Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 442 – 450

Procedia
Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

The Making of Lingala Corpus: An Under-resourced Language and the Internet

Bienvenu Sene-Mongaba*

Université Pédagogique Nationale, Kinshasa, DR Congo

Abstract

Lingala is now the most widespread language in Congo. The Internet provides a great amount of data. This paper has attempted to elucidate the issues that are involved with building a corpus for an under-resourced language where access to internet texts is difficult. To extract Lingala text from a mass of French text, it has been necessary to go through a process of selection by seed words list. The raw corpus is composed of 6,080,426 tokens. I have intervened on the data from internet sources by standardizing the spelling. This standardized corpus is stored separately from the raw corpus.

Keywords: Lingala; Congo; Unitex; NLP; spelling standardization; corpus cleaning; under-resourced language; African languages

1. Introduction

Lingala is now the most widespread language of daily communication both in the cities of Kinshasa and Brazzaville, which are respectively the capital of DR Congo and the capital of Republic of Congo. It has been spreading much more rapidly than its national counterparts (i.e., Kikongo, Kiswahili, and Ciluba) in the rest of both countries and among the Congolese Diaspora. Around 10 million people use Lingala as their first language, 20 million as their second language and more than 50 million use it as one of their languages of daily communication. However, like in most African countries, former colonial languages continue to be used as languages of instruction and languages of administration. This is the case, for example, of Kinshasa students, who speak Lingala but, in the classroom, are taught in French. As a logical consequence of this dichotomy, most available books and other writings (elaborate or popular) in Congo are in French. Thus, Lingala is a relatively less documented language (less

* Corresponding author. Tel.: +32-495-48-97-50.
E-mail address: senemongaba@yahoo.fr

doi:10.1016/j.sbspro.2015.07.464

than 1000 books published to date). For historical reasons (the Christianization of Africa), most texts in Lingala are religious texts, although there is a growing trend of non-religious literature in Lingala, as well as a widening tendency to translating documents and reports of international organizations into Lingala. The irruption of the Internet in the cultural life of our day and age has introduced an important element in this scenario: the ever-mounting trend of pdf or html documents and debates in social networks. This provides the researcher with a great amount of data. However, the fact that Lingala is predominantly used in oral communication has a very important effect on the nature of such text: the spelling is often unstable and inconsistent. To that, one should add the ever-present lexical and grammatical influence of the French educational background of most Congolese speakers. Thirdly, in general, Congolese websites are in French and texts in other Congolese languages are all over the websites. Access to texts in Congolese languages require additional pre-processes to what is described (Scanell 2007, Kilgarriff 2010) for other under-resourced languages where the whole website is in the under-resourced language. For this reason, Lingala can be qualified as an under-resourced language where access to internet texts is particularly difficult. Otlogetswe used the terminology of Language with Limited Written Traditions or LWT (2004) for this group of languages.

The intrinsic nature of religious texts shifts the balance of a corpus towards a set of terms which are not widely used in today's everyday life. Adding internet sources to the mix has improved the representativeness and balance of a corpus otherwise dominated by religious texts.

This paper is a contribution to corpus building of under-resourced languages with limited access to internet texts. It describes a way to build a corpus using data from websites where the under-resourced language is a secondary language disseminated in main language pages. This is the case of Lingala as an under-resourced language and French as a main language of Congolese websites and social networks. As affirmed by Prinsloo for Bantu languages spoken in South Africa and I apply it for Congolese languages: 'The crucial development steps to future corpus-based lexicography, in chronological order, are: corpus creation, corpus annotation, qualitative corpus queries outputs and advanced dictionary writing systems capable of extracting relevant data from corpora and other lexicographic sources'.

My work of compiling a Lingala corpus aims to build a corpus allowing me to identify and analyze: the morphosemantic structures of Lingala affixes; syntax (structures, styles and strategies of disambiguation); lexicons; examples illustrating cases studied; spelling used by speakers.

These data will also allow researchers to create efficient dictionaries, schoolbooks and to coin new terms. The final objective of this work is to allow a better use of Lingala as a language of instruction.

Discussion and analysis in this paper are structured as follows: Section (2) presents an overview of Lingala variations. Section (3) discusses internet data extraction and cleaning issues I have faced. Section (4) explains the architecture we have adopted for building the corpus. Section (5) examines the spelling issues due to practical constraints. Section (6) outlines preliminary annotations and analyses obtained by processing the corpus with Unitex software. In the final part, we will then draw some conclusions and indicate some perspectives.

2. Lingala variations

Compiling a Lingala corpus means dealing with the problem of language variations. My intention is that this Lingala corpus represents a range of registers. In this section I will briefly describe Lingala varieties and its registers. As shown by Sene-Mongaba (2013a), Lingala has two main varieties: Lingala lya Mankanza (henceforth LM) and the variety which I am going to refer to in this paper as Lingala ya leló (today's Lingala, henceforth LL).

LM, which is considered as the classic or ‘pure’ variety, uses a full range of subject-verb agreement (SVA), as well as a full range of noun class grammatical agreement involving all modifiers (i.e., adjectives, demonstratives, quantifiers, and possessives). That means that verbs and all modifiers take the prefix determined by the head noun of the NPs subject. This is a general characteristic of Bantu Languages. LM also uses object markers, vocalic harmony and a 7-vowels system (a, i, e, ɛ, o, ɔ,u). Current or Spoken Lingala (henceforth SL) is the variety spoken in the