilistic Context-Free Grammars, http://www.cs.columbia.edu/~mcollins/ courses/nlp2011/notes/lexpcfgs.pdf
[12] A. Carnie, Syntax: A Generative Introduction, 2013 Andrew Carnie. Published 2013 by John Wiley & Sons, Inc.
[13] Y. Falk, Lexical-Functional Grammar, CSLI Pub, 2001,ISNB: 1-57586-340-5
[14] Z. Güngördü, K. Oflazer, Parsing Turkish Using the Lexical Functional Grammar Formalism, Yük. Lisans Tezi, cmp-lg/9406008, 1994, Bilkent
[15] O. T. Şehitoğlu, A Sign-Based Phrase Structure Grammar for Turkish, METU, Yük. Lisans tezi, 1996
[16] Ü. D. Turan, G. Durmuşoğlu, Turkish Syntax, Semantics, Pragmatics and Discourse, T.C. Anadolu Üniv. Yayını No: 2421, 2011
[17] K. Oflazer, B. Say, D. Z. Hakkani, G. Tür, Building a Turkish Treebank, Treebanks: Building and Using Parsed C0I7)()ra, 261-277. © 2003 Kluwer Academic Publishers.
[18] S. Harlow, Introduction to Head-driven Phrase Structure Grammar, http://www-users.york.ac.uk/ ~sjh1/ hpsgcourse.pdf, 2009
[19] M. Collins, A New Statistical Parser Based on Bigram Lexical Dpendencies, Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL), Santa Cruz, CA, 24-27 June 1996, 184–191.
[20] M. Collins, Three Generative, Lexicalised Models for Statistical Parsing , Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL), Madrid, 7-12 July 1997, 16–23.
[21] M. Collins, Head-driven statistical models for natural language parsing. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA. 1999
[22] E. Charniak, A Maximum-entropy-inspired Parser, Proceedings of the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL), Seattle, Washington, 132–139.2000
[23] G. G. Şahin, Building of Turkish Propbank and Semantic Role Labeling of Turkish, Doktora Tezi, İTÜ FEn Bilimleri Ens. 2018.
[24] V. H. Yngve, Syntax and the Problem of Multiple Meaning, Machine Translation of Languages, pp: 208-226, MIT Press, 1955
[25] A. V. Aho, J. D. Ullman, The Theory of Parsing, Translation and Compiling, Printice-Hall, 1972
[26] F. Jelinek, J. Lafferty, D. M. Magerman, R. Mercer, A. Ratnaparkhi, S. Roukos, Decision Tree Parsing using a Hidden Derivation Model, Proc. of the 1994 Human Language Tech. Workshop, PP 272-277, 1994
[27] D. M. Magerman, Natural Language Parsing as Statistical Pattern Recognition, pHD Thesis, Stanford Univ, 1994
[28] D. M. Magerman, Statistical Decision-Tree Models for Parsing, Proc. of the 33rd Annual Meeting of the ACL, PP 276-283, 1995
[29] G. Eryiğit, J. Nivre, K. Oflazer, Dependency Parsing of Turkish, Association for Computational Linguistics, Vol 34, No:3, 2008
[30] G. Erci, Türkçe'nin Bağlılık Ayrıştırması, Doktora Tezi, İTÜ Fen Bilimleri Ens. 2006
[31] G. Eryiğit, E. Adalı, K. Oflazer, Türkçe cümlelerin kural tabanlı bağlılık analizi, Proceedings of the 15th Turkish Symposium on Artificial Intelligence and Neural Networks, Muğla, 21-24 June, 2006, 17–24.
[32] T. Kudo, Y. Matsumoto, Japanese Dependency Analysis Using Cascaded Chunking, Proceedings of the 6th Conference on Computational Natural Language Learning (CoNLL-2002), Taipei, 31 August-1 September, 63–69.
[33] N. Coşkun, Türkçe Tümcelerin Ögelerinin Bulunması, Yük. Lis. Tezi İTÜ Fen Bilimleri Ens. 2013.