THE LINGALA CORPUS:

A TOOL FOR DESIGNING A LANGUAGE OF INSTRUCTION

Bienvenu Sene Mongaba
Université Pédagogique Nationale-Kinshasa
DR Congo
senemongaba@yahoo.fr
CILC 2015
1. Introduction
2. Lingála variations
3. Representativeness and balancing issues
4. The Cleaning operation
5. Building the Corpus
6. Perspectives
7. Conclusions
Introduction
Lingala

• the most widespread language of daily communication both in the cities of Kinshasa and Brazzaville
African countries pluriglossia

- Former colonial languages
- Languages of instruction and languages of administration.

- African languages
- Language of daily communications

- Kinshasa students:
- French vs Lingala
Lingala : LWT
Language with Limited Written Traditions

• most available books and other writings are in French.

• Lingala: less documented language

• Less than 1000 books published to date
LWT:
Language with Limited Written Traditions

- most texts in Lingala: Religious texts
- a growing trend of non-religious literature in Lingala
- a widening tendency to translating documents and reports of international organisations into Lingala.
Internet

• Internet: Website and social networks provides a great amount of data.
• Lingala: oral communication
• effect on the nature of such text:
• spelling unstable and inconsistent.
• lexical and grammatical influence of the French
corpus linguistics of LWT

• a way to build a corpus using LWT data from websites

• The LWT is a secondary language disseminated in main language pages.

• E.g.

• Congolese websites :

• Lingala : LWT

• French : the main language.
Internet documents

• Congolese websites are generally in French with disseminated texts in Lingala or in other Congolese languages.

• Search with Google and other search engines

• Keywords allowing to access documents in Lingala: ezali, biso, ngai, ekolo,...
Congolese websites are French designed
5 min d’histoire des noirs au Canada: les années 1900 à nos jours

5 min d'histoire des noirs au Canada: les années 1900 à nos jours.

Diasporablacknews.com ...

mars 02, 2015 | parcongomikili.com
| 0 commentaire

Vérité Choc Marie Lor: Scandale na Montréal, pasteur na muasi na ye ba yibi mbongo ya batu

Vérité Choc Marie Lor: Scandale na Montréal, pasteur na muasi na ye ba yibi mbongo ya batu ...

février 03, 2015 | parcongomikili.com
| 0 commentaire
• internet sources improve the representativeness and balance of a corpus otherwise dominated by religious texts
Corpus Objectives

- Identify and Analyse
- Morphosemantic structures of Lingala affixes
- Lingala Syntax (structures, styles and strategies of disambiguation)
- Lexicons
- Examples illustrating cases studied
- Spelling used by speakers
Corpus Applications

• efficient dictionaries

• schoolbooks

• coining new terms

• Lingala to be better used as a language of instruction
Lingala variations

• 2 main varieties:

• Lingala lya Mankanza (LM)

• Lingala ya Leló (today's Lingala) (LL).
Lingala speakers language competence: a sociolinguistic conundrum

- French lexicon in Lingala Facile can be challenging for some of those Lingala speakers
- Certain Lingala terms are not known to other Lingala speakers
- Less familiarity with numbers, colours, terms of specialty in Lingala
Scholars representations

• LM:acrolectal pole

• CL: mesolectal pole

• LL: basilectal pole
• few texts are entirely produced in LM or in LL and most texts show an inconsistent application of agreement

• LL does not enjoy high consideration by scholars.

• The LL variety : the variety of Lingala now in full spread
A representative and balanced corpus

- All varieties and registers of Lingala.
- Different domains not only religious
The project

• **Initially**:
  Finding deverbative nouns to generate neologisms for scientific purposes.

• **Broadening**:
  Making Dictionaries, Lingala learning books and Lingala schoolbooks
Social networks text

- Forums, Facebook, Youtube and Skype
- Comments written in spoken register (LF).
- Anonymity guarantees (name deleting)
- Popular TV or webTV on Website or Youtube.
- JTLF, KINSHASA MAKAMBO, CONGOMIKILI
Data Extraction

• Manual extraction (identify-select-copy-paste): from Facebook, Skype or Youtube.

• Crawling with HTtrack: Data from websites.

• Data in text format (Unicode UTF-8) with Notepad++.

• Cleaning.
robi
février 23, 2015 at 7:57
qu il soit petit ou grand on ne l a pas pris par force pour epouser henriette c leur vie priver

fils de l'homme
décembre 23, 2014 at 2:04
vraiment sans avoir motema mabe mariage na bino bosali yango bien mais en effet je constate que il ya decalage d'age entre ce couple ici etant donne que ns sommes de bantous et africains henriette aza mukolo ya mobali na mosika, mais ezala bolingo ya vraie ezala bolingo ya interet te, po kk aza joer po ofiba ye mbongo, eza juste remarque la femme abimi neti maman ya mobali i am sorry to say that but i have to.

Socofere
décembre 20, 2014 at 1:49
Salle Romeo golf sur l'avenue de la justice à Gombe
@MegaPatriote les films porno ne passent pas sur les chaînes congolaises, avant d'écrire des accusations, informez vous. Don't be silly guys your skin is black and will always be. Your are all from african origin.. respect your background.

Mega patriote a za na réson po na eloko moko ba clip americain et puis bo wébi ke musique e za reflé ya societer il ne sufis pas de cencurer mais il faut plutot ke ba parent ba kengela banna na bango na kin nga na sali consta ke sexe eza nango banaliser na niveau ke bana a 12an ba webi deja sé pas faute ya ba musicien mais ya education

En tout cas lingala facile oyo eza mosala malamu, bakonzi batala nanu likambo oyo, po soki musicien ayeli bango mwa 50us ndenge nini bako empeche musique naye? tango baza na bureau ya mabe boye, na ba salaire pe toyebite soki eza ndenge nini

bjr bravo a lingala facile car c'est en montrant des âneries comme ca que notre gouvernement prendra conscience et que les musiciens changerons ce ne pas en acceptant les mabangas que nos autorités ferons quelques choses nous devons nous sensibiliser pour arrêter ca. merci
• **MP aza** na réson po na eloko moko ba clip américain et puis **bo wébi** ke musique e za reflé ya societer il ne sufit pas de cencurer mais il faut plutot ke **ba parent ba kengela banna** na bango na kin nga na **sali consta** ke sexe eza **nango banaliser**.

• **MP aza** na raison po na eloko moko : **baclip** américain. Et puis, **boyébi** ke musique eza reflet ya société. il ne suffit pas de censurer mais il faut plutôt ke **baparent bakengela bana** na bango. Na kin nga **nasali constat** ke sexe eza **na yango banalisé**.
Spelling decision

• Each file crawled from internet was standardized by manual and automatic replacements.
User: user

Choose the language you want to work on:

Lingala

[OK] [Exit]
Preprocessing & Lexical parsing

Preprocessing

- No separator normalization (allows preprocessing graphs to match multi-separators)
- Apply graph in MERGE mode:
  - Set...
- Apply graph in REPLACE mode:
  - Set...

Tokenizing
The text is automatically tokenized. This operation is language-dependant, so that Unitex can handle languages with special spacing rules.

Lexical Parsing

- Apply All default Dictionaries
- Analyse unknown words as free compound words (this option is available only for Dutch, German, Norwegian & Russian)
- Construct Text Automaton

GO!
Cancel but tokenize text
Cancel and close text
bonjour FABREGAS, vraiment le monde est pétant, félicitation, meme si les gens il vutt parle, rien a dire, vraiment félicitation, tala danse non ya ya ya Mado, ensuite nrembo ya ya ya MASCARA, change a nrembo niwos obima ya 2014, félicitation, obti album na yo tiko trop, ningie ya lobako trop avant a ba album nana bango obima, je crois que regrette leol, courage FABREGAS Kikoo trop he nrembo ya ya félicitation :

Bonjour!

We are happy to share our successful launch of our new album. We would like to express our gratitude to everyone who supported us throughout this journey. Thank you for your love and support. We promise to give you more of the same.

Please share your thoughts and comments with us.

FABREGAS

Select the dictionaries to be applied. You can sort them one by one using the arrows. Note that system dictionaries are given to the Dico program before the user ones.

Right-click a dictionary to get information about it:
Combining Lingala and French Dictionaries, I identify non-standardized spelling
Aim of standardizing spelling

- make the corpus user-friendly for lexical checking or syntactic analyses.

- characterization of users spelling : basis for establishing a Lingala spellchecker.
Unitex software accounting

• 225 files

• 142 652 sentences

• 5 081 323 tokens

• 110 502 types
Domains

• Domains
  social, politics, religion, cuisine, music and education

• File name structure
  Domain, Number of this file in the sub-corpus, Sub-corpus abbreviation and File proper name

• E.g.
  social 2 EL LM boyangeli zamba
Corpus storage

• 4 forms:
  a) Files without standardized spelling.
  b) Files with standardized spelling.
  c) Corpus without XML tags
  d) Corpus with XML tags.
Preliminary results

• Setting Lingala dictionary and syntactic and semantics annotations in Unitex

• Each word annotated (automatic) with no disambiguation.

• Frequency

• Extraction of deverbative nouns and derivative verbs
Perspectives

- The creation of this Lingala corpus is still in progress.

- design an automatic Spell-checker.

- Translations from French / English to Lingala.

- Continue adapting the Unitex software to enhance its uses for Lingala.

- Disambiguation possibilities; develop the annotation (manual and automatic)
Conclusions

• Issues when building a LWT corpus
• First Lingala corpus 5-millions words
• Useful tool for researchers in
• Computational linguistics,
• Lexicography
• Schoolbooks writers
• ICT sectors operating in Congo,
• Useful tool for general speakers of Lingala
Matondo mingi!
Thank you!