EACL 2009

Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages

AfLaT 2009

31 March 2009
Megaron Athens International Conference Centre
Athens, Greece

Production and Manufacturing by
TEHNOGRAFIA DIGITAL PRESS
7 Ektoros Street
152 35 Vrilissia
Athens, Greece

This workshop is sponsored by

ISBN 1-932432-25-6

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

Preface

In multilingual situations, language technologies are crucial for providing access to information and opportunities for economic development. With somewhere between 1,000 and 2,000 different languages, Africa is a multilingual continent *par excellence* and presents acute challenges for those seeking to promote and use African languages in the areas of business development, education, research, and relief aid.

In recent times a number of researchers and institutions, both from Africa and elsewhere, have come forward to share the common goal of developing capabilities in language technologies for the African languages. The goal of the workshop, then, is to provide a forum to meet and share the latest developments in this field. It also seeks to attract linguists who specialize in African languages and would like to leverage the tools and approaches of computational linguistics, as well as computational linguists who are interested in learning about the particular linguistic challenges posed by the African languages.

The workshop consists of an invited talk on African language families and their structural properties by Prof. Sonja Bosch (UNISA, South Africa), followed by refereed research papers in computational linguistics. The call for papers specified that the focus would be on the less-commonly studied and lesser-resourced languages, such as those of sub-Saharan Africa. These could include languages from all four families: Niger-Congo, Nilo-Saharan, Khoisan and Afro-Asiatic, with the exception of Arabic (which is covered by the 'Computational Approaches to Semitic Languages' workshop). Variants of European languages such as African French, African English or Afrikaans were also excluded from the call. The call was well answered, with 24 proposals being submitted. Following a rigorous review process, nine were withheld as full papers, and a further seven as poster presentations. These proceedings contain the double-blind, peer reviewed texts of the contributions that were withheld.

We would like to thank the US Army RDECOM International Technology Center - Atlantic for generously sponsoring this workshop.

We wish you a very enjoyable time at AfLaT 2009!

— The Organizers.

AfLaT 2009 Organizers

Proceedings Editors:

Guy De Pauw
Gilles-Maurice de Schryver
Lori Levin

Organizers:

Lori Levin, Language Technologies Institute, Carnegie Mellon University (USA)
Guy De Pauw, CNTS - Language Technology Group, University of Antwerp (Belgium) - School of Computing and Informatics, University of Nairobi (Kenya) - AfLaT.org
John Kiango, Institute of Kiswahili Research, University of Dar Es Salaam (Tanzania)
Judith Klavans, Institute for Advanced Computer Studies, University of Maryland (USA)
Manuela Noske, Microsoft Corporation (USA)
Gilles-Maurice de Schryver, African Languages and Cultures, Ghent University (Belgium) - Xhosa Department, University of the Western Cape (South Africa) - AfLaT.org
Peter Waiganjo Wagacha, School of Computing and Informatics, University of Nairobi (Kenya) - AfLaT.org

Program Committee:

Alan Black, Carnegie Mellon University (USA)
Sonja Bosch, University of South Africa (South Africa)
Christopher Cieri, University of Pennsylvania, Linguistic Data Consortium (USA)
Guy De Pauw, University of Antwerp (Belgium)
Gilles-Maurice de Schryver, Ghent University (Belgium)
Robert Frederking, Carnegie Mellon University (USA)
Dafydd Gibbon, University of Bielefeld (Germany)
Jeff Good, SUNY Buffalo (USA)
Mike Gasser, Indiana University (USA)
Gregory Iverson, University of Maryland, Center for Advanced Study of Language (USA)
Stephen Larocca, US Army Research Lab (USA)
Michael Maxwell, University of Maryland, Center for Advanced Study of Language (USA)
Manuela Noske, Microsoft Corporation (USA)
Tristan Purvis, University of Maryland, Center for Advanced Study of Language (USA)
Clare Voss, US Army Research Lab (USA)
Peter Waiganjo Wagacha, University of Nairobi (Kenya)
Briony Williams, University of Wales, Bangor (Great Britain)

Invited Speaker:

Sonja Bosch, Department of African Languages, University of South Africa (South Africa)

Table of Contents

Conference Program

Tuesday, March 31, 2009

09:00–10:30 Invited Talk: ”African Language Families and their Structural Properties” by Sonja Bosch

10:30–11:00 Coffee Break

Session 1: Corpora (11:00–12:30)

11:00–11:30 *Collecting and Evaluating Speech Recognition Corpora for Nine Southern Bantu Languages*
Jaco Badenhorst, Charl Van Heerden, Marelie Davel and Etienne Barnard

11:30–12:00 *The SAWA Corpus: A Parallel Corpus English - Swahili*
Guy De Pauw, Peter Waiganjo Wagacha and Gilles-Maurice de Schryver

12:00–12:30 *Information Structure in African Languages: Corpora and Tools*
Christian Chiarcos, Ines Fiedler, Mira Grubic, Andreas Haida, Katharina Hartmann, Julia Ritz, Anne Schwarz, Amir Zeldes and Malte Zimmermann

12:30–14:00 Lunch Break

Session 2: Morphology, Speech and Part-of-Speech Tagging (14:00–16:00)

14:00–14:30 *A Computational Approach to Yorùbá Morphology*
Raphael Finkel and Odetunji Ajadi Odejobi

14:30–15:00 *Using Technology Transfer to Advance Automatic Lemmatisation for Setswana*
Hendrik Johannes Groenewald

15:00–15:30 *Part-of-Speech Tagging of Northern Sotho: Disambiguating Polysemous Function Words*
Gertrud Faaß, Ulrich Heid, Elsabé Taljard and Danie Prinsloo

15:30–16:00 *Development of an Amharic Text-to-Speech System Using Cepstral Method*
Tadesse Anberbir and Tomio Takara

16:00–16:30 Coffee Break

Tuesday, March 31, 2009 (continued)

Session 3: General Papers and Discussion (16:30–18:00)

16:30–17:00 *Building Capacities in Human Language Technology for African Languages*
Tunde Adegbola

17:00–17:30 *Initial Fieldwork for LWAZI: A Telephone-Based Spoken Dialog System for Rural South Africa*
Tebogo Gumede and Madelaine Plauché

17:30–18:00 Discussion

Poster Session (During Coffee and Lunch Breaks)

Setswana Tokenisation and Computational Verb Morphology: Facing the Challenge of a Disjunctive Orthography
Rigardt Pretorius, Ansu Berg, Laurette Pretorius and Biffie Viljoen

Interlinear Glossing and its Role in Theoretical and Descriptive Studies of African and other Lesser–Documented Languages
Dorothee Beermann and Pavel Mihaylov

Towards an Electronic Dictionary of Tamajaq Language in Niger
Chantal Enguehard and Issouf Modi

A Repository of Free Lexical Resources for African Languages: The Project and the Method
Piotr Bański and Beata Wójtowicz

Exploiting Cross-Linguistic Similarities in Zulu and Xhosa Computational Morphology
Laurette Pretorius and Sonja Bosch

Methods for Amharic Part-of-Speech Tagging
Björn Gambäck, Fredrik Olsson, Atelach Alemu Argaw and Lars Asker

An Ontology for Accessing Transcription Systems (OATS)
Steven Moran