



**ADDIS ABABA SCIENCE AND TECHNOLOGY UNIVERSITY**  
**(AASTU)**

**COLLEGE OF ELECTRICAL AND MECHANICAL ENGINEERING**  
**DEPARTMENT OF SOFTWARE ENGINEERING**

**Formal Language & Automata Theory(SWEG3101)**

- ☐ Name: Teklu Moges
- ☐ ID: ETS1531/14
- ☐ Section: E
- ☐ Submission date: 1 / 6 / 2024
- ☐ Submitted to: Inst. Sebahadin
- ☐ Academic Year: 2024

## Table of content

<b>Introduction.....</b>	<b>2</b>
<b>Type 0 Languages ( Recursively Enumerable Languages).....</b>	<b>3</b>
Amharic-English Machine Translation.....	3
Extended Parallel Corpus for Amharic-English Machine Translation.....	4
The Effect of Normalization for Bi-directional Amharic-English Neural Machine Translation.....	5
<b>Type 1 Languages (Context-Sensitive Languages).....</b>	<b>6</b>
Application Area: Morphological analysis and generation for Ethiopian languages.....	6
Main Functionalities of Morphological Analysis and Generation.....	6
Why FLAT Concepts Are Needed?.....	6
Approaches and Examples.....	7
<b>Type 2 Languages (Context-Free Languages ).....</b>	<b>8</b>
Context-Free Languages and Amharic Part-of-Speech Tagging.....	8
Importance of PoS Tagging in NLP.....	8
Methodology.....	8
Technological Implementations.....	9
Accuracy and Performance.....	9
<b>Type 3 Languages (Regular Languages).....</b>	<b>10</b>
Application Area: Text Summarization for Ethiopian Languages.....	10
Importance of Text Summarization in NLP.....	10
Methodology.....	10
Technological Implementations.....	11
Accuracy and Performance.....	11
Text Summarization for Afaan Oromo.....	11
Exploring Amharic Text Summarization.....	12
<b>Conclusion.....</b>	<b>13</b>
<b>Reference.....</b>	<b>14</b>

## Introduction

Natural language processing (NLP) is a subfield of Artificial Intelligence that equips computer systems with the ability to analyze and synthesize spoken and written languages similar to human beings. The term ‘Natural’ is used to distinguish these formal languages from programming languages, emphasizing the unique challenges and tasks performed by NLP. The exponential growth of computing power, coupled with the significant decrease in hardware costs, has paved the way for advancements in fields requiring substantial computational resources. This progress has been further enhanced by the Web, which serves as a vast repository of information. NLP is one of the fields that has greatly benefited from these developments.

Below are some of the applications that utilize NLP, allowing users to interact with computer systems using natural languages:

- Information Retrieval: The ability to retrieve relevant documents from a repository of documents in response to user requests.
- Machine Translation: The ability to translate text from one language into another.
- Question and Answering: The ability to provide specific answers to given questions.
- Speech Recognition: Recognizing spoken language and transcribing it into written form.

Advancements in research and development of NLP applications have shown rapid progress, especially for languages considered resource-rich, such as English. However, there is a growing need to develop NLP applications for low-resource languages, including those spoken in Ethiopia. This document explores the application of the Chomsky hierarchy in the context of Ethiopian languages, demonstrating how different types of formal languages can be employed to address specific linguistic challenges.

## Type 0 Languages ( Recursively Enumerable Languages)

### Amharic-English Machine Translation

Machine translation between Amharic and English presents unique challenges due to the significant linguistic differences and complexities of Amharic. A crucial effort to address these challenges is detailed in the paper "Extended Parallel Corpus for Amharic-English Machine Translation," which discusses the creation of an extensive parallel corpus. This corpus is a foundational resource that collects and aligns a substantial volume of bilingual texts to support translation tasks. By incorporating a large dataset, the system can better capture the nuances of both languages, thereby improving translation accuracy.

The necessity of formal language and automata theory (FLAT) concepts in this context is evident. Unrestricted grammars are employed to handle the complex syntactic structures inherent in Amharic. These grammars facilitate the representation of intricate linguistic patterns that simpler models might overlook. Additionally, probabilistic models are utilized to manage linguistic variations and predict the most likely translations, enhancing the reliability of the machine translation system.

Normalization techniques play a crucial role in refining these translation systems. The paper "The Effect of Normalization for Bi-directional Amharic-English Neural Machine Translation" explores how normalization impacts translation performance. Text normalization reduces variability in the input data, making it easier for neural network models to learn effective translation mappings. This process involves standardizing the text to minimize differences in spelling, formatting, and other inconsistencies, which can significantly affect translation accuracy.

Improving Amharic-English machine translation involves several key steps:

- **Development of a Comprehensive Parallel Corpus:** Collecting and aligning bilingual texts to create an extensive dataset that forms the backbone of the translation model.

- **Application of Unrestricted Grammars:** Utilizing these to represent and manage the complex syntactic structures of Amharic.
- **Implementation of Probabilistic Models:** Employing these models to predict translations based on linguistic data and manage variations effectively.
- **Normalization Techniques:** Standardizing text to reduce variability and improve the learning efficiency of neural network models.

By integrating these methodologies, the system can significantly enhance the accuracy and reliability of Amharic-English translations.

### **Extended Parallel Corpus for Amharic-English Machine Translation**

The paper "Extended Parallel Corpus for Amharic-English Machine Translation" submitted by Andargachew Mekonnen Gezmu, Andreas Nürnberger, Tesfaye Bayu Bati explores the development and utilization of a comprehensive parallel corpus aimed at enhancing machine translation between Amharic and English. The authors embarked on compiling a substantial corpus by aggregating and aligning texts from diverse sources, such as news articles, religious texts, and official documents. This endeavor addresses the significant resource gap for Amharic, a language with relatively limited computational resources.

The study evaluated various neural machine translation (NMT) models trained on the newly compiled corpus. Among the models tested, Transformer-based architectures were highlighted for their superior performance, particularly in handling low-resource languages like Amharic. The effectiveness of these models was assessed using standard evaluation metrics such as BLEU scores, which demonstrated significant improvements in translation quality compared to previous Amharic-English translation efforts.

## **The Effect of Normalization for Bi-directional Amharic-English Neural Machine Translation**

The paper "The Effect of Normalization for Bi-directional Amharic-English Neural Machine Translation" submitted by Tadesse Destaw Belay, Atnafu Lambebo Tonja, Abinew Ali Ayele and others mainly investigates on how normalization techniques impact the performance of neural machine translation (NMT) systems when translating between Amharic and English. The study is motivated by the unique linguistic characteristics of Amharic, such as its complex script and the presence of homophones, which pose significant challenges for machine translation systems.

The researchers explored various normalization strategies to preprocess the Amharic text before feeding it into the NMT models. These strategies included removing punctuation, normalizing homophones, and applying script normalization techniques to handle different representations of the same characters. The goal was to determine which normalization techniques could enhance the translation quality by reducing the complexity of the input text.

Their experiments showed that normalization significantly improves the translation accuracy for both directions (Amharic to English and English to Amharic). The study used the BLEU (Bilingual Evaluation Understudy) score as a metric to evaluate the performance of the translations. They found that homophone normalization had a particularly positive effect on the translation quality, as it reduced ambiguity and helped the model better understand the semantic context of the sentences.

## Type 1 Languages (Context-Sensitive Languages)

Application Area: Morphological analysis and generation for Ethiopian languages

### Main Functionalities of Morphological Analysis and Generation

#### ❖ Morphological Analysis:

- Determining a word's minimal units, or morphemes.
- Identifying the grammatical characteristics (such as case, gender, and tense) connected to each morpheme.
- Managing intricate derivations and inflections.
- Putting words into grammatical categories (such as noun, verb, and adjective) according to their morphology and syntactic context.

#### ❖ Morphological Generation

- Constructing from morphemes legitimate word forms.
- Word inflecting according to context (e.g., noun declensions, verb conjugations).
  - o Guaranteeing grammatical accuracy.
- Adding affixes to already-existing words to alter their meaning or grammatical classification in order to create new words.

### Why FLAT Concepts Are Needed?

#### i. Complex Morphology and Word Forms:

- Ethiopian languages, including Amharic, exhibit rich morphological complexity.
- FLAT provides a theoretical framework to understand and model this complexity.
- Morphological rules involve intricate interactions between morphemes (minimal meaning-bearing units).

#### Example:

As an illustration, take the Amharic word "ሰላም" (selam), which means "peace." An examination of morphology reveals:

Stem: “ሰላ” (selā)

Suffix: “ም” (m)

Grammatical function: Noun (peace)

## ii. Expressive Power of Formal Grammars:

- Context-Free Grammars (CFGs): Used to explain how words and sentences are put together.
- Word categories (such as nouns and verbs) are represented by non-terminal symbols.
- Morphological patterns (such as affixation and inflection) are captured by CFG rules.

Example:

CFG rule:

Noun → Stem + Suffix

This rule generates valid Amharic nouns by combining stems and suffixes.

## Approaches and Examples

- **Finite-State Transducers (FST) for Creation and Analysis of Amharic Morphology:**

Algorithm: Create finite-state transducers that encode Amharic morphological rules, enabling effective text processing and word form generation. Example: Putting in place an FST-based system to examine and produce Amharic verb, noun, and adjective inflected forms

- **Morphological Rule-Based Systems for Oromo Tokenization: Algorithm:**

Create rule-based systems that tokenize Oromo text by recognizing word and morpheme boundaries using formal grammars and regular expressions. Example: Creating an Oromo text rule-based tokenizer that divides words into morphemes in accordance with grammatical rules and patterns.

- **Morpheme-based Approaches for Tigrinya Stemming:** Stemming is the process of eliminating affixes (prefixes, suffixes, and infixes) to determine a word's base or root form. Algorithm: Use FLAT concepts to model morphological processes at the morpheme level, with a focus on directly analysing and producing morphemes in Tigrinya text. For Tigrinya, an example would be to create a morpheme-based stemming algorithm that finds and eliminates affixes to extract word roots or stems.



## Type 2 Languages (Context-Free Languages )

### Context-Free Languages and Amharic Part-of-Speech Tagging

#### Importance of PoS Tagging in NLP


Part-of-Speech (PoS) tagging, the process of categorizing words in a sentence as nouns, verbs, adjectives, etc., is crucial in natural language processing (NLP). It provides essential syntactic information for advanced NLP tasks like parsing, machine translation, and sentiment analysis. For Amharic, PoS tagging is especially important due to its rich morphology and complex syntax. Amharic, a Semitic language, uses various affixes and root patterns that affect word forms and functions. Accurate PoS tagging is vital for disambiguating these forms and enhancing the performance of applications such as machine translation and information retrieval.

#### Methodology

Context-free grammars (CFGs) are often used in PoS tagging to define syntactic structure rules, helping systems understand and process sentences correctly. CFGs consist of production rules that describe how words and phrases combine to form valid sentences, which are crucial for generating parse trees that represent sentence structures.

For Amharic, CFGs must consider its morphological complexity. For example, CFGs for Amharic might include rules for how verb roots combine with tense and aspect markers or how nouns combine with gender and number markers. Examples of CFG rules for Amharic PoS tagging include:

- 'S -> NP VP' (A sentence consists of a noun phrase followed by a verb phrase)
- 'NP -> Det N' (A noun phrase consists of a determiner followed by a noun)
- 'VP -> V NP' (A verb phrase consists of a verb followed by a noun phrase)



These rules help parsers generate all possible valid sentence structures, from which the most likely one can be selected based on context.

## Technological Implementations

Several tools and algorithms have been developed for Amharic PoS tagging, using both traditional and modern approaches:

- **CRFTagger:** This tool uses Conditional Random Fields (CRF) for sequence labeling tasks like PoS tagging, predicting PoS tags for each word based on the context provided by surrounding words.
- **NLTK and spaCy:** These popular NLP libraries have been adapted for Amharic. They offer frameworks for training PoS taggers on annotated corpora using machine learning techniques to improve accuracy.
- **Deep Learning Models:** Recently, approaches like Long Short-Term Memory (LSTM) networks and Transformer-based models have been applied to Amharic PoS tagging. These models capture long-range dependencies and morphological patterns more effectively than traditional methods.

## Accuracy and Performance

The accuracy of PoS tagging tools for Amharic depends on the training corpus quality and size, model complexity, and specific linguistic features of the language. Traditional models like CRFTagger generally achieve reasonable accuracy but may struggle with rare or ambiguous word forms. Deep learning models, though more powerful, require larger datasets and more computational resources.

A comparative study of different PoS tagging approaches for Amharic found that deep learning models, especially those based on the Transformer architecture, achieved the highest accuracy.

However, these models also demanded extensive training data and computational power. The study emphasized the importance of high-quality, annotated corpora and the need for ongoing research to optimize models for Amharic's unique linguistic features.

## Type 3 Languages (Regular Languages)

### Application Area: Text Summarization for Ethiopian Languages

#### Importance of Text Summarization in NLP

Text summarization is an NLP task that involves creating a concise and coherent summary of a longer text while retaining its essential information and meaning. This task is vital for managing and understanding large volumes of information, making it easier to digest and extract key insights from extensive documents, news articles, and other textual data. In the context of Ethiopian languages, text summarization is particularly useful for enhancing information accessibility and enabling efficient content consumption, given the linguistic diversity and the need for localized content.

#### Methodology

Text summarization can be approached using various techniques, including extractive and abstractive methods. Extractive summarization involves selecting key sentences from the original text, while abstractive summarization generates new sentences that convey the main ideas of the text.

For Ethiopian languages, context-free grammars (CFGs) can be employed to define syntactic rules that help identify and structure the important parts of a text. CFGs consist of production rules that describe how words and phrases can be combined, facilitating the generation of coherent summaries.

Examples of CFG rules for Ethiopian languages text summarization might include:

- 'Summary -> Sentence Summary' (A summary can be composed of a sentence followed by another summary segment)
- 'Sentence -> NP VP' (A sentence consists of a noun phrase followed by a verb phrase)
- 'NP -> Det N' (A noun phrase consists of a determiner followed by a noun)

These rules help the summarization system identify the main components of a text and organize them into a coherent summary.

## Technological Implementations

Several tools and algorithms have been developed for text summarization in Ethiopian languages, utilizing both traditional and modern approaches:

- **TF-IDF and LexRank:** Traditional extractive summarization methods like Term Frequency-Inverse Document Frequency (TF-IDF) and LexRank, which rank sentences based on their importance and relevance.
- **NLTK and spaCy:** These NLP libraries have been adapted to support Ethiopian languages. They provide frameworks for implementing summarization algorithms, leveraging annotated corpora and rule-based techniques.
- **Deep Learning Models:** Recently, neural network-based approaches such as sequence-to-sequence (Seq2Seq) models and Transformer-based models like BERT and GPT have been applied to text summarization. These models can generate more coherent and contextually accurate summaries by capturing long-range dependencies and semantic patterns.

## Accuracy and Performance

The accuracy of text summarization tools for Ethiopian languages depends on the quality of the training data, the complexity of the model, and the linguistic characteristics of the language. Traditional extractive methods like TF-IDF and LexRank are straightforward and efficient but may not always produce the most coherent summaries. Deep learning models, although more powerful, require extensive training data and computational resources.

A comparative study of different summarization approaches for Ethiopian languages found that deep learning models, especially Transformer-based architectures, achieved the highest coherence and informativeness in their summaries. However, these models also required large annotated datasets and significant computational power. The study highlighted the importance of developing high-quality, language-specific datasets and ongoing research to optimize summarization models for Ethiopian languages.

## Text Summarization for Afaan Oromo

Text summarization for Afaan Oromo is a nascent area of research. Initial efforts have focused on creating summarization systems for news articles and educational content. Researchers have employed extractive methods and simple rule-based approaches to generate summaries. Despite

achieving promising results, the lack of comprehensive linguistic resources and annotated datasets poses significant challenges.

## Exploring Amharic Text Summarization

For Amharic, text summarization research has primarily focused on summarizing news articles and social media content. Early studies employed extractive methods, leveraging TF-IDF and LexRank to identify key sentences. Recent efforts have explored neural network-based approaches, such as using Seq2Seq models and Transformer-based architectures, to generate more coherent and contextually accurate summaries.

### Text Summarization Approaches

- Extractive Methods: Using techniques like TF-IDF and LexRank to identify and rank key sentences from the text.
- Abstractive Methods: Generating new sentences that capture the main ideas of the text using Seq2Seq models and Transformer-based architectures.
- Hybrid Methods: Combining extractive and abstractive techniques to enhance the quality and coherence of summaries.

By integrating these methodologies, text summarization systems for Ethiopian languages can significantly improve, leading to more accurate, coherent, and informative summaries across various applications.

## Conclusion

In this assignment, the four types of Chomsky hierarchy—Type 0, Type 1, Type 2, and Type 3—are described in detail. The document also lists the application areas of these types in the Ethiopian context. Here is a short summary of the document:

Recursively Enumerable Languages (Type 0) are languages for which there exists a Turing machine that can enumerate all the strings belonging to the language. An application of Type 0 languages in the Ethiopian context is Amharic-English Machine Translation. This involves creating an extensive parallel corpus and utilizing unrestricted grammars to handle the complex syntactic structures of Amharic, enhancing translation accuracy through probabilistic models and normalization techniques.

Context-sensitive languages (Type 1) represent a crucial class within formal language theory, characterized by grammar rules that consider the context surrounding symbols for production. In the Ethiopian context, an application area is Morphological Analysis and Generation for Ethiopian languages such as Amharic and Oromo. This involves determining and generating valid word forms from morphemes, managing intricate derivations and inflections using finite-state transducers and rule-based systems.

Context-free languages (Type 2) constitute a fundamental category within formal language theory, distinguished by grammar rules where productions are solely determined by non-terminal symbols, independent of context. For Ethiopian languages, an application is Amharic Part-of-Speech Tagging. This is essential for syntactic parsing and involves using context-free grammars to define syntactic structures, with tools like CRFTagger, NLTK, spaCy, and deep learning models enhancing tagging accuracy.

Regular Languages (Type 3) represent a foundational class within formal language theory, characterized by simple and efficient grammar rules that can be expressed through regular expressions or finite automata. An application in the Ethiopian context is Text Summarization for Ethiopian languages. This task involves creating concise summaries of texts using both extractive and abstractive methods, leveraging context-free grammars and modern neural network-based approaches to improve coherence and contextual accuracy.

Each application demonstrates the practical utility of formal language theory in addressing the unique linguistic features and computational challenges associated with Ethiopian languages.

## Reference

The following references are used to do this project. You can get this research papers in the following GitHub repository. <https://github.com/EthioNLP/Ethiopian-Language-Survey>

- AMHARIC NAMED ENTITY RECOGNITION USING A HYBRID  
BY: MIKIYAS TADELE BELAY October 2010
- Named Entity Recognition for Afan Oromo  
By: Mandefro Legesse Kejela August 2014
- Machine Learning Approaches for Amharic Parts-of-speech Tagging  
By Ibrahim Gashaw H L Shashirekha
- Sentiment Analysis of Afaan Oromoo Facebook Media Using Deep Learning Approach  
By Megersa Oljira Rase
- Part of Speech Tagging for Wolaita Language using Transformation-Based Learning  
By Birhanesh Fikre Shirko
- Exploring Amharic Sentiment Analysis from Social Media Texts: Building Annotation Tools and Classification Models  
By Seid Muhie Yimam<sup>1</sup> , Hizkiel Mitiku Alemayehu, Abinew Ali Ayele<sup>1</sup>