AfLaT 2010
Proceedings of the Second Workshop on
African Language Technology

Held in collocation with the
Seventh International Conference on
Language Resources and Evaluation (LREC 2010)

May 18, 2010
Valletta, Malta

Edited by
Guy De Pauw
Handré Groenewald
Gilles-Maurice de Schryver
African Language Technology is rapidly becoming one of the hottest new topics in computational linguistics. The increasing availability of digital resources, an exponentially growing number of publications and a myriad of exciting new projects are just some of the indications that African Language Technology has been firmly established as a mature field of research. The AfLaT workshops attempt to bring together researchers in the field of African Language Technology and provide a forum to present on-going efforts and discuss common obstacles and goals.

We are pleased to present to you the proceedings of the Second Workshop on African Language Technology (AfLaT 2010), which is held in collocation with the Seventh International Conference on Language Resources and Evaluation (LREC 2010). We were overwhelmed by the quantity and quality of the submissions we received this year, but were lucky enough to have a wonderful program committee, who sacrificed their valuable time to help us pick the cream of the crop. We pay tribute to their efforts by highlighting reviewers’ quotes in the next paragraphs.

Grover et al. kick off the proceedings with a comprehensive overview of the HLT situation in South Africa, followed by Bański and Wójtowicz’s description of an initiative that is beneficial to the creation of resources [...] for African languages. De Pauw et al. describe techniques that could be used to develop a plethora of [...] HLT resources with minimal human effort, while Shah et al. present impressive results on tackling the problem of NER in MT systems between languages, one of which at least is poorly resourced. Groenewald and du Plooy’s paper tackles the all too-often overlooked problem of text anonymization in corpus collection, followed by Chege et al.’s effort that is significant [...] to the open source community, not just for Gikuyu but for the African languages in general. Faass presents a useful resource for further computational processing of the language of Northern Sotho.

Tachbelie and Menzel provide a clear and concise overview of the general issues affecting language models for morphologically rich languages, while Van der Merwe et al. go into an informative discussion of the properties of the Zulu verb, its extensions, and deverbatives. The paper by Oosthuizen et al. aptly discusses the issue of quantifying and correcting transcription differences between inexperienced transcribers, while Davydov’s paper is an interesting case study for collecting corpora for “languages recently put into writing”. Ng’ang’a presents the key resource for the identification of a machine-readable dialectal dictionary for Igbo and Purvis concludes by discussing a corpus that contributes to the development of HLT tools for Dagbani.

We are proud to have Justus Roux as the invited speaker for this year’s edition of AfLaT to discuss one of the most often asked and rarely answered questions in our field of research: Do we need linguistic knowledge for speech technology applications in African languages? We hope you enjoy the AfLat 2010 workshop and look forward to meeting you again at AfLaT 2011.

The AfLaT 2010 Organizers
Guy De Pauw
Handré Groenewald
Gilles-Maurice de Schryver
Peter Waiganjo Wagacha
HTTP://AFLAT.ORG/AFLAT2010
Workshop Organizers

Guy De Pauw  
CLiPS - Computational Linguistics Group, University of Antwerp, Belgium  
School of Computing & Informatics, University of Nairobi, Kenya

Handré Groenewald  
Centre for Text Technology (CTexT), North-West University, South Africa

Gilles-Maurice de Schryver  
African Languages and Cultures, Ghent University, Belgium  
Xhosa Department, University of the Western Cape, South Africa  
TswaneDJe HLT, South Africa

Peter Waiganjo Wagacha  
School of Computing & Informatics, University of Nairobi, Kenya
Workshop Program Committee

Tunde Adegbola - African Languages Technology Initiative (Alt-i), Nigeria
  Tadesse Anberbir - Addis Ababa University, Ethiopia
  Winston Anderson - University of South Africa, South Africa
    Lars Asker - Stockholm University, Sweden
  Etienne Barnard - Meraka Institute, South Africa
  Piotr Bąński - University of Warsaw, Poland
  Ansu Berg - North-West University, South Africa
  Sonja Bosch - University of South Africa, South Africa
  Chantal Enguehard - LINA - UMR CNRS, France
  Gertrud Faaß - Universität Stuttgart, Germany
  Björn Gambäck - Swedish Institute of Computer Science, Sweden
  Katherine Getao - NEPAD e-Africa Commission, South Africa
    Dafydd Gibbon - Universität Bielefeld, Germany
    Arvi Hurskainen - University of Helsinki, Finland
    Fred Kitoogo - Makerere University, Uganda
  Roser Morante - University of Antwerp, Belgium
  Lawrence Muchemi - University of Nairobi, Kenya
  Wanjiku Ng‘ang’a - University of Nairobi, Kenya
  Oδέιμνι Oδέινβι - University College Cork, Ireland
  Chinyere Ohiri-Anichie - University of Lagos, Nigeria
  Sulene Pilon - North-West University, South Africa
  Laurette Pretorius - University of South Africa, South Africa
  Rigardt Pretorius - North-West University, South Africa
    Danie Prinsloo - University of Pretoria, South Africa
    Justus Roux - North-West University, South Africa
    Kevin Scannell - Saint Louis University, United States
  Gerhard Van Huyssteen - Meraka Institute, South Africa
# Table of Contents

*Do we need linguistic knowledge for speech technology applications in African languages?*
Justus Roux ................................................................. 1

*An HLT profile of the official South African languages*
Aditi Sharma Grover, Gerhard B. van Huyssteen and Marthinus W. Pretorius ............... 3

*Open-Content Text Corpus for African languages*
Piotr Bański and Beata Wójtowicz ........................................ 9

*A Knowledge-Light Approach to Luo Machine Translation and Part-of-Speech Tagging*
Guy De Pauw, Naomi Maajabu and Peter Waiganjo Wagacha .............................. 15

*SYNERGY: A Named Entity Recognition System for Resource-scarce Languages such as Swahili using Online Machine Translation*
Rushin Shah, Bo Lin, Anatole Gershman and Robert Frederking ......................... 21

*Processing Parallel Text Corpora for Three South African Language Pairs in the Autshumato Project*
Hendrik J. Groenewald and Liza du Plooy .................................... 27

*Developing an Open source spell checker for Gikuyu*
Kamau Chege, Peter Waiganjo Wagacha, Guy De Pauw, Lawrence Muchemi, Wanjiku Ng’ang’a, Kenneth Ngure and Jayne Mutiga ................................. 31

*The verbal phrase of Northern Sotho: A morpho-syntactic perspective*
Gertrud Faaß ................................................................. 37

*Capturing Word-level Dependencies in Morpheme-based Language Modeling*
Martha Yifiru Tachbelie and Wolfgang Menzel ...................................... 43

*Towards the Implementation of a Refined Data Model for a Zulu Machine-Readable Lexicon*
Ronell Van der Merwe, Laurette Pretorius and Sonja Bosch .......................... 49

*Improving Orthographic Transcriptions of Speech Corpora*
Nicolas Laurenz Oosthuizen, Martin Johannes Puttkammer and Martin Schlemmer ...... 55

*Towards The Manding Corpus: Texts Selection Principles and Metatext Markup*
Artem Davydov ............................................................. 59

*Towards a Comprehensive, Machine-readable Dialectal Dictionary of Igbo*
Wanjiku Ng’ang’a ........................................................ 63

*Corpus Building in a Predominantly Oral Culture: Notes on the Development of a Multi-Genre Tagged Corpus of Dagbani*
Tristan Purvis .............................................................. 69

*Author Index* ............................................................... 73
# The Workshop Program

**Tuesday, May 18, 2010**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00–10:00</td>
<td><strong>Invited Talk</strong> - <em>Do we need linguistic knowledge for speech technology applications in African languages?</em> Justus Roux</td>
</tr>
<tr>
<td>10:00–10:30</td>
<td><em>An HLT profile of the official South African languages</em> Aditi Sharma Grover, Gerhard B. van Huyssteen and Marthinus W. Pretorius</td>
</tr>
<tr>
<td>10:30–11:00</td>
<td><strong>Coffee Break</strong></td>
</tr>
<tr>
<td>11:00–11:30</td>
<td><em>Open-Content Text Corpus for African languages</em> Plotr Bański and Beata Wójtowicz</td>
</tr>
<tr>
<td>11:30–12:00</td>
<td><em>A Knowledge-Light Approach to Luo Machine Translation and Part-of-Speech Tagging</em> Guy De Pauw, Naomi Maajabu and Peter Waiganjo Wagacha</td>
</tr>
<tr>
<td>12:00–12:30</td>
<td><em>SYNERGY: A Named Entity Recognition System for Resource-scarce Languages such as Swahili using Online Machine Translation</em> Rushin Shah, Bo Lin, Anatole Gershman and Robert Frederking</td>
</tr>
<tr>
<td>12:30–13:00</td>
<td><em>Processing Parallel Text Corpora for Three South African Language Pairs in the Autshumato Project</em> Hendrik J. Groenewald and Liza du Plooy</td>
</tr>
<tr>
<td>13:00–14:30</td>
<td><strong>Lunch Break</strong></td>
</tr>
</tbody>
</table>
Tuesday, May 18, 2010 (continued)

14:30–15:00  Developing an Open source spell checker for Gikuyu
Kamau Chege, Peter Waiganjo Wagacha, Guy De Pauw, Lawrence
Muchemi, Wanjiku Ng’ang’a, Kenneth Ngure and Jayne Mutiga

15:00–15:30  The verbal phrase of Northern Sotho: A morpho-syntactic perspective
Gertrud Faß

15:30–16:00  Capturing Word-level Dependencies in Morpheme-based Language
Modeling
Martha Yifiru Tachbelie and Wolfgang Menzel

16:00–16:30  Coffee Break

16:30–17:00  Towards the Implementation of a Refined Data Model for a Zulu Machine-
Readable Lexicon
Ronell Van der Merwe, Laurette Pretorius and Sonja Bosch

17:00–17:30  Improving Orthographic Transcriptions of Speech Corpora
Nicolas Laurens Oosthuizen, Martin Johannes Puttkammer and Martin
Schlemmer

17:30–18:00  Towards The Manding Corpus: Texts Selection Principles and Metatext
Markup
Artem Davydov

18:00–18:30  Towards a Comprehensive, Machine-readable Dialectal Dictionary of Igbo
Wanjiku Ng’ang’a

18:30–19:00  Corpus Building in a Predominantly Oral Culture: Notes on the
Development of a Multi-Genre Tagged Corpus of Dagbani
Tristan Purvis