
# ST. MARY'S UNIVERSITY
# SCHOOL OF GRADUATE STUDIES

## ENGLISH-WOLAYTTA MACHINE TRANSLATION USING STATISTICAL APPROACH

**BY**
**MELAKU MARA**

**JULY 2018**
**ADDIS ABABA, ETHIOPIA**


# English-Wolaytta Machine Translation Using Statistical Approach

**A Thesis Presented**

**by**

**Melaku Mara**

**to**

**The Faculty of Informatics**

**of**

**St. Mary's University**

**In Partial Fulfillment of the Requirements**
**for the Degree of Master of Science**

**in**

**Computer Science**

**July, 2018**

# ACCEPTANCE

# ENGLISH-WOLAYTTA MACHINE TRANSLATION USING STATISTICAL APPROACH

**BY**

**MELAKU MARA**

**Accepted by the Faculty of Informatics, St. Mary's University, in partial fulfillment of the requirements for the degree of Master of Science in Computer Science**

**Thesis Examination Committee:**

_____

**Internal Examiner**

_____

**External Examiner**

_____

**Dean, Faculty of Informatics**

**July 2018**

## DECLARATION

I declare that this research is my original work and has not been presented for a degree in any university, and that all sources of material used for the research have been properly acknowledged.

_____ _____

Full Name of Student

_____

Signature

Addis Ababa
Ethiopia

This thesis has been submitted for examination with my approval as advisor.

_____ _____

Full Name of Advisor

_____

Signature

Addis Ababa
Ethiopia

July 2018

**Acknowledgment**

Above all I would like to thank the almighty God, who gave me the opportunity and strength to achieve whatever I have achieved so far. I would like to express my gratitude to all the people who supported and accompanied me during the progress of this thesis.

First, I would like to express my deep-felt gratitude to my advisor, Ato Michael Melese, whose excellent and enduring support shaped this work considerably and made the process of creating this thesis an invaluable learning experience.

I want to thank Wolaytta Sodo University language department and Wolaytta Zone education office who helped me in the process of data collection of Wolaytta language.

Special thanks to my family; their endless motivation and unconditional love have been influential in whatever I have achieved so far.

Finally, all my friends deserve special thanks. They are the ones who are always there to spend time with, and share my joys and sorrows.

# Abstract

Machine translation is a technology for the automatic translation of text or speech from one natural language to another. Since there is a need for translation of sentences between English-Wolaytta language to make available the English documents in Wolaytta language and minimize the language barrier. Thus, this study in the development of a English-Wolaytta machine translation system using statistical approach.

In order to achieve the objective of this research work, 30,000 bilingual corpus is collected from spiritual domain and 39,893 monolingual corpus from different sources. And also prepared in a format suitable for use in the development process (normalization, tokenization, lower-case and clean) and classified as training, tunning and testing set. Aligned parallel sentences manually and used freely available tools for the different purposes such as SRILM toolkit for language model, MGIZA++ align the corpus at word level by using IBM models (1-5), Decoding has been done using Moses, and Ubuntu operating system which is suitable for Moses environment has been used. In addition, unsupervised morpheme segmentation tool Morfessor is used for segmentation of Wolaytta text.

The experiments were taken separately, one for the unsegmented and the other for segmented corpus. The parallel sentences divided by 5,000, 10,000, 15,000, 20,000, 25,000 and 30,000. The unsegmented corpus performs BLEU score of 4.91%, 6.30%, 7.21%, 7.60%, 7.96% and 8.46% used the above divided parallel sentences. The segmented corpus performs BLEU score of 9.83%, 11.38%, 12.70%, 12.77%, 12.93% and 13.21% used the above divided parallel sentences. Its performance improved by increased the size of the corpus and segmented parallel sentences.

Base on the experiments done, the researcher observed that there will be a better performance when increase the size of the corpus and morphological segmentation. Therefore future research should focus to further improve the performance of the system increase the size of the corpus and morphological segmentation.