AfLaT 5 (@ GAPSYM 7)

5th Workshop on African Language Technology

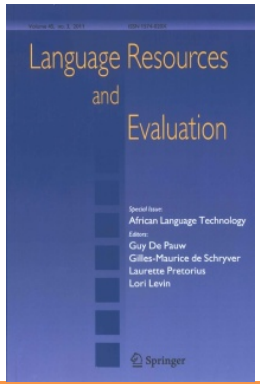
Friday 6 December 2013 – Ghent University

Provinciehuis, Gouvernementstraat 1, 9000 Ghent, Belgium

Programme and Abstracts Booklet

ed. G-M de Schryver

WORKSHOP SERIES		
1st, AfLaT 2009	Athens, Greece	@ EACL 2009
2nd, AfLaT 2010	Valletta, Malta	@ LREC 2010
3rd, AfLaT 2011	Addis Ababa, Ethiopia	@ AGIS 2011
4th, AfLaT 2012	Istanbul, Turkey	@ LREC 2012
5th, AfLat 2013	Ghent, Belgium	@ GAPSYM 2013



AfLaT was conceptualized at the *5th World Congress of African Linguistics* (WOCAL 5), African Union Conference Center, Addis Ababa, Ethiopia, in August 2006, and is led by the AfLaT Team: Guy De Pauw (U Antwerp), Gilles-Maurice de Schryver (U Ghent), and Peter Waiganjo Wagacha (U Nairobi).

Main AfLaT publication, in addition to the workshop proceedings, is a special issue of the A1-journal *Language Resources and Evaluation* in 2010 on African Language Technology.

Word of Welcome from the Workshop Organizer

This year's workshop dealing with *African Language Technology* (AfLaT) takes place on Friday 6 December 2013, at Ghent University. It is the fifth in the series, and conceived differently from the earlier ones, in that we wish to broaden our activities by reaching out to all colleagues who have lexical resources for the African languages, and already work with those resources, but have not yet necessarily made the move to using advanced computational routines to speed up the analysis or the building of tools.

In other words:

- Are you a corpus linguist working on the African languages, and would you like to know how computational linguistics could further your work, then this workshop is for you.
- Or, do you have African language data (corpora), and would you like to know about the state-of-the-art software tools to annotate and search those data, then this workshop is for you.
- Of course, if you are already into computational linguistics for the African languages, then you are most welcome to present your latest research results.

This year's workshop, then, is conceived as a MasterClass, led by the founding members of AfLaT: Guy De Pauw (U Antwerp), Gilles-Maurice de Schryver (U Ghent), and Peter Wagacha (U Nairobi).

The programme for the day, as well as the abstracts, follow herewith. In each talk the current data sets and/or research is presented during max. 20 minutes, to be followed by a discussion and advice from those present for max. 10 minutes. The working languages are English and French.

We thank the organizers of GAPSYM, with which we co-locate this year.

Workshop Organizer:

Gilles-Maurice de Schryver

KongoKing Research Group

Department of Languages and Cultures

Ghent University

Rozier 44

B-9000 Ghent (Belgium)

E-mail: gillesmaurice.deschryver@ugent.be

Programme			
Time slot	Activity	Presenter	Language(s)
8:30 – 9:00	**REGISTRATION AND COFFEE**		
9:00 – 9:05	Welcome	Gilles-Maurice de Schryver	African languages
9:10 – 10:10	[KEYNOTE] Activating your corpus	Guy De Pauw	African languages
10:15 – 10:45	Simple tools for corpus processing	Georges Mertens	Swahili
10:50 – 11:20	Using a diachronic Swahili prose & poetry corpus for linguistic research	Maud Devos	Swahili
11:20 – 11:50	**COFFEE BREAK**		
11:50 – 12:20	Semantic classification of Dutch and Afrikaans noun-noun compounds	Ben Verhoeven, Walter Daelemans and Gerhard van Huyssteen	Dutch and Afrikaans
12:25 – 12:55	Exploiting corpora and multimedia databases: A first study on double and triple negation in Nsong (B85d)	Joseph Koni Muluwa	Nsong
13:00 – 13:30	A corpus-based study of the Lusoga grammar	Minah Nabirye	Lusoga
13:30 – 14:30	**LUNCH**		
14:30 – 15:00	Building a hybrid set of diachronic Kikongo corpora	Gilles-Maurice de Schryver	Kikongo
15:00 – 15:30	Negation markers, focus markers and Jespersen cycles in Kikongo (Bantu, H16): a comparative and diachronic corpus-based approach	Jasper De Kind	Kikongo
15:30 – 16:00	L'étude comparative et de contacts des langues africaines par l'ordinateur, un casse-tête !!!	Jean-Pierre Donzo Bunza	NW DRC
16:00 – 16:30	**COFFEE BREAK**		
16:30 – 17:00	Un corpus du kirundi. Description, usages et perspectives	Ferdinand Mberamihigo	Kirundi
17:00 – 17:30	A corpus-driven study of the expression of modality in Luganda	Deo Kawalya	Luganda
18:30	**RECEPTION**		
Last-minute cancellation (literally, six hours before the workshop; the paper will be distributed at the workshop but not read):			
The Somali Corpus: State of the art, and tools for linguistic analysis		Jama Musse Jama	Somali

Activating your corpus

GUY DE PAUW
CLIPS – COMPUTATIONAL LINGUISTICS GROUP
UNIVERSITY OF ANTWERP
BELGIUM

Thanks to the many localization efforts and improved ICT access in urban as well as rural areas in sub-Saharan Africa, a vast amount of vernacular content is now finding its way to the Internet. This has not only served African languages to increase their visibility in the digital world, but this freely available language data also constitutes an amazing resource for corpus linguists. Being able to illustrate one's theories and hypotheses on the basis of empirical data from a natural-language corpus is a tremendous advantage in the current research climate in linguistics, as we slowly and quite rightly move away from its armchair.

But we are not the only ones that can learn new insights from corpora. So can computers. Language corpora nowadays have become essential for the development of language technology tools and applications. In this presentation, I will provide a brief tutorial on how machine learning techniques can automatically extract knowledge from large amounts of data and through an overview of our research efforts in data-driven African language technology, I will illustrate how, using established approaches, well-chosen knowledge representation techniques and just a little bit of elbow grease, you can *activate your corpus* and turn it into an essential building block for the development of state-of-the-art language technology tools.

Simple tools for corpus processing

GEORGES MERTENS
DEPARTMENT OF LANGUAGES AND CULTURES
GHENT UNIVERSITY
BELGIUM

I am very happy to read that this year the organizers “wish to broaden their activities by reaching out to all colleagues who have lexical resources for the African languages, and already work with those resources.” That far I go along. But then “are you a corpus linguist working on African languages, and would you like to know how computational linguistics could further your work.” I have to disagree. No, I am not a corpus linguist. I am a lecturer and have been teaching Kiswahili for the last thirty-five years of my live. I do not know much about “the state-of-the-art software tools” but have been using computers and informatics since the first year I have been teaching Kiswahili and was making a Kiswahili-Dutch dictionary.

The title of my talk is “Simple tools for corpus processing.” My aim is to demonstrate how a few simple tools which are open source and thus freely available on any MS Windows, Apple Mac OS or Linux/Unix machine helps me with my teaching of Kiswahili.

The first tool is Regular Expressions (RE). According to Wikipedia a regular expression is a sequence of characters that forms a search pattern, mainly for use in pattern matching with strings. If you can define what you need, regular expressions will provide you with all the occurrences of your pattern in any text or corpus. A lot of examples will be given so that you will be convinced that RE should be in the curriculum of any linguistics / literature student.

A second tool is a simple script which gives you a frequency list of all the words in any text. When you are teaching a language it is important to know which words are used more than others. And even a frequency list can give you some indication about the contents of the text.

The last tool is a program which allows me to extract the root of any Kiswahili word. As you know Kiswahili is a '*langue agglutinante*' and if you are making a wordlist or dictionary you are not interested in *ninakupenda* but only in *-penda*. It is still a work in progress.

I am aware that some very intelligent people prefer state-of-the-art software tools. I guess these tools are very useful but I am unable to judge. My experience is that sometimes they are rather heavy and complicated and cost a lot of money. That is why I like simple tools which are free and open source and which do your job much faster. *Des goûts et des couleurs ...*

Using a diachronic Swahili prose & poetry corpus for linguistic research

MAUD DEVOS
LINGUISTICS SERVICE
ROYAL MUSEUM FOR CENTRAL AFRICA
TERVUREN, BELGIUM

In this paper we want to present a new dedicated, custom-made Swahili prose & poetry corpus containing the full text of 45 works. It contains a synchronic corpus of over 1 million tokens (i.e., running words) as well as two smaller diachronic components, totalling over 300,000 tokens. An important aspect of the synchronic corpus is that it is geographically balanced containing materials from the Kenyan coast and inland as well as from the Tanzanian coast and inland.

In order to illustrate the utility of a diachronic as well as a geographically balanced synchronic Swahili corpus we will briefly discuss two case studies. The first looks at a typologically remarkable epistemic sentence adverb derived from a verb from meaning 'someone / something habitually goes'. The typological particularity of this sentence adverb makes one wonder about its syntagmatic as well as semantic history which should be retrievable in the diachronic corpus. The second case study discusses a verb form which has been assumed to be a regionally bound alternative to the Standard Swahili perfect *me*. The geographically balanced corpus is used to verify whether there is quantitative evidence for a regionally bound substrate analysis.

Semantic classification of Dutch and Afrikaans noun-noun compounds

BEN VERHOEVEN, WALTER DAELEMANS AND GERHARD VAN HUYSSTEEN
CLIPS – COMPUTATIONAL LINGUISTICS GROUP
UNIVERSITY OF ANTWERP
BELGIUM

The meaning of compound words is often ambiguous because there is no explicit description of the relation between the compound constituents. A newly produced compound like '*donut seat*' can be interpreted in different ways, such as '*seat with donut nearby*', '*seat that looks like donut*' or even '*seat made of donuts*'. An automatic semantic analysis of these compounds may shed more light on this issue.

Building on previous research by Ó Séaghdha for English, the task of semantically analysing noun-noun compounds was considered a supervised machine learning problem. We adopted and adapted Ó Séaghdha's semantic classification scheme and guidelines for noun-noun compounds. This scheme describes 11 classes, of which 6 are semantically specific. Lists of noun-noun compounds were annotated according to this classification scheme. Following the distributional hypothesis that states that the set of contexts of a word can be used as a representation of its meaning, vectors with co-occurrence information on the compound constituent nouns were used to construct feature vectors for our classifier. We present results of our experiments on Dutch and Afrikaans compounds that confirm the learnability of this classification task. Different experiments vary in the number and kind of co-occurrence words they select (e.g. content words vs. function words). Our results are promising and approach the accuracies reached by similar systems for English.

Exploiting corpora and multimedia databases: A first study on double and triple negation in Nsong (B85d)

JOSEPH KONI MULUWA

DEPARTMENT OF LANGUAGES AND CULTURES

GHENT UNIVERSITY

BELGIUM

This presentation is part of an endangered language documentation project funded by the DoBeS program of the Volkswagen Foundation through a 3-year grant (2012-2015) and involves collaboration between Ghent University, Kinshasa University and the Humboldt University of Berlin.

The planned research outcomes include:

1. Systematic documentation of Nsambaan, Nsong, and Ngong;
2. Corpus-based lexicons of the languages under study;
3. Guide of useful plants, mushrooms and animals from the Kwilu;
4. Development of a practical orthography for the languages under study;
5. Drafting of local MA dissertations at the ISP of Kikwit;
6. Raising awareness on language endangerment;
7. Drafting of high-quality academic papers on interesting lexical and grammatical phenomena observed in the documentation in order to continue linguistic research on these languages.

The present presentation mainly belongs to goal number 7, whereby corpora and multimedia databases are exploited, to show how negative utterances are expressed in the Bantu language Nsong. Four negative markers are used in three positions: pre-initial, post-initial and post-verbal. These morphemes are used to express single, double and triple negation. If the two morphemes seem to be inherited (the pre-initial) or old (the post-initial), the post-verbal negative marker could be borrowed by language contact. However, this particle is currently used most often to reinforce negation and to focus different sentence constituents.

A corpus-based study of the Lusoga grammar

MINAH NABIRYE
DEPARTMENT OF LANGUAGES AND CULTURES
GHENT UNIVERSITY
BELGIUM

This paper presents part of my PhD study aimed at writing a grammar of Lusoga. The language is largely oral so its description in the current study is mostly based on natural language sources. A 1.7-million-word Lusoga corpus collected over a period of over two years, is the focus of this presentation. The corpus covers both the oral and written genres. Over 100 hours of oral recordings were transcribed and amount to a full one third of the entire corpus. The other two-thirds of the corpus was collected from sources such as translations and written Lusoga texts which were either downloaded or scanned and OCRed.

There are three main areas to describe the Lusoga grammar, namely: sounds, words and constructions. Of the three, only the description of Lusoga sound is partially corpus-based. The description of words and constructions will be directly based on corpus data. Selection of the words, word classes and construction types to include in the PhD study has only just begun. So far, words appearing in the corpus at least 12 times have been translated and part-of-speech tagged using TLex. In the process, the material was lemmatized. Further clean-up is ongoing. The analysis of concordances for a selected set of words will be the foundation for the description of the grammatical usage of some of these word structures in constructions.

This paper focuses on presenting the Lusoga corpus and its analysis that has been done to date. The paper will also show how this analysis will be used in the description of the grammar of Lusoga.

Building a hybrid set of diachronic Kikongo corpora

GILLES-MAURICE DE SCHRYVER
DEPARTMENT OF LANGUAGES AND CULTURES
GHENT UNIVERSITY
BELGIUM

Few sub-Saharan African languages have a long history of written texts. Most are not even written on a regular basis today. Compiling a corpus for sub-Saharan African languages is thus a difficult task; a diachronic corpus in most cases a theoretical impossibility. However, some languages have been the subject of early attention by outsiders, and have historical documentation. Such is the case of Kikongo, a Bantu language spoken in West-Central Africa. Spoken in Kongo and neighbouring kingdoms, which from the end of the 15th century had strong trade and diplomatic contacts with Portugal and other European nations, Kikongo gained the interest of missionaries. A small number of dictionaries, grammars and texts, document Kikongo varieties as spoken in the 17th and 18th century. Travel accounts also supply some linguistic information. While historians have extensively exploited these early European documents, linguists thus far made little use of the data. The KongoKing research group is the first to build a diachronic language corpus for Kikongo.

The Kikongo corpus brings together a variety of text types, starting with the first manuscripts dating back to the early 17th century (religious texts, grammatical sketches, vocabulary lists, etc.), toponyms and hydronyms as found on the first and later maps, Kikongo words and phrases mentioned in (official) correspondence, travel accounts and the first descriptions of the region, on to the later proper dictionaries, language textbooks, bibles, historical accounts, and so on. From the hand-written manuscripts of four centuries ago to the printed material that followed, and the online material produced today, all is being digitized, classified and tagged with metadata. A single corpus-query system, and knowledge of the changing and unstable orthographies, allows for diachronic searches through all this data.

In this presentation this corpus—which is actually a hybrid set of sub-corpora—will be presented. It is hybrid because it contains both diachronic and synchronic parts, for several dozen Kikongo varieties, with for most sub-corpora also parallel translations. The largest part is still unannotated, although the very first text (dating from 1624) has already been POS-tagged, and is being exploited.

Negation markers, focus markers and Jespersen cycles in Kikongo (Bantu, H16): a comparative and diachronic corpus-based approach

JASPER DE KIND
DEPARTMENT OF LANGUAGES AND CULTURES
GHENT UNIVERSITY
BELGIUM

The present paper builds upon recent research by Devos & Van der Auwera (2013) on double and triple negation in Bantu. In this paper, we give a more detailed account of this phenomenon in the Kikongo dialect continuum (Bantu H10). In Kikongo, negation is commonly doubly marked, i.e. by a verbal negative marker *ka/ke* and the post-verbal marker *ko* also appears. This negation strategy is reminiscent of what is commonly called a Jespersen cycle: in a first stage, only one negative marker is used. This negation marker is subsequently strengthened by some kind of focus marker whose emphatic value may become neutralized through extensive use (cf. *ne…pas* in French). In a final stage, the initial negative marker may be dropped (cf. colloquial French *pas*). In this paper, we investigate the syntactic behaviour of the post-verbal marker *ko*, as well as its possible origin, in order to see whether we are dealing with a Jespersen cycle comparable to the well-known French example. Devos & Van der Auwera (2013) also observe a possible third negation marker, the locative possessive pronoun of class 17, i.e. *kwandi*. This marker is also treated in the present paper. It seems as if it has not (yet) received an intrinsic negative value in Kikongo, contrary to *ko*.

L'étude comparative et de contacts des langues africaines par l'ordinateur, un casse-tête !!!

JEAN-PIERRE DONZO BUNZA
UNIVERSITÉ LIBRE DE BRUXELLES
BELGIQUE
&
DEPARTMENT OF LANGUAGES AND CULTURES
GHENT UNIVERSITY
BELGIUM

Je présente ici une petite expérience concernant une étude doctorale en cours sur la parenté linguistique, la variation et les contacts linguistiques de quelques langues du nord-ouest de la République Démocratique du Congo ; les difficultés technologiques, depuis la récolte des données de terrain à l'analyse et la rédaction des textes sur les résultats à l'aide de logiciels piqués, au hasard, sur le net.

Les objectifs principaux de notre étude sont :

- Étudier l'évolution de 10 langues bantoues (Bolondo, Bonyange, Ebudza, Ebwela, Libobi, Mondongo, Monyongo, Mosange, Ngombɛ-bobo, Pakabete) à partir des correspondances phonologiques et lexicales des vocabulaires culturels ;
- Déterminer l'influence de 7 langues oubanguiennes (Gbanziri, Mbanza, Mono, Monzombo, Ngbaka, Ngbandi, Yango) sur les langues bantoues à partir de la comparaison de leur vocabulaire commun.

En effet, les recherches de terrain ont consisté à récolter des lexiques et des phrases sur base d'un questionnaire de lexique de 800 mots et celui des phrases d'environ 250 phrases françaises que le locuteur devrait traduire.

SCHÉMA DU TRAVAIL

- Étape 1 : Récolte des données orales se sont faites par enregistrement pour certaines données sur bande cassette et pour d'autres à l'aide d'un dictaphone ;
- Étape 2 : Transcription des données ;
- Étape 3 : Lexiques et phrases ;
- Étape 4 : Notre étude consiste à décrire les langues et à comparer les lexiques ;
- Étape 5 : La comparaison des lexiques des différentes langues pour la lexicostatistique ;
- Étape 6 : Établir les arbres génétiques ;
- Étape 7 : Détecter les emprunts phonologiques et lexicaux des langues oubanguiennes dans les langues bantoues ;
- Étape 8 : Cartographie des emprunts.

Un corpus du kirundi. Description, usages et perspectives

FERDINAND MBERAMIHIGO
UNIVERSITÉ LIBRE DE BRUXELLES
BELGIQUE
&
DEPARTMENT OF LANGUAGES AND CULTURES
GHENT UNIVERSITY
BELGIUM

Dans le cadre de notre projet doctoral en cours sur « L'expression de la modalité en kirundi : exploitation d'un corpus», nous avons été initié à la compilation d'un corpus électronique. Ainsi, nous avons pu compiler, à ce jour, un corpus du kirundi de 3 470 704 tokens. Ce travail est une œuvre pionnière pour cette langue, comme pour la plupart des langues africaines, à quelques exceptions près. Ce résultat a été obtenu grâce à un concours de plusieurs outils et techniques chaque fois adaptés aux données d'origine, qu'elles soient orales, écrites ou électroniques. Il fallait en effet rassembler les diverses ressources sous un même format, le format électronique. Dans le cadre de notre sujet, il a été rendu possible par le logiciel OmniPage pour la numérisation.

Tant dans la méthodologie qu'au niveau de l'analyse, le travail que nous menons concrétise ce à quoi un corpus peut servir pour des études dans le domaine de langues sans traditions écrites, parmi lesquelles figure le kirundi. Tout d'abord, dans la description des phénomènes linguistiques que nous étudions, les faits sont illustrés par des données issues du corpus, celles-ci ayant une marque d'authenticité car elles saisissent la langue dans son expression naturelle, contrairement à la méthode traditionnelle d'élicitation qui consiste à demander à un locuteur natif de construire un énoncé qui correspond au phénomène à illustrer. Plus encore l'outil électronique permet d'exprimer les faits en termes statistiques en comparant les phénomènes linguistiques du point de vue de leur fréquence. C'est là le principal apport de la linguistique de corpus. C'est ainsi que, dans le cadre de notre étude de la modalité, nous établissons les fréquences des diverses catégories modales au sein de notre corpus et nous établissons les comparaisons, tant sur le plan synchronique qu'au niveau diachronique.

De telles opérations sont rendues possibles par l'existence de logiciels permettant une génération automatique des données en fonction de paramètres bien déterminés. L'analyse est faite grâce au logiciel WordSmith Tools pour le traitement des données. Les perspectives offertes par notre corpus sont très nombreuses. Un corpus de près de trois millions et demi de tokens constitue un point de départ assez satisfaisant pour notre sujet. Mais pour l'avenir, il sera nécessaire de l'enrichir de plus en plus pour l'équilibrer davantage et le rendre plus représentatif. En outre, il sera nécessaire que, d'un corpus brut, il passe à un état de corpus annoté. Pour ce faire, il nous faudra acquérir des techniques d'annotation, que nous n'avons pas pour le moment. De la même façon, un corpus plus grand nous permettra également de perfectionner de plus un plus le correcteur d'orthographe (spellchecker) que la taille actuelle nous a permis de mettre au point dans une version modeste mais très utile. Ainsi, d'un corpus pour usage individuel nous pourrons passer à une version susceptible de servir de ressource à des demandeurs de profils variés, avec plus de fonctionnalités.

A corpus-driven study of the expression of modality in Luganda

DEO KAWALYA
DEPARTMENT OF LANGUAGES AND CULTURES
GHENT UNIVERSITY
BELGIUM
&
MAKERERE UNIVERSITY
KAMPALA, UGANDA

In this presentation, I report on the current process of assembling a corpus that will be used in the study of the expression of modality in Luganda. In the same vein, and as part of this more encompassing ongoing research, I show how a 1.5m-token corpus has been used in a pilot study on the possibility marker *-sóból-* in Luganda (i.e. *Diachronic semantics of the modal verb -sóból- in Luganda: A corpus-driven approach*).

For this fully corpus-driven study, corpus material from the internet, electronic transfers, scans (+OCR), transcriptions and translations, covering a time-depth of 120 years, was organized into 17 genres. It was queried with WordSmith Tools and concordance lines were exported into a spreadsheet where they were analysed, tagged, glossed and annotated. For very frequent forms, standard sampling techniques were used, and for the lesser frequent ones all instances were considered.

Although the methodology produced feasible results, especially with regard to synchronic analyses, the corpus itself, in its current state, suffers from both a diachronic and author end even genre imbalance (e.g. there are no materials from the 1930s – 1950s and all texts from the 1900s are not only by a single author but also belong to only two genres). Secondly, at 1.5m tokens, the corpus is still small and this leads to some non-statistically relevant results. To cater for these problems, in the current corpus expansion process more emphasis is being put on old materials (i.e. before 1960) and on underrepresented authors and genres. At present, the corpus is not tagged for parts of speech or any other features, for which I hope to acquire skills from this forum on how best to approach it.