

# Development of Ethiopic OCR Technology for an Improved Information Processing

The Second PID Workshop on ICT4D  
March 14, 2007

Stockholm, Sweden

Yaregal Assabie  
Halmstad University  
Halmstad, Sweden



# Presentation Outline

- Introduction
- Ethiopic Writing System
- OCR Technology
- Applications in Information Processing

# Introduction

## Language and Technology

- ❖ Information and Communication Technology (ICT) is a widely understood as a key for development.
- ❖ The speed of information processing has a direct effect on the speed of development.
- ❖ The written form of language is an important method of information exchange in human communication.
- ❖ The invention of printing machine by Gutenberg (in the 15<sup>th</sup> C) and computers & electronic printers (in the current era) are milestones in the history of written languages.
- ❖ Recently, written languages have acquired a new dimension in their history—*the reverse process of printing*.

# Ethiopic Writing System

## The Ethiopic Alphabet

- ❖ Conveniently written in a tabular format of 7 columns representing sounds

1 <sup>st</sup> (ä) order	2 <sup>nd</sup> (u) order	3 <sup>rd</sup> (i) order	4 <sup>th</sup> (a) order	5 <sup>th</sup> (e) order	6 <sup>th</sup> (ə) order	7 <sup>th</sup> (o) order
ሀ hä	ሁ hu	ሂ hi	ሃ ha	ሄ he	ህ hə	ሆ ho
ለ lä	ሉ lu	ሊ li	ላ la	ሌ le	ል lə	ሎ lo
ሐ hä	ሑ hu	ሒ hi	ሓ ha	ሔ he	ሐ hə	ሐ ho
መ mä	ሙ mu	ሚ mi	ማ ma	ሜ me	ሞ mə	ሞ mo
ሠ sä	ሡ su	ሢ si	ሣ sa	ሤ se	ሥ sə	ሦ so
ረ rä	ሩ ru	ሪ ri	ራ ra	ራ re	ር rə	ሮ ro
ሰ sä	ሱ su	ሲ si	ሳ sa	ሴ se	ሰ sə	ሰ so
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
ሸ šä	ሹ šu	ሺ ši	ሻ ša	ሼ še	ሽ šə	ሾ šo
ጸ pä	ጹ pu	ጺ pi	ጻ pa	ጼ pe	ጽ pə	ጾ po
ፈ fä	ፉ fu	ፊ fi	ፋ fa	ፌ fe	ፍ fə	ፎ fo
ፐ pä	ፑ pu	ፒ pi	ፓ pa	ፔ pe	ፕ pə	ፖ po

# Ethiopic Writing System

## Literature in Ethiopia

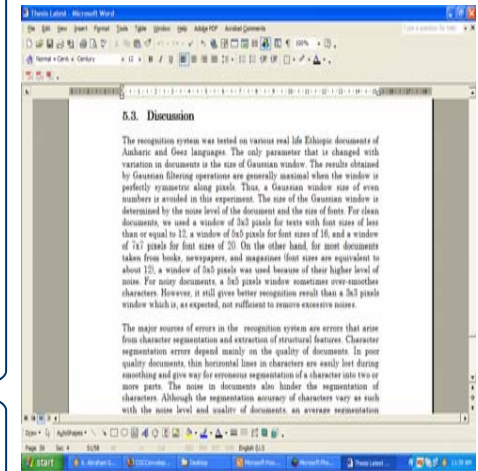
- ❖ Ethiopic alphabet has been in use since the 5<sup>th</sup> BC.
- ❖ Literature has flourished in Ethiopia with the introduction of Christianity in the country in the 4<sup>th</sup> century.
  - Geez is used as the liturgical language of Ethiopian Orthodox Church
  - Most of the ancient documents are written in Geez
- ❖ Currently, the alphabet is used widely by Amharic which is the official language of Ethiopia (pop. over 75 million).
  - Amharic has grown to become the second most spoken Semitic language, next to Arabic.
  - Recent documents are usually written in Amharic.

# OCR Technology

## What is OCR (Optical Character Recognition) Technology?



OCR System



# OCR Technology

## Types of OCR Systems

### ❖ Online OCR

- An OCR system that recognizes text at the time of writing
- The text is captured by an electronic tablet with a special pen



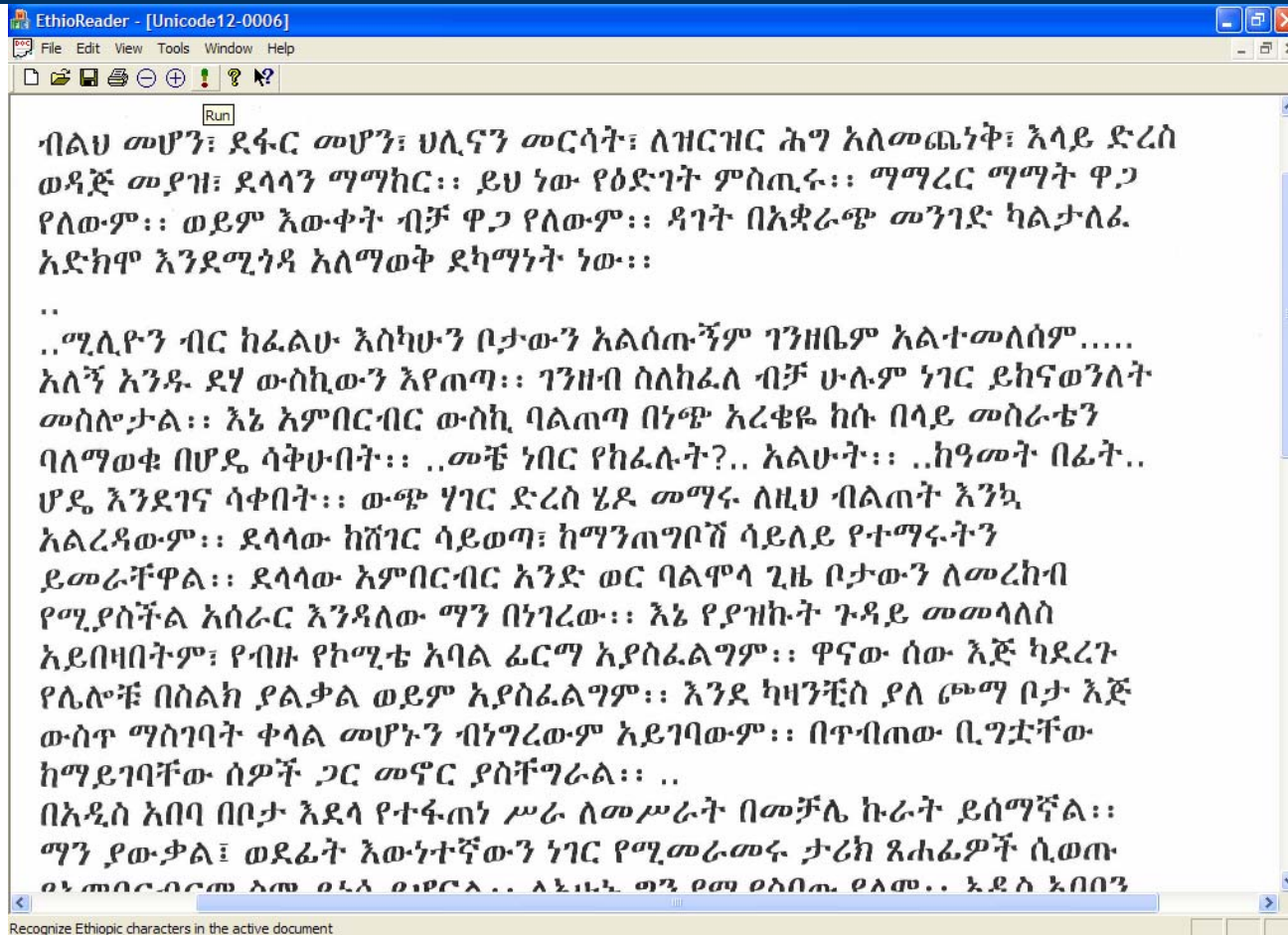
### ❖ Offline OCR

- An OCR system that recognizes text after the writing is completed
- The text is captured by a scanner



# OCR Technology

## Ethiopic OCR (...under development)



# Applications in Information Processing

- ❖ Saves time (and man power) for data entry
  - converts existing documents into electronic copy faster than human beings
  - Online OCR systems are more convenient and natural than the keyboard
- ❖ Storage and Retrieval
  - Used to preserve historical documents
  - Compact space and convenient access
  - Used to curb scarcity of documents (e.g., books in schools)
  - Reduces the cost of production of documents
- ❖ Office automation
  - Mail sorting
  - Automatic banking system (checks, credit cards, etc...)
  - Form processing
- ❖ Speech Application
  - Helps the blind and illiterate to 'read' documents
  - Helps for automatic machine translation of documents



**Thank you**