Indigenous Scripts of African Languages

Million Meshesha, C. V. Jawahar
Center for Visual Information Technology,
International Institute of Information Technology,
Gachibowli, Hyderabad - 500 019
jawahar@iiit.net

Abstract

*In Africa there are a number of languages. Some of these languages have their own indigenous scripts. In this paper we present script analysis for the indigenous scripts of African languages, with particular emphasis to Amharic language. Amharic is the official and working language of Ethiopia, which has its own writing system. This is the first attempt to analyze scripts of African language to ease document analysis and understanding. We believe researchers will continue exploring African indigenous languages and their scripts to be part of the revolving information technology for local development. W*e also highlighted problems related to the scripts that have bearings on Amharic document analysis and understanding. Especially availability of large number of characters and similarity among characters makes the task of document understanding research much tougher than that of most Latin-based scripts.

1. Introduction

Africa is the second largest continent in the world, next to Asia, covering about one-fifth of the total surface area of the Earth. Africa is not one country with a uniform culture. Africa has a very rich diversity on culture, history and languages (The Columbia Encyclopedia, 2001). Document analysis and understanding research has not yet addressed the indigenous African scripts, as much they deserve.

There are more than 55 independent countries (including islands) in Africa with approximately 800 million people and over 800 ethnic groups. Its many languages testify to the vast diversity of the African people. In all, more than 2,500 languages (including regional dialects) are spoken in Africa, an estimated one third of the world's total.

The languages of Africa are grouped into four language families: Niger-Congo, Nilo-Saharan, Khoisan, and Afro-Asiatic (Microsoft Encarta Online Encyclopedia, 2006). The Niger-Congo Family is the largest African languages family comprising over 55 percent of the total languages spoken. This family includes several subfamilies, including Kordofanian, Mande, and Atlantic-Congo, which is further sub-categorized into subfamilies including Benue-Congo, Atlantic, Gur, Kwa, and Ijoid. In the Benue-Congo subfamily, a relationship exists among most of the languages of southern and central Africa. These languages have become widely known as Bantu. Some of Bantu languages are Zulu and Xhosa in South Africa; Makua in Mozambique; Nyanja in Malawi; Shona in Zimbabwe; Bemba in Zambia; Swahili and Sukuma in Tanzania; Kikuyu in Kenya; Fang and Bulu in Cameroon, Yoruba, Igbo, and Efik in Nigeria.

From south-eastern Nigeria to Liberia are found the languages of the Kwa branch. This branch includes such important languages as Ewe in Togo and Ghana; Akan in Ghana; and Anyin in Cote d'Ivoire. From Liberia to the desert north of Dakar, are several languages of the Atlantic branch. These include Themne in Sierra Leone, Wolof in Dakar, and Fulani spoken in Guinea, eastern Nigeria (Nigerian Fulfulde), and Senegal (Pulaar). Speakers of languages of the Mande branch inhabit in the West Africa. One Mande language, known as Bambara, is spoken by up to 3 million people in Senegal, Mali, Guinea and Cote d'Ivoire. Other important Mande languages are Mende in Sierra Leone and Kpelle in Liberia. The Mande languages are believed to be the oldest offshoots of the parent Niger-Congo language spoken more than 5,000 years ago.

The second group of languages is Nilo-Saharan. Around 200 Nilo-Saharan languages are found in a broken chain from the great bend of the Niger River in West Africa to Ethiopia, throughout most of the upper Nile valley, and in parts of Uganda and Kenya. The western most branch of this family is Songhai spoken in Mali and Niger. The Saharan branch of this family includes languages spoken in

Nigeria, Chad and Libya. Along the River Nile near Egypt and in south-west are the Nubian languages, Chari-Nile languages spoken by about 1 million people. The Nubian alphabet was derived from that of the Coptic language. In Sudan, Uganda and Kenya a group of Nilotic languages such as Dinka, Nuer, Shilluk, and Acholi (or Luo) also belongs to this branch of family.

The third group of languages is the Khoisan family. The Khoisan languages comprise the smallest language family in Africa, with only around 200,000 speakers of the 30 languages altogether. Most of these languages are spoken by the peoples of southern Africa; the largest of them is Nama. In Tanzania there are two other representatives of this family: Sandawe and Hadza. The Khoisan languages are best known for the unusual click consonants characteristic of most of them; in some Khoisan languages nearly every word begins with a click. Some of the Khoisan languages have a system of grammatical gender, which is found elsewhere in Africa only in the Afro-Asiatic family.

The final group of languages constitutes Afro-Asiatic family. It has many sub branches. Almost 400 Afro-Asiatic languages constitute the most important group of languages spoken in northern Africa. The Semitic branch of the family includes languages spoken in Asia as well as in Africa. The many Arabic languages, the leading members of this branch, are the major languages of North Africa (Tunisia, Morocco, etc.) and East Africa (Sudan, Ethiopia, etc.). An Ethiopian language, Amharic is grouped under Afro-Asiatic language. Other Semitic languages spoken in East Africa include Tigre and Tigrigna in Ethiopia and Eritrea, respectively. Languages of the Berber branch of the Afro-Asiatic family are spoken by a substantial portion of the population in Morocco, Algeria, and Tunisia. The Cushitic branch, confined to Ethiopia, Eritrea, Somalia, Kenya, Sudan, and Tanzania, includes such major languages as Oromo and Somali.

1.1 Multilingualism

Africans have traditionally spoken not only their birth tongue but also a local or regional lingua franca, such as Hausa, Swahili, or Arabic, associated with trade. Basically multilingualism is extensive throughout the continent. In this regard, Arabic is a major world language. In the same way, some African languages are important transnational languages which function as lingua francas. Apart from Arabic which is not confined to Africa, the most widely spoken African tongues are Hausa, Swahili and Amharic, both of which are used over wide areas as lingua francas (Wikipedia Encyclopedia, 2006). Hausa has the largest number of speakers. It is spoken by around 39 million people. It is followed by Swahili which is spoken by 35 million speakers and Amharic which has around 34 million speakers.

Native speakers of Hausa are mostly found in Niger and Nigeria, but the language is widely used as a lingua franca in a much larger part of West Africa, including Benin, Burkina Faso, Cameroon, Ghana, Togo, etc.

Swahili is the mother tongue of the Swahili people in the East African coast from Somalia to Mozambique. It is widely used by most countries in East Africa. Swahili is an official language in Tanzania and Kenya. It is also spoken in Uganda, Rwanda, Burundi, Democratic Republic of Congo (DRC), Somalia, Comoros Islands (including Mayotte), Mozambique and Malawi.

Amharic language is mainly spoken in Ethiopia and Eritrea. It is the working language of Ethiopia and the most commonly learned language next to English throughout the country. An indigenous Amharic script is used for writing in the various languages in Ethiopia and Eritrea, including Amharic, Tigre and Tigrigna. Amharic language is also spoken by a number of people living in countries such as Ethiopian Jews in Israel, Egypt, Sweden, etc.

No.	Writing systems	Country of Origin
1	Egyptian writing systems	Egypt
2	Amharic script	Ethiopia
3	Vai script	West Africa
4	Bassa script	Liberia
5	Mende script	Sierra Leone
6	Nsibidi/Nsibiri script	Nigeria and Cameroon
7	Shumom script	Cameroon
8	Meroitic script	Sudan

Table 1: Indigenous scripts of Africa

2. African Indigenous Scripts

There are many languages spoken in Africa; almost an estimated one third of the world's total. Some of these are indigenous languages, while others are installed by conquerors of the past. English, French, Portuguese, Spanish and Arabic are official languages of many of the African countries. Most African languages with a writing system use a modification of the Latin and Arabic scripts. There are also many languages in Africa with their own indigenous scripts that vary considerably in shapes (Mafundikwa, 2000). Some of these scripts are presented in Table 1.

- **Egyptian Writing System**: Egyptian writing system is an ancient pictographic writing now dated to be 3400B.C. It consists of approximately 121 bi-literals, 75 tri-literals, and various determinants and phonetic complements. The bi-literals were individual symbols which expressed two sounds and the tri-literals were individual symbols which express three sounds. Phonetic complements are mono-literals found in front of and/or behind multi-consonantal signs in order to provide clarity and also to complete the meaning of the word. They normally repeat sounds already found in the word, but have no separate sound value.

- **Amharic script**: It is designed as a meaningful and graphic representation of knowledge. Amharic script is a component of the African knowledge systems and one of the signal contributions made by Africans to the world history and cultures. It is created to holistically symbolize and locate the cultural and historical parameters of the Ethiopian people. The System, in its classic state, has a total of 182 syllographs, which are arranged in seven columns, each column containing 26 syllographs. Through time many characters are added to give it the present shape. Detailed discussion is made in Section 3.
- **Mende script**: This script was used by the Mende people of Sierra Leone. It is not only considered a writing system, it is a work of art.
- **Nsibidi script**: Nsibidi is a writing system of the Ejagham people of Nigeria. It is seen on tombstones, secret society buildings, costumes, ritual fans, headdresses, textiles, and in gestures, body and ground painting.
- **Vai Syllabry**: The Vai Syllabry is a writing system used by the Vai people of West Africa since the 20th century. It is one of the many indigenous secret writing systems in Africa.
- **Meroitic Script**: The Meroitic script is very similar to the Egyptian Writing System. It was used by the Meroe people of the Sudan. The system is written from right to left, unlike the Egyptian system which is written both from right to left, left to right, and vertically.
- **Shumom Writing System**: The Shumom people are the people of Cameroon in West Africa. Cameroonians use the Shumom writing systems, perhaps beginning with the hieroglyphics of the Ancient Egyptians writing.
- **Bassa Script**: Bassa is the most commonly spoken languages in Liberia which has its own written script. Bassa script is phonemic rather than syllabic.

There are also other languages in Africa known to have their own written language. These scripts include the Kpelle, Gola, Lorma, Grebo, and Kissi. Most of these scripts have diminished over time, as a result of abandonment.

3. Amharic Language

Ethiopia is located in Eastern Africa with a population of approximately 67 million. Ethiopia is a mosaic of ethnicities with 83 languages and there are more than 200 different dialects spoken. The Ethiopian languages are divided into four major language groups, such as Cushitic, Omotic, Nilo-Saharan and Semitic. Amharic belongs to the Semitic family of languages.

Amharic is written in the unique and ancient Ethiopic script (inherited from Geez, a Semitic language). Amharic is the second most spoken Semitic language in the world, next to Arabic. Amharic is spoken by roughly 30% of the population as a first language, and an additional 20% as a second language, totaling about half of the population. (The Columbia Encyclopedia, 2001). Outside Ethiopia, Amharic is the language of some 2.7 million people living in Egypt, Israel and Sweden, and is spoken in Eritrea. In general, there are more than 34 million speakers of Amharic (The Columbia Encyclopedia, 2001; Ethnologue, 2006).

Amharic is the official and working language of Ethiopia and thus has official status nationwide. It is also the working language of several of the states within the federal system of Ethiopia. It has been the working language of government, the military, and of the Ethiopian Orthodox Church throughout modern times. Accordingly, there is a bulk of printed documents (such as correspondence letters, books, newspapers, and magazines) available in government and private offices, libraries and museums. Digitization of these documents enables to harness already available information technologies to local information needs and developments, thereby, among others, to facilitate indexing and retrieval, to save storage space, and to preserve historical documents.

All these stress the importance and tremendous need for document analysis and understanding systems such as optical character recognition (OCR) systems (OCR systems are used to convert large-scale scanned printed/handwritten text documents into computer understandable format to ease processing and retrieval of information), if at all one is to harness already available information technology to local information needs and developments. The encouraging efforts being made and the advancements registered in the area of application software capable of interfacing with the scripts is also an additional motivation for seeking information technology solution.

Those African languages that use modified scripts of Latin and Arabic language can be integrated to the already available Latin and Arabic document analysis and understanding systems with the same additional language processing modules. It is suffix to mention Commercial OCR packages such as ABBYY FineReader which recognizes texts written in some of the African languages such as Afrikaans, Xhosa and Zulu (South Africa), Kongo (Congo), Swahili (East Africa), Swazi (Swaziland), etc. that uses Latin script. Therefore, researchers in African or else where need to give more emphasis to indigenous African scripts. To the best of our knowledge, this is the first work that reports the challenges toward the recognition of indigenous African scripts and a possible solution for Amharic script.

Since script characteristics have considerable effect in document analysis and understanding technologies especially for OCR system development, in the following sections a detailed analysis of Amharic scripts is presented.

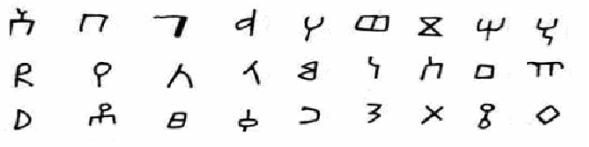


Figure 1: Sabean writing system

4. Amharic Writing System

Amharic language started to be used in Ethiopia as early as the 14th century, although the language of literature at that time and until the 19th century was Geez, from which Amharic evolved. Geez is now mainly used in Ethiopian Orthodox Church as liturgical language.

Amharic has its own writing system called FIDEL. The oldest Amharic inscription was derived from the Sabean writing (shown in Figure 1) which has had twenty-seven symbols in its unvocalized shape. But later Geez pursued the most original course taken Semitic script in denoting vowels by a variety of changes in the structure of the consonantal symbol. Vowels have thus become an integral part of Amharic writing which now assumed the character of a syllabary.

Amharic script has also evolved through the centuries and has undergone changes in shape and number of symbols (by dropping some of the characters and adding others as depicted in Figure 2). The need for such transformation through time is due to (Ullendorff, 1973):

- the tendency towards round forms,
- the changed direction of writing,
- the turn of some characters by ninety degrees.

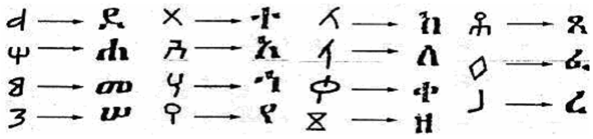


Figure 2: Some of the transformation in Amharic scripts

.

Undergoing many transformations through the ages, the Amharic script has now 33 core characters each of which occurs in seven orders (one basic form and six non-basic forms) (as shown in Figure 3). The seven orders represent syllable combinations consisting of a consonant and following vowel (Bender, 1976). This is why the Amharic writing system is often called a syllabary rather than an alphabet. The non-basic forms are derived from the basic forms by more-or-less regular modifications. Other symbols representing labialization, numerals, and punctuation marks are also available. These bring the total number of Amharic scripts to 310. Table 2 shows the number of characters in each group. The availability of large number of Amharic characters in the writing system is, among others, a great challenge in the development of document analysis and understanding systems, such as OCR system for Amharic language as compared to Latin scripts.

No.	Type of Amharic Characters	Number of Characters
1	Core characters	231
2	Labialized characters	51
3	Punctuation marks	8
4	Numerals	20
	Total	310

Table 2: Total number of symbols in Amharic writing system

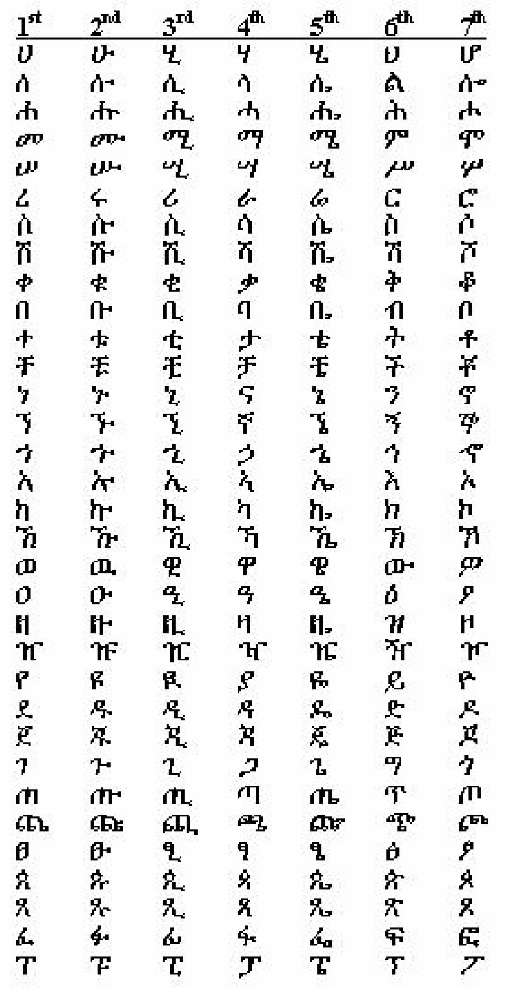
Since Amharic writing system does not have a symbol for zero, negative, decimal point, and mathematical operators, the Hindu-Arabic numerals and Latin mathematical operators are used for computational purpose.

5. Amharic Script Analysis

Amharic scripts formation has certain notable features as discussed below.

Shape similarities among characters: As pointed out by Bender (Bender, 1976), the shape of many Amharic characters shows similarities with few distinctions among them, for example, ዶ and ጆ , ተ and ቸ, ከ and ኸ.

Structural relation: Many basic characters are also clearly related in graphical structure, for instance, ነ and ከ, ረ and ሬ, etc.



1st	2nd	3rd	4th	5th	6th	7th
ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ
ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ
ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ
ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ
ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ
የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ
ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ
ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ
ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ
ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ

Figure 3: List of Amharic alphabets (called FIDEL). The figure Shows Amharic scripts with their seven orders. The orders are shown at the top. The first order indicates list of basic characters and other orders are vowels derived from them.

Shape differences: There are also remarkable differences in shapes among the basic characters. Consider ሀ and በ (both are open in one side but in opposite direction), መ and ወ (both are formed from two loops but differ in the connection of the loops), ሠ and ጠ (both have three legs which end in different direction), etc.

Vowel formation: An interesting peculiarity of the Amharic writing system is the way vowels are formed. Vowels are written with small appendages to the consonant letters, with modifications of their shapes. This method of writing vowels is similar to that of Indic alphabets. Specifically speaking, vowels are derived from consonants in two ways. Some vowels (such as the fourth and seventh orders) take a modified shape of the base character by shortening/ lengthening one of its main strokes. On the other hand, adding small appendages, such as strokes, loops to the right, left, top or bottom of each base character forms the remaining vowels (like second, third and fifth orders). As shown in Figure 3, the second, third, and fifth orders are formed (with few exceptions) according to patterns of great regularity, while the fourth, sixth and seventh orders are highly irregular. For instance, the second order is mostly constructed by adding a horizontal stroke at the middle of the right side of the base character; where as, the sixth order is formed by adding a stroke, loop or other forms in either side of the base character.

For instance, the second order is mostly constructed by adding a horizontal stroke at the middle of the right side of the base character; where as, the sixth order is formed by adding a stroke, loop or other forms in either side of the base character.

Size and width differences: Amharic characters can differ in size both vertically and horizontally. There are very short characters (such as ሬ, ሠ , ጠ) and there are very long characters (such as ኟ, ዠ, ኸ). There is also noticeable variance in width, for instance between ገ , ጢ, and ፊ. . As compared to Latin scripts, the concepts of upper case and lower-case letters are absent in Amharic writing system.

Other features: As compared to Latin scripts, the concepts of upper-case and lower-case letters are absent in Amharic writing system. On the other hand, like English, the writing mode is from left to right and top-to-bottom. Words are separated with blank space/Ethiopic two-dots, sentences end with Ethiopic four-dots, and paragraphs with recognized horizontal space.

6. Conclusions

There are a number of indigenous scripts in African in which a pile of paper-based information written and available in government and private organizations. Document analysis and understanding research has not yet addressed the indigenous African scripts, as much they deserve. So the necessity of conversion of such paper-based data (including poor quality ones) to computer readable format is greatly required for easy indexing, searching and retrieval with the help of recent advancement in information technology.

This paper discusses issues related to African language scripts with specific emphasis to Amharic scripts. The availability of large number of Amharic characters in the writing system, existence of similarity between character shapes are, among others, a great challenge in the development of document analysis and understanding systems for African scripts as compared to Latin scripts.

Our future work will explore other African scripts, in addition to designing methodologies for development of document analysis and understanding systems for the scripts.

Reference

Bender, M.L. 1976. *Language in Ethiopia.* London: Oxford University Press

Ethnologue. 2006. *Languages of the World*. 14th ed., at http://www.ethnologue.com/

Mafundikwa, S.. 2000. *African Alphabets*. Harare, Zimbabwe, available at http://www.ziva.org.zw/afrikan.htm (accessed May 2006).

Microsoft Encarta Online Encyclopedia. 2006. *African Languages*. at http://uk.encarta.msn.com (accessed November 2006).)

The Columbia Encyclopedia. 2001. *Africa*, 6th ed., available at http://www.bartleby.com/65/af/Africa.html (accessed May 2006).

Ullendorff, E. 1973. *The Ethiopians: An Introduction to the Country and People.* 3rd ed., London: Oxford University Press

Wikipedia Encyclopedia. 2006. at: http://en.wikipedia.org/wiki. (accessed December 2006).