Ambo University
woliso campus
Department of Computer Science

Selected Topics in Computer Science

Chapter 1:

Natural Language Processing (NLP)

Instructor Name: Chaltu.M

NLP(natural language processing)

Natural Language

- **Natural language** refers to human languages (Amharic, Afaan Oromo, Tigrigna, English, Arabic, Chinese, etc.), as opposed to artificial/programming languages such as C++,Java, Pascal, etc.
- **Natural language** is represented using texts in spoken or written forms.

Natural Language Processing

- NLP is the computerized approach to analyzing text that is based on both a set of theories and a set of technologies.

Cont…

Natural Language Processing

- A more comprehensive definition of NLP is given as:
- An interdisciplinary field of study dealing with computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.

Cont…

...interdisciplinary field...

➢Several fields including linguistics, psycholinguistics, mathematics, computer science, and electrical engineering contribute to the research and development of NLP.

...computational techniques...

➢Multiple models, methods and algorithms are employed to accomplish a particular type of language analysis.

...naturally occurring texts...

➢Texts can be in spoken or written forms representing natural languages used by humans to communicate to one another.

Cont…

...levels of linguistic analysis...

- Multiple types of language processing are known to be at work when humans produce or comprehend language.

...human-like language processing...

- NLP strives for human-like performance, and thus considered as a discipline within Artificial Intelligence.

...tasks or applications...

- The goal of NLP is to accomplish human-like language processing for various tasks and applications such as **machine translation**, **information retrieval**, **question-answering**, etc.

Cont…

- Closely related (and overlapping) fields are Natural Language Understanding and Computational Linguistics.
- The field of NLP was originally referred to as Natural Language Understanding (NLU) in the early days of Artificial Intelligence. A full NLU system would be able to:
    - ✓paraphrase an input text
    - ✓translate the text into another language
    - ✓answer questions about the contents of the text
    - ✓draw inferences from the text.
- Computational Linguistics emerged as a field of study in linguistics with the purpose of providing computational models for various linguistics phenomena.

Cont…

- An alternative view on NLP is that it is a computer system which uses natural language as input and/or output. In this view, NLP is considered to have two distinct focuses-Natural Language Understanding and Natural Language Generation.
- The task of Natural Language Understanding is equivalent to the role of reader/listener, whereas the task of Natural Language Generation is that of the writer/speaker.

Importance of NLP

- ➢Natural language is the preferred medium of communication for people.
  - ✓People communicate with each other in natural languages.
  - ✓Scientific articles, magazines etc. are all in natural languages.
  - ✓Billions of web pages are also in natural languages
- ➢Computers can do useful things for us if:
  - ✓Data is in structured form, e.g. databases, knowledge bases.
  - ✓Specifications are in formal language, e.g. programming languages.
- ➢NLP bridges the communication gap between people and computers.
  - ✓ lead to a better and a more natural communication with computers.
  - ✓process an ever increasing amount of natural language data generated by people
    e.g. extract required information from web.

Difficulty of NLP

- People generally don’t appreciate how intelligent they are as natural language processors.

- For them natural language processing is deceptively simple because no conscious effort is required.
  Since computers are orders of magnitude faster, many find it hard to believe that computers are not good at processing natural languages.

- NLP is hard because of:
  Ambiguity - A word, term, phrase or sentence could mean several possible things.
  - Computer languages are designed to be unambiguous.
  Variability - Lots of ways to express the same thing.
  - Computer languages have variability but the equivalence of expressions can be automatically detected.

Levels of Linguistic Analysis

1. Morphology:

- is the study of the structure of words.is the study of word formation from smallest unit of words.
- At morphological level, the smallest parts of words that carry meanings, affixes are analyzed.
- **Morphology** is important in NLP because language is **productive**: in any given text we will encounter words and word forms that we haven’t seen before and that are not in our precompiled dictionary.
- **Morpheme**: The minimal units of morphology. e.g. helpfulness.
- **Stem:** part of the word that never changes even when morphologically
- inflected.

For example, walk is the stem for the words walk, walks, walking, and walked.

Cont…

- **Root/Lemma** is citation form of a set of words, e.g. *break* is the root form for the words *break*, *breaks*, *breaking*, *broke*, and *broken*.
- **Part-of-Speech/Lexical Category/Word Class** : is a linguistic category of words that explains how the word is used in a sentence.
- Although different languages may have different classification schemes, English and Amharic words are usually classified into eight lexical categories: noun, pronoun, adjective, verb, adverb, preposition, conjunction and interjection.
- Morphologically important parts-of-speech in English and Amharic include: **nouns, adjectives and verbs**.

Cont…

- **Morphological Analysis -** the process of finding morphemes of a word.
- It is an important component of Spelling Correction, Machine Translation, Information Retrieval, Text Generation and other natural language systems.
- **Morphological Generation:** the process of generating different words from a morpheme.
- **Lemmatization:** the process of finding the root/lemma of a word.
- **Stemming:** the process of finding the stems of a word.
- **Morphemes** can be classified in two ways:
  1. **Free** versus **Bound**
  2. **Roots, Affixes** versus **Combining Forms**

Cont…

1.Free versus Bound

1. **Free morphemes** - morphemes that can stand on their own to give meaning.

e.g. friend in friendly

large in enlarge
help in helpfulness
perform in performance

2. **Bound morphemes** - morphemes that cannot stand on their own as a word

e.g. -ly in friendly
en- in enlarge
-ful and -ness in helpfulness

Cont…

2. Roots, Affixes versus Combining Forms

- **Roots** - morphemes (within a non-compound word) that makes the most precise and concrete contribution to the word's meaning, and is either the sole morpheme or else the only one that is not an affix.

  e.g. break in breaks

  help in unhelpfulness

- **Affixes -** bound morphemes that either precede, follow or are inserted inside the root or stem.

  e.g. **Prefix**: en- in enlarge is an affix that precedes the root *large*

  Suffix: -ly in *l*argely is an affix that follows the root *large*

  Infix: is an affix that is inserted inside the root.

  Circumfix: is an affix that precedes and follows the stem.

Cont…

- **Combining Forms -** morphemes that are formed from two bound or free-like roots.

e.g. two free roots: photo and graph in photograph

two bound roots: electro- and -lysis in electrolysis

bound and free roots: Ethio- and America in Ethio-American

- **The major types of morphological process:**

1. ***Inflections*** are the systematic modifications of a ***root form*** by means of prefixes and suffixes to indicate grammatical distinctions like singular and plural.

- Inflection does not change word class or meaning significantly, but varies features such as **tense**, **number**, and **plurality.**

Cont…

➢All the inflectional forms of a word are often grouped as manifestations of a single lexeme.

2. **Derivation** is less systematic.

- This derivation process usually changes the **part-of-speech category**.

➢An example is the derivation of the adverb **widely** from the adjective **wide**.

3.Compounding refers to the merging of two or more words into a new word.

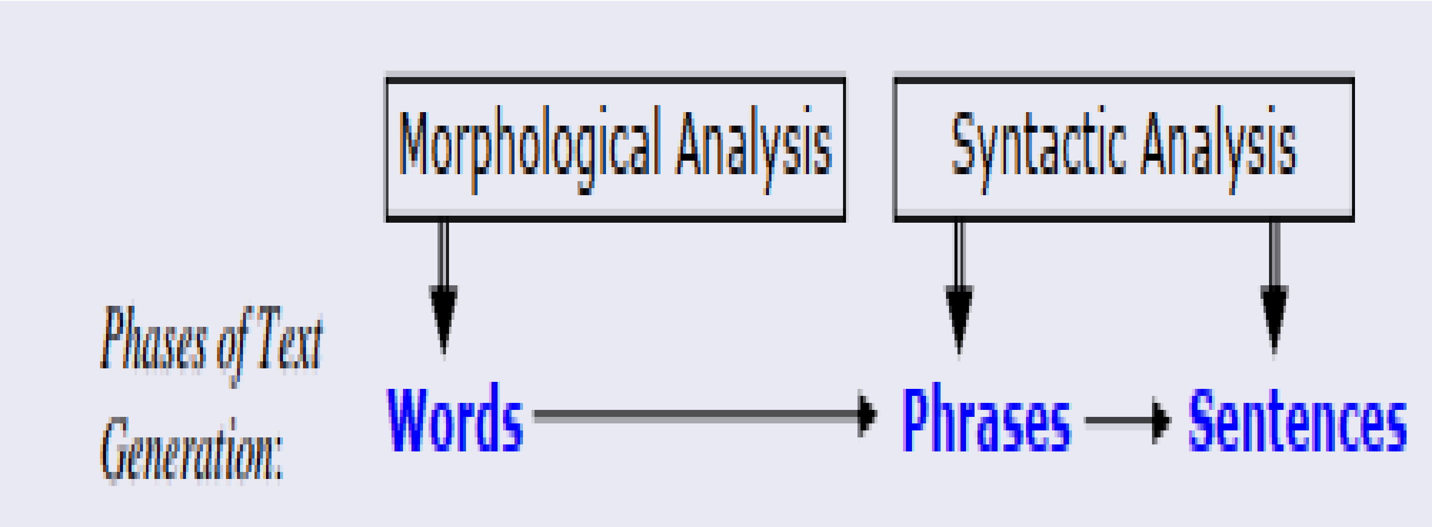
➢English has many noun-noun compounds, nouns that are carbonations of two other nouns.

➢Examples are disk drive, college degree, high school, etc.….

Cont…

2. Syntax- refers to the way words are related to each other in a sentence.

- **Syntactic Analysis-** analyzes:
  - how words are grouped together into phrases;
  - what words modify other words;
  - what words are of central importance to the sentence.
- **Syntactic Analysis** is used in many NLP applications such as:
  - Grammar Checking
  - Question Answering
  - Information Extraction
  - Machine Translation



Cont…

English Noun Phrases

Student, the student, that student, two students, many students, Clever student, A student of computer science

English Verb Phrases

- ✓turn, turn on, is turning on, have been working
- ✓threatened to throw himself into the window
- ✓was an understandable reaction by the visitors
- ✓is amazingly rich in minerals

Cont…

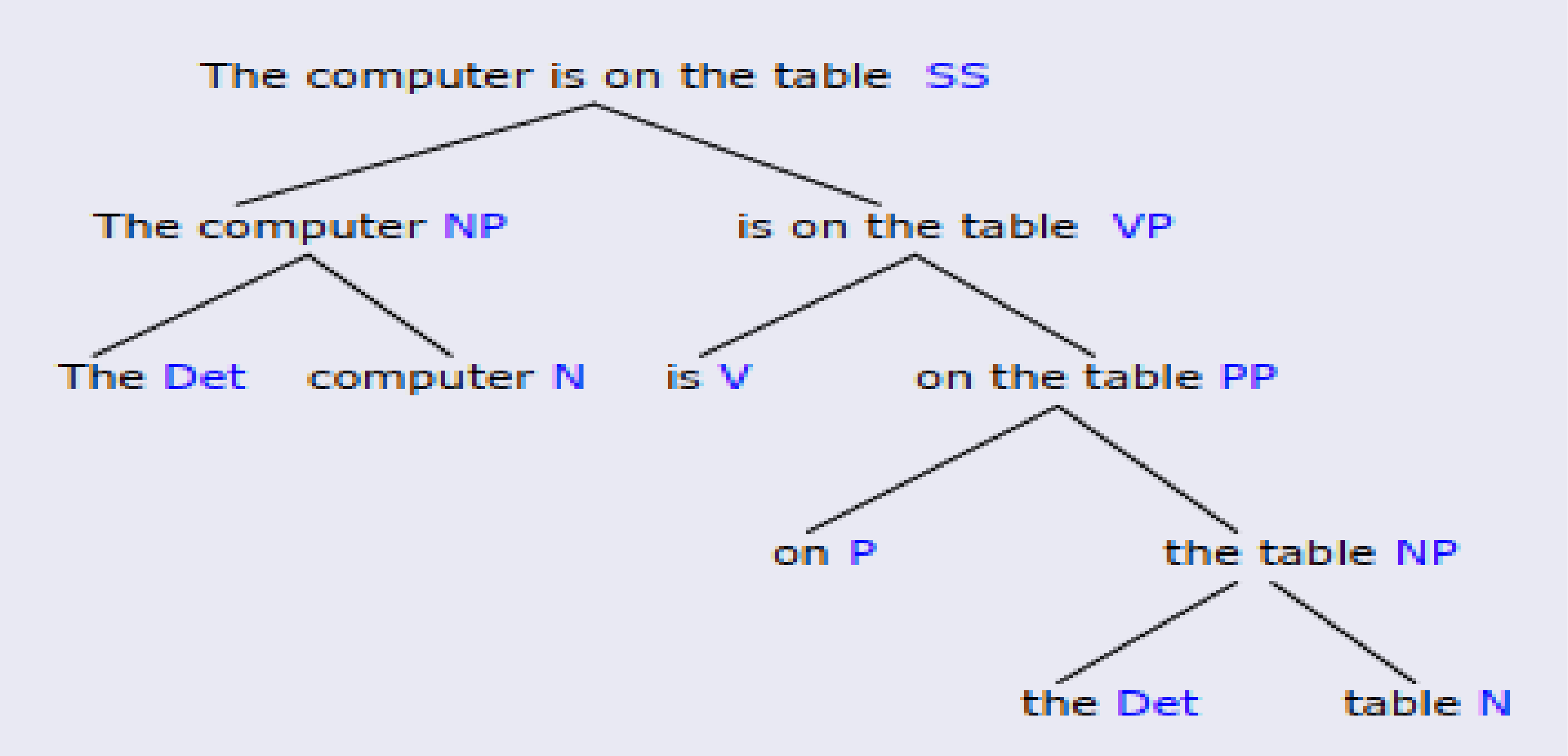
3. Parsing: is a derivation process which identifies the structure of sentences using a given grammar.

- ➢considered as a special case of a search problem.
  - o two basic methods of searching are used
    - ✓**top-down strategy**
    - ✓**bottom-up strategy**
  - o methods of improving efficiency
    - ✓**storing lexical rules separately**
    - ✓**chunking**

Cont…

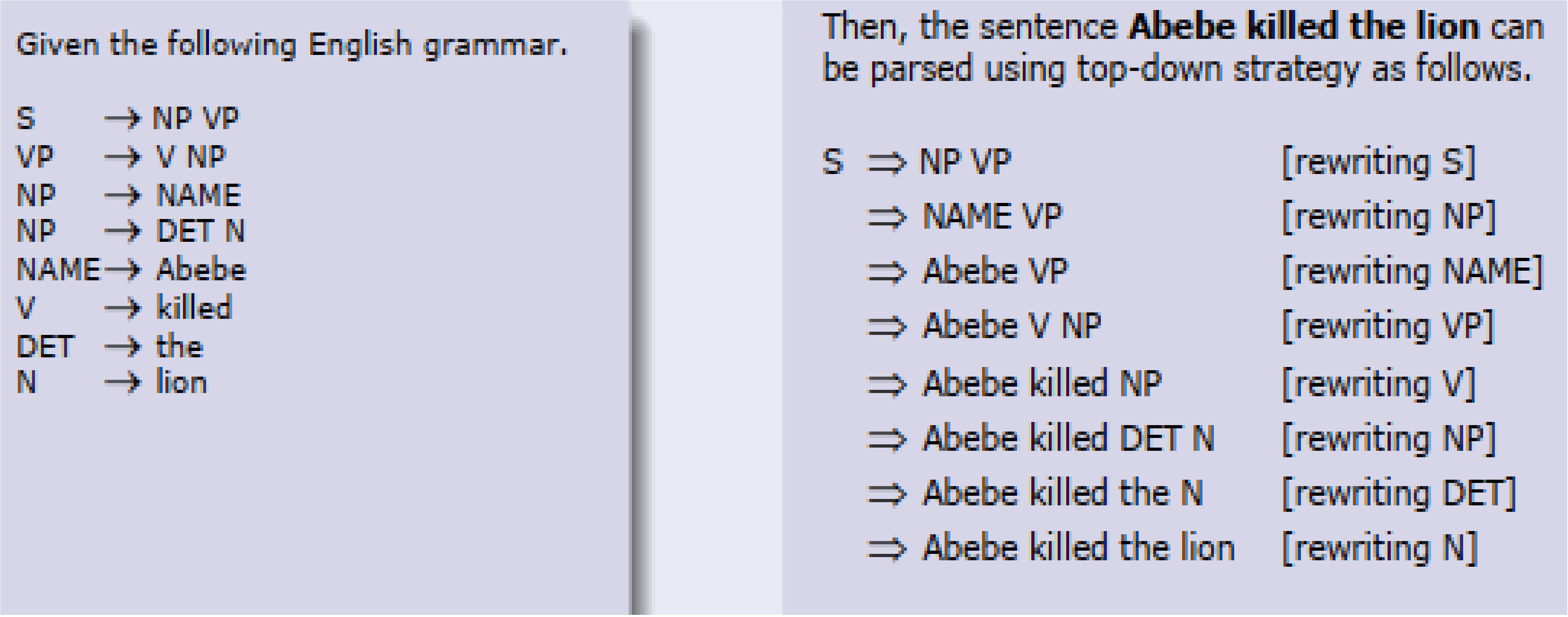
Simple Sentences (English)

The computer is on the table …represent using parse tree



Cont…

- **Top-down parsing** starts with the symbol **S** and then searches through different ways to rewrite the symbols until the input sentence is generated.



Given the following English grammar.

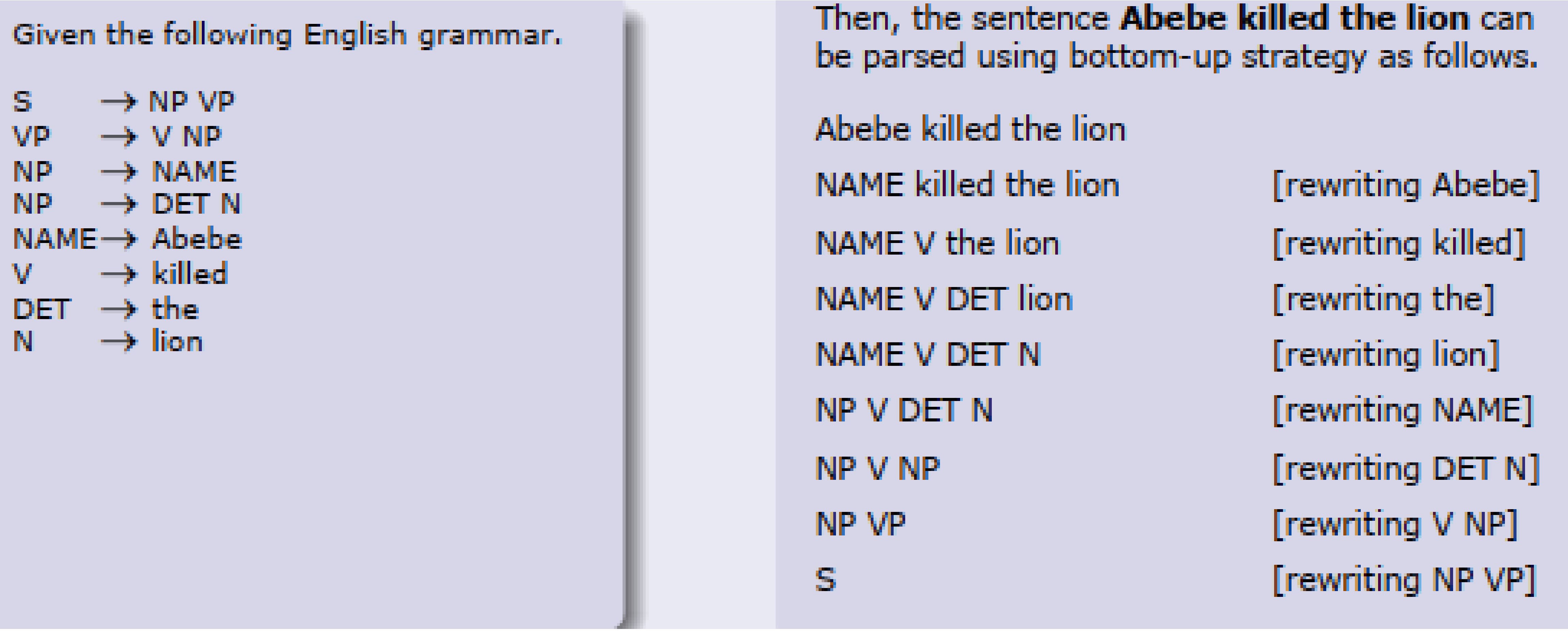
S	→ NP VP
VP	→ V NP
NP	→ NAME
NP	→ DET N
NAME	→ Abebe
V	→ killed
DET	→ the
N	→ lion

Then, the sentence **Abebe killed the lion** can be parsed using top-down strategy as follows.

S	⇒ NP VP	[rewriting S]
	⇒ NAME VP	[rewriting NP]
	⇒ Abebe VP	[rewriting NAME]
	⇒ Abebe V NP	[rewriting VP]
	⇒ Abebe killed NP	[rewriting V]
	⇒ Abebe killed DET N	[rewriting NP]
	⇒ Abebe killed the N	[rewriting DET]
	⇒ Abebe killed the lion	[rewriting N]

.. Cont’d

- **Bottom-up parsing** starts with words in a sentence and uses production rules backward to reduce the sequence of symbols until it consists solely of **S.**



Given the following English grammar.

S → NP VP
VP → V NP
NP → NAME
NP → DET N
NAME→ Abebe
V → killed
DET → the
N → lion

Then, the sentence **Abebe killed the lion** can be parsed using bottom-up strategy as follows.

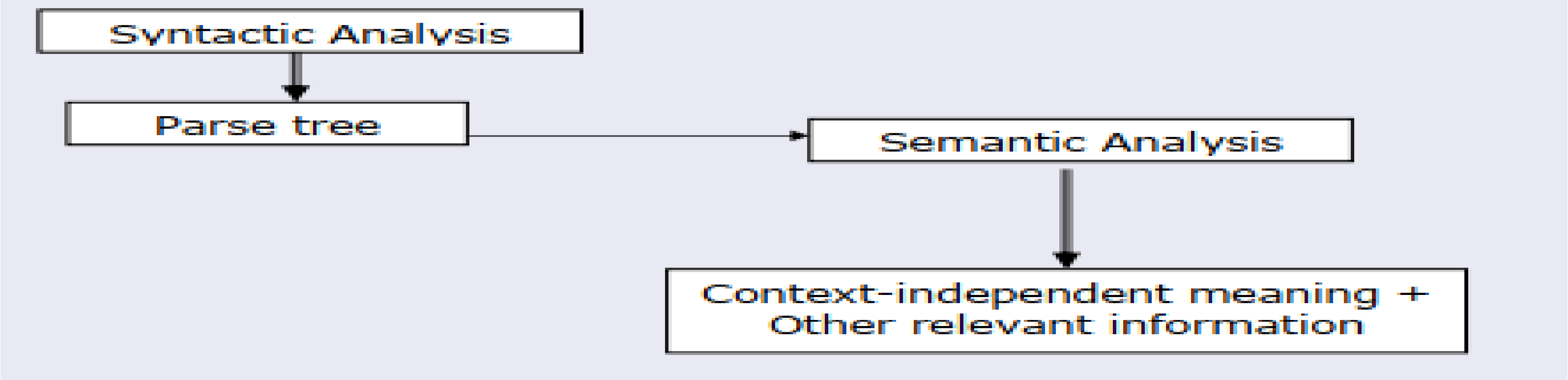
Abebe killed the lion	
NAME killed the lion	[rewriting Abebe]
NAME V the lion	[rewriting killed]
NAME V DET lion	[rewriting the]
NAME V DET N	[rewriting lion]
NP V DET N	[rewriting NAME]
NP V NP	[rewriting DET N]
NP VP	[rewriting V NP]
S	[rewriting NP VP]

Cont…

- **Chunking,** also called **partial parsing**, is a technique which attempts to model human parsing by breaking the text up into small pieces, each parsed separately.
- Chunk boundaries correspond roughly to the pauses in everyday speech.
- For example, consider the following sentence.
  - ✓When I read a sentence, I read it a chunk at a time.
- Then, the following chunks can be identified.
  - ✓[**When I read**] [**a sentence**], [**I read it**] [**a chunk**] [**at a time**].

Cont…

4. **Semantic Analysis** involves extraction of context-independent aspects of a sentence's meaning, including the semantic roles of entities mentioned in the sentence, and quantification information, such as cardinality, iteration, and dependency.



- **Semantic Analysis** is an important component for many NLP applications.
- **A semantic role** is the underlying relationship that a participant has with the main verb in a clause.
- Semantic roles are identified from the grammatical relations

Cont…

5. Discourse deals with the properties of the text as a whole that convey meaning by making connections between component sentences.

- Much of language interpretation is dependent on the preceding discourse/dialogue.
- Discourse imposes meaning and structure on individual sentences (or utterances) that go well beyond the compositional meaning of sentences in isolation.
- The most common methods applied for discourse processing are discourse segmentation and reference resolution.
- Discourse processing is required for:
- Natural language understanding , Text summarization, Machine translation, Natural language generation

Cont…

- ➢**References** in a given text can be anaphoric or coreferential noun phrases.
- ➢The task of reference resolution is to determine which noun phrases refer to each real world entity mentioned in the text.
- ➢Anaphora:
  - ✓An expression α1 is in an anaphoric relation with expression α2 if and only if the interpretation of α1 depends on α2.
  - ✓The relation holds within a text.
- ➢Coreference:
  - ✓Two expressions α1 and α2 are coreferential if and only if Referent (α1) = Referent (α2).
  - ✓The expressions can be in the same text or different texts, in the same language or different language.

Cont…

- Some expressions are both coreferential and anaphoric.
- *A bus had to divert to the local hospital when one of the passengers had a heart attack. It go to the hospital in time and the man's life was saved.*
- **Coreferential**: {*the local hospital, the hospital*} , {*bus, it*}, {*one of the passengers, the man*}
- **Anaphoric**: {it}
- **Pragmatics** is the study of how linguistic properties and contextual factors interact in the interpretation of utterances, enabling hearers to bridge the gap between sentence meaning and speaker's meaning.

Cont…

- **6.Disambiguation**
- A text is said to be **ambiguous** if multiple or alternative linguistic structures can be built for it.
- For example, given the following lexical entry in a lexicon
- Ambiguity may occur at:
  - ✓Phonological level - multiple orthographic representations
  - ✓Morphological level - multiple word classes
  - ✓Syntactic level - different ways to parse the tree
  - ✓Semantic level - different meanings of the same parse tree
  - ✓Discourse level - different references of the same anaphora
  - ✓Pragmatic level - cannot be clearly interpreted

Approaches to NLP

- **Rule-based Approach**
- Rule-based systems are based on explicit representation of facts about language through well-understood knowledge representation schemes and associated algorithms.
- Rule-based systems usually consist of a set of rules, an inference engine, and a workspace or working memory.
- Knowledge is represented as facts or rules in the rule-based approach.
- The inference engine repeatedly selects a rule whose condition is satisfied and executes the rule.
- The primary source of evidence in rule-based systems comes from human-developed rules (e.g. grammatical rules) and lexicons.
- Rule-based approaches have been used tasks such as information extraction, text categorization, ambiguity resolution, and so on.

Cont…

2. Statistical Approach

- Statistical approaches employ various mathematical techniques and often use large text corpora to develop approximate generalized models of linguistic phenomena based on actual examples of these phenomena provided by the text corpora without adding significant linguistic or world knowledge.
- The primary source of evidence in statistical systems comes from observable data (e.g. Large text corpora).
- Statistical approaches have typically been used in tasks such as speech recognition, parsing, part-of-speech tagging, statistical machine translation, statistical grammar learning, and so on.

Cont…

3. Connectionist Approach

- **A connectionist model** is a network of interconnected simple processing units with knowledge stored in the weights of the connections between units.
- Similar to the statistical approaches, connectionist approaches also develop generalized models from examples of linguistic phenomena.
- What separates connectionism from other statistical methods is that connectionist models combine statistical learning with various theories of representation.
- Connectionist approaches have been used in tasks such as word-sense disambiguation, language generation, syntactic parsing, limited domain translation tasks, and so on.

Application of NLP

1. Spelling Correction and Grammar Checking

- **Spelling Correction** is a process of detecting and sometimes providing suggestions for incorrectly spelled words in a text.
- **Spell Checker** is an application program that flags words in a document that may not be spelled correctly.
- **Grammar Checking is** an application program that checks whether the sentence is constructed correctly or not.. subject-Verb agreement and others..

2. Information retrieval

- provides a list of potentially relevant documents in response to a user's query.

Cont…

3. Information Extraction focuses on the recognition, tagging, and extraction of certain key elements of information (e.g. persons, companies, locations, organizations, etc.) from large collections of text into a structured representation.

➢It has the following subtasks:

- ✓Named Entity Recognition: recognition of entity names.
- ✓Relation Detection and Classification: identification of relations between entities.
- ✓Coreference and Anaphoric Resolution: resolving links to previously named entities.
- ✓Temporal and Event Processing: recognizing temporal expressions and analyzing events.
- ✓Template Filling: filling in the extracted information.
- ✓

Cont…

4. **Machine Translation** is an automatic translation of text from one language to another.

5. **Question-Answering** provides the user with either just the text of the answer itself or answer-providing passages.

6. **Dialogue Systems** are agents that converse with human beings in a coherent structure using several modes of communication such as text, speech, gesture, etc.

7. **Text Summarization:** reduces a larger text into a shorter, yet richly constituted representation of the original document.

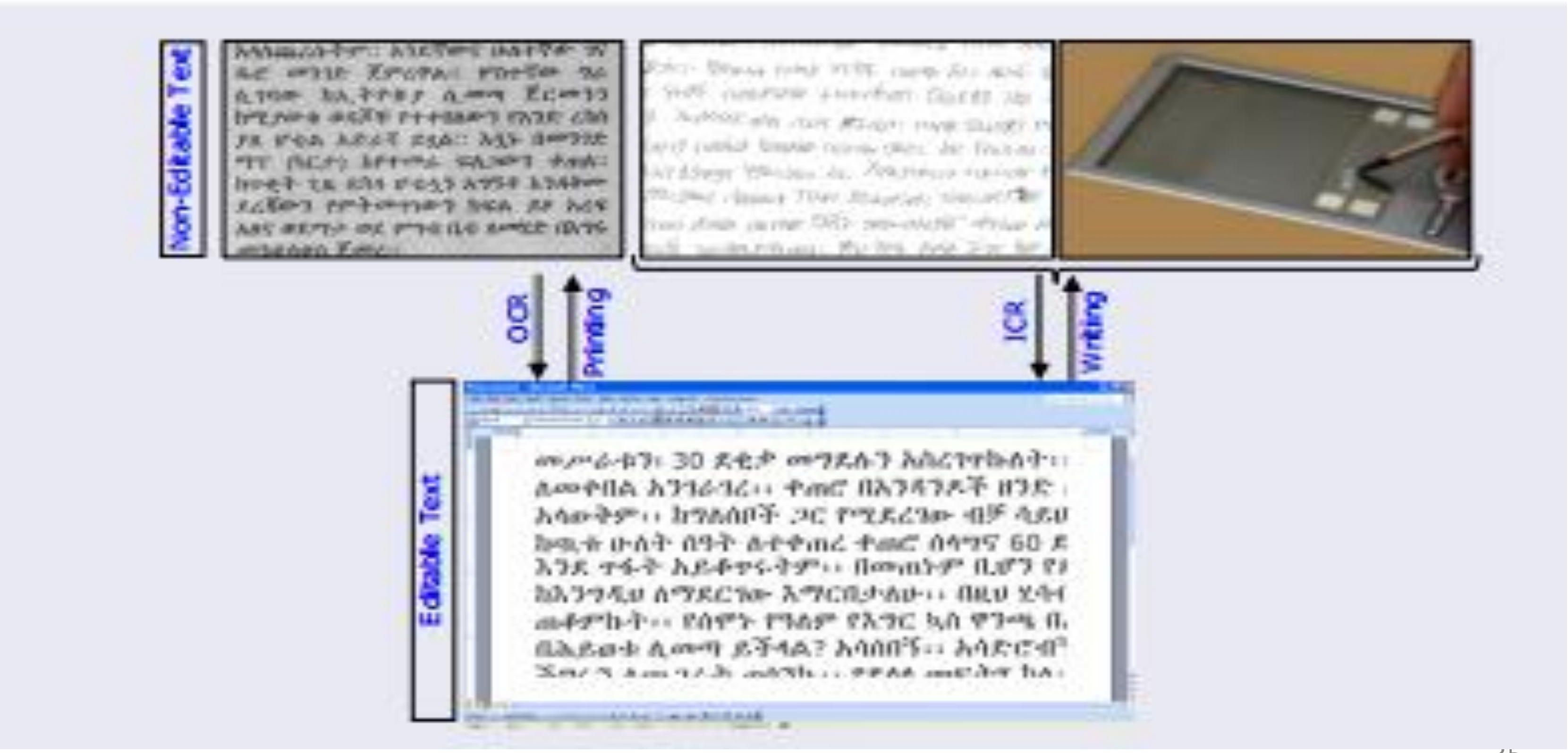
8. **Speech Recognition** is the process of converting spoken words (acoustic signals) into equivalent text.

9.**Speech Synthesis,** also known as Text-to-Speech system, performs the reverse process, i.e. artificially produces human speech from a given text.

Cont…

10. Optical Character Recognition (OCR) is a computerized system that converts non-editable text to machine-encoded text.

- ➢If the text to be converted is handwritten, the system is also known as Intelligent Character Recognition (ICR).


Non-Editable Text
OCR
Printing
ICR
Writing
Editable Text