# *Language Technology for Normalisation of Less-Resourced Languages*

*The 8th International Workshop of the ISCA Special Interest Group on Speech and Language Technology for Minority Languages (SaLTMiL2012)*
*and*
*the 4th Workshop on African Language Technology (AfLaT2012)*

**22 May 2012**

# ABSTRACTS

**Editors:**

Guy De Pauw, Kepa Sarasola and Francis M. Tyers

# Workshop Programme

09:15–09:30 Welcome / Opening Session

09:30–10:30 Invited Talk

- Sjur Moshagen Nørstebø. *How to build language technology resources for the next 100 years*

10:30–11:00 Coffee Break

11:00–13:00 Resource Creation

- Elaine Uí Dhonnchadha, Alessio Frenda and Brian Vaughan, *Issues in Designing a Spoken Corpus of Irish*.
- Wondwossen Mulugeta and Michael Gasser, *Learning Morphological Rules for Amharic Verbs Using Inductive LogicProgramming*
- Kristín Bjarnadóttir, *The Database of Modern Icelandic Inflection*
- Fadoua Ataa Allah and Siham Boulaknadel, *Natural Language Processing for Amazigh Language: Challenges and Future Directions*

13:00–14:00 Lunch Break

14:00–16:00 Resource Use

- Tommi A. Pirinen and Francis M. Tyers. *Compiling Apertium morphological dictionaries with HFST and using them in HFST applications*.
- Borbóla Siklósi, György Orosz, Attila Novák and Gábor Prószéky. *Automatic structuring and correction suggestion system for Hungarian clinical records*.
- Linda Wiechetek. *Constraint Grammar based Correction of Grammatical Errors for North Sàmi*.
- Michael Gasser, *Toward a Rule-Based System for English-Amharic Translation*.

16:00–16:30 Coffee Break

16:30–17:30 Poster Session

- Paola Carrión González and Emmanuel Cartier, *Technological Tools for Dictionary and Corpora Building for Minority Languages: Example of the French-based Creoles*.
- Denys Duchier, Brunelle Magnana Ekoukou, Yannick Parmentier, Simon Petitjean and Emannuel Schang, *Describing Morphologically-rich Languages using Metagrammars: a Look at Verbs in Ikota*.
- Tjerk Hagemeijer, Iris Hendrickx, Abigail Tiny and Haldane Amaro, *A Corpus of Santomé*.
- Sigrún Helgadóttir, Asta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir and Hrafn Loftsson, *The Tagged Icelandic Corpus (MM)*.
- Laurette Pretorius and Sonja Bosch, Semi-automated extraction of morphological grammars for Nguni with special reference to Southern Ndebele.
- Björn Gambäck, Tagging and Verifying an Amharic News Corpus.
- Guy De Pauw, Gilles-Maurice de Schryver and Janneke van de Loo. *Resource-Light Bantu Part-of-Speech Tagging*.
- Gulshan Dovudov, Vít Suchomel and Pavel Smerk, *POS Annotated 50M Corpus of Tajik Language*.

# Workshop Organizers

*[Please insert the name(s) and affiliation(s) of the Organizing Committee Members using font Times New Roman, 12 pts]*

| | |
|---|---|
| Guy De Pauw (AfLaT) | CLiPS - Computational Linguistics Group, University of Antwerp, Belgium |
| Gilles-Maurice de Schryver (AfLaT) | African Languages and Cultures, Ghent University, Belgium<br>Xhosa Department, University of the Western Cape, South Africa |
| Mikel L. Forcada (SaLTMiL) | Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Spain |
| Kepa Sarasola (SaLTMiL) | Dept. of Computer Languages, University of the Basque Country |
| Francis M. Tyers (SaLTMiL) | Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Spain |
| Peter Waiganjo Wagacha (AfLaT) | School of Computing & Informatics, University of Nairobi, Kenya |

# Workshop Programme Committee

| | |
|---|---|
| Iñaki Alegria | University of the Basque Country, Spain |
| Nuria Bel | Universitat Pompeu Fabra, Barcelona, Spain |
| Lars Borin | Göteborgs universitet, Sweden |
| Sonja Bosch | University of South Africa, South Africa |
| Khalid Choukri | ELRA/ELDA, France |
| Guy De Pauw | Universiteit Antwerpen, Belgium |
| Gilles-Maurice de Schryver | Universiteit Gent |
| Mikel L. Forcada | Universitat d'Alacant, Spain |
| Dafydd Gibbon | Universität Bielefeld, Germany |
| Lori Levin | Carnegie Mellon University, USA |
| Hrafn Loftsson | University of Reykjavik, Iceland |
| Girish Nath Jha | Jawaharlal Nehru University, India |
| Odétúnjí Odéjobi | Obafemi Awolowo University, Nigeria |
| Laurette Pretorius | University of South Africa, South Africa |
| Benoît Sagot | INRIA, France |
| Felipe Sánchez-Martínez | Universitat d'Alacant, Spain |
| Kepa Sarasola | University of the Basque Country, Spain |
| Kevin Scannell | Saint Louis University, United States |
| Trond Trosterud | University of Tromsø, Norway |
| Francis M. Tyers | Universitat d'Alacant, Spain |
| Peter Waiganjo Wagacha | University of Nairobi, Kenya |

# Preface

The 8th International Workshop of the ISCA Special Interest Group on Speech and Language Technology for Minority Languages (SALTMIL)1 and the Fourth Workshop on African Language Technology (AfLaT2012)2 are jointly held as part of the 2012 International Language Resources and Evaluation Conference (LREC 2012). Entitled “Language technology for normalisation of less-resourced languages”, the workshop is intended to continue the series of SALTMIL/LREC workshops on computational language resources for minority languages, held in Granada (1998), Athens (2000), Las Palmas de Gran Canaria (2002), Lisbon (2004), Genoa (2006), Marrakech (2008) and Malta (2010), and the series of AfLaT workshops, held in Athens (EACL2009), Malta (LREC2010) and Addis Ababa (AGIS11).

The Istanbul 2012 workshop aims to share information on tools and best practices, so that isolated researchers will not need to start their work from scratch. An important aspect will be the forming of personal contacts, which can minimize duplication of effort. There will be a balance between presentations of existing language resources, and more general presentations designed to give background information needed by all researchers.

While less-resourced languages and minority languages often struggle to find their place in a digital world dominated by only a handful of commercially interesting languages, a growing number of researchers are working on alleviating this linguistic digital divide, through localisation efforts, the development of BLARKs (basic language resource kits) and practical applications of human language technologies. The joint SaLTMiL/AfLaT workshop on “Language technology for normalisation of less-resourced languages” provides a unique opportunity to connect these researchers and set up a common forum to meet and share the latest developments in the field.

The workshop takes an inclusive approach to the word normalisation, considering it to include both technologies that help make languages more “normal” in society and everyday life, as well as technologies that normalise languages, i.e. help create or maintain a written standard or support diversity in standards. We particularly focus on the challenges less-resourced and minority languages face in the digital world.

## Resource Creation

Tuesday 22 May, 11:00 – 13:00
Chairperson: *Francis M. Tyers*

### Issues in Designing a Spoken Corpus of Irish

*Elaine Uí Dhonnchadha, Alessio Frenda and Brian Vaughan*

Abstract
This paper describes the stages involved in implementing a corpus of spoken Irish. This pilot project (consisting of approximately 140K words of transcribed data) implements part of the design of a larger corpus of spoken Irish which it is hoped will contain approximately 2 million words when complete. It hoped that such a corpus will provide material for linguistic research, lexicography, the teaching of Irish and for development of language technology for the Irish language.

### Learning Morphological Rules for Amharic Verbs Using Inductive LogicProgramming

*Wondwossen Mulugeta and Michael Gasser*

Abstract
This paper presents a supervised machine learning approach to morphological analysis of Amharic verbs. We use Inductive Logic Programming (ILP), implemented in CLOG. CLOG learns rules as a first order predicate decision list. Amharic, an under-resourced African language, has very complex inflectional and derivational verb morphology, with four and five possible prefixes and suffixes respectively. While the affixes are used to show various grammatical features, this paper addresses only subject prefixes and suffixes. The training data used to learn the morphological rules are manually prepared according to the structure of the background predicates used for the learning process. The training resulted in 108 stem extraction and 19 root template extraction rules from the examples provided. After combining the various rules generated, the program has been tested using a test set containing 1,784 Amharic verbs. An accuracy of 86.99% has been achieved, encouraging further application of the method for complex Amharic verbs and other parts of speech.

### The Database of Modern Icelandic Inflection

*Kristín Bjarnadóttir*

Abstract
The topic of this paper is the Database of Modern Icelandic Inflection (DMII), containing about 270,000 paradigms from Modern Icelandic, with over 5.8 million inflectional forms. The DMII was created as a multipurpose resource, for use in language technology, lexicography, and as an online resource for the general public. Icelandic is a morphologically rich language with a complex inflectional system, commonly exhibiting idiosyncratic inflectional variants. In spite of a long history of morphological research, none of the available sources had the necessary information for the making of a comprehensive and productive rule-based system with the coverage needed. Thus, the DMII was created as a database of paradigms showing all and only the inflectional variants of each word. The initial data used for the project was mostly lexicographic. The creation of a 25 million token corpus of Icelandic, the MÍM Corpus, has made it possible to use empirical data in the development of the DMII, resulting in extensive additions to the vocabulary. The data scarcity in the corpus, due to the enormous number of possible inflectional forms, proves how important it is to use both lexicographic data and a corpus to complement each other in an undertaking such as the DMII.

**Natural Language Processing for Amazigh Language: Challenges and Future Directions**

*Fadoua Ataa Allah and Siham Boulaknadel*

Abstract
Amazigh language, as one of the indo-European languages, poses many challenges on natural language processing. The writing system, the morphology based on unique word formation process of roots and patterns, and the lack of linguistic corpora make computational approaches to Amazigh language challenging.
In this paper, we give an overview of the current state of the art in Natural Language Processing for Amazigh language in Morocco, and we suggest the development of other technologies needed for the Amazigh language to live in "information society".

## Resource Use

Tuesday 22 May, 14:00 – 16:00
Chairperson: *Guy De Pauw*

**Compiling Apertium morphological dictionaries with HFST and using them in HFST applications.**

*Tommi A. Pirinen and Francis M. Tyers*

Abstract
In this paper we aim to improve interoperability and re-usability of the morphological dictionaries of Apertium machine translation system by formulating a generic finite-state compilation formula that is implemented in HFST finite-state system to compile Apertium dictionaries into general purpose finite-state automata. We demonstrate the use of the resulting automaton in FST-based spell-checking system.

**Automatic structuring and correction suggestion system for Hungarian clinical records.**

*Borbóla Siklósi, György Orosz, Attila Novák and Gábor Prószéky*

Abstract
The first steps of processing clinical documents are structuring and normalization. In this paper we demonstrate how we compensate the lack of any structure in the raw data by transforming simple formatting features automatically to structural units. Then we developed an algorithm to separate running text from tabular and numerical data. Finally we generated correcting suggestions for word forms recognized to be incorrect. Some evaluation results are also provided for using the system as automatically correcting input texts by choosing the best possible suggestion from the generated list. Our method is based on the statistical characteristics of our Hungarian clinical data set and on the HUMor Hungarian morphological analyzer. The conclusions claim that our algorithm is not able to correct all mistakes by itself, but is a very powerful tool to help manually correcting Hungarian medical texts in order to produce a correct text corpus of such a domain.

**Constraint Grammar based Correction of Grammatical Errors for North Sámi.**

*Linda Wiechetek*

Abstract
The article describes a grammar checker prototype for North Sámi, a language with agglutinative and inflective features. The grammar checker has been constructed using the rule-based Constraint Grammar formalism. The focus is on the setup of a prototype and diagnosing and correcting grammatical case errors, mostly those that appear with adpositions. Case errors in writing are typical even for native speakers as case errors can result from spelling mistakes. Typical candidates for spelling mistakes are forms containing the letter á and those with double consonants. Alternating double and single consonants is a possible case marker. Case errors in an adpositional phrase are common mistakes. Adpositions are typically homonymous (preposition, postposition, adverb) and ask for a genitive case to the left or right of it. Therefore, finding case errors requires a disambiguation of the adposition itself, a correct dependency mapping between the adposition and its dependent and a diagnosis of the case error, which can require homonymy disambiguation of the dependent itself. A deep linguistic analysis including a module for disambiguation, syntactic analysis and dependency annotation is necessary for correcting case errors in adpositional phrases.

**Toward a Rule-Based System for English-Amharic Translation.**

*Michael Gasser*

Abstract
We describe key aspects of an ongoing project to implement a rule-based English-to-Amharic and Amharic-to-English machine translation system within our L3 framework. L3 is based on Extensible Dependency Grammar (Debusmann, 2007), a multi-layered dependency grammar formalism that relies on constraint satisfaction for parsing and generation. In L3 , we extend XDG to multiple languages and translation. This requires a mechanism to handle cross-lingual relationships and mismatches in the number of words between source and target languages. In this paper, we focus on these features as well as the advantages that L3 offers for handling structural divergences between English and Amharic and its capacity to accommodate shallow and deep translation within a single system.

## Poster session

Tuesday 22 May, 16:30 – 17:30
Chairperson:

**Technological Tools for Dictionary and Corpora Building for Minority Languages: Example of the French-based Creoles.**

*Paola Carrión González and Emmanuel Cartier*

Abstract
In this paper, we present a project which aims at building and maintaining a lexicographical resource of contemporary French-based creoles, still considered as minority languages, especially those situated in American-Caribbean zones. These objectives are achieved through three main steps: 1) Compilation of existing lexicographical resources (lexicons and digitized dictionaries, available on the Internet); 2) Constitution of a corpus in Creole languages with literary, educational and journalistic documents, some of them retrieved automatically with web spiders; 3) Dictionary maintenance: through automatic morphosyntactic analysis of the corpus and determination of the

frequency of unknown words. Those unknown words will help us to improve the database by searching relevant lexical resources that we had not included before. This final task could be done iteratively in order to complete the database and show language variations within the same Creole-speaking community. Practical results of this work will consist in 1/ A lexicographical database, explicitating variations in French-based creoles, as well as helping normalizing the written form of this language; 2/ An annotated corpora that could be used for further linguistic research and NLP applications.

**Describing Morphologically-rich Languages using Metagrammars: a Look at Verbs in Ikota.**

*Denys Duchier, Brunelle Magnana Ekoukou, Yannick Parmentier, Simon Petitjean and Emannuel Schang.*

Abstract
In this paper, we show how the concept of metagrammar originally introduced by Candito (1996) to design large Tree-Adjoining Grammars describing the syntax of French and Italian, can be used to describe the morphology of Ikota, a Bantu language spoken in Gabon. Here, we make use of the expressivity of the XMG (eXtensible MetaGrammar) formalism to describe the morphological variations of verbs in Ikota. This XMG specification captures generalizations over these morphological variations. In order to produce the inflected forms, one can compile the XMG specification, and save the resulting electronic lexicon in an XML file, thus favorising its reuse in dedicated applications.

**A Corpus of Santomé.**

*Tjerk Hagemeijer, Iris Hendrickx, Abigail Tiny and Haldane Amaro.*

Abstract
We present the process of constructing a corpus of spoken and written material for Santome, a Portuguese-related creole language spoken on the island of S. Tomé in the Gulf of Guinea (Africa). Since the language lacks an official status, we faced the typical difficulties, such as language variation, lack of standard spelling, lack of basic language instruments, and only a limited data set. The corpus comprises data from the second half of the 19 th century until the present. For the corpus compilation we followed corpus linguistics standards and used UTF-8 character encoding and XML to encode meta information. We discuss how we normalized all material to one spelling, how we dealt with cases of language variation, and what type of meta data is used. We also present a POS-tag set developed for the Santome language that will be used to annotate the data with linguistic information.

**The Tagged Icelandic Corpus (MM).**

*Sigrún Helgadóttir, Asta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir and Hrafn Loftsson,*

Abstract
In this paper, we describe the development of a morphosyntactically tagged corpus of Icelandic, the MÍM corpus. The corpus consists of about 25 million tokens of contemporary Icelandic texts collected from varied sources during the years 2006–2010. The corpus is intended for use in Language Technology projects and for linguistic research. We describe briefly other Icelandic corpora and how they differ from the MÍM corpus. We describe the text selection and collection for MÍM, both for written and spoken text, and how metadata was created. Furthermore, copyright

issues are discussed and how permission clearance was obtained for texts from different sources. Text cleaning and annotation phases are also described. The corpus is available for search through a web interface and for download in TEI-conformant XML format. Examples are given of the use of the corpus and some spin-offs of the corpus project are described. We believe that the care with which we secured copyright clearance for the texts will make the corpus a valuable resource for Icelandic Language Technology projects. We hope that our work will inspire those wishing to develop similar resources for less-resourced languages.

**Semi-automated extraction of morphological grammars for Nguni with special reference to Southern Ndebele.**

*Laurette Pretorius and Sonja Bosch.*

Abstract
A finite-state morphological grammar for Southern Ndebele, a seriously under-resourced language, has been semi-automatically obtained from a general Nguni morphological analyser, which was bootstrapped from a mature hand-written morphological analyser for Zulu. The results for Southern Ndebele morphological analysis, using the Nguni analyser, are surprisingly good, showing that the Nguni languages (Zulu, Xhosa, Swati and Southern Ndebele) display significant cross-linguistic similarities that can be exploited to accelerate documentation, resource-building and software development. The project embraces recognized best practices for the encoding of resources to ensure sustainability, access, and easy adaptability to future formats, lingware packages and development platforms.

**Tagging and Verifying an Amharic News Corpus.**

*Björn Gambäck.*

Abstract
The paper describes work on verifying, correcting and retagging a corpus of Amharic news texts. A total of 8715 Amharic news articles had previously been collected from a web site, and part of the corpus (1065 articles; 210,000 words) then morphologically analysed and manually part-of-speech tagged. The tagged corpus has been used as the basis for testing the application to Amharic of machine learning techniques and tools developed for other languages. This process made it possible to spot several errors and inconsistencies in the corpus which has been iteratively refined, cleaned, normalised, split into folds, and partially re-tagged by both automatic and manual means.

**Resource-Light Bantu Part-of-Speech Tagging.**

*Guy De Pauw, Gilles-Maurice de Schryver and Janneke van de Loo.*

Abstract
Recent scientific publications on data-driven part-of-speech tagging of Sub-Saharan African languages have reported encouraging accuracy scores, using off-the-shelf tools and often fairly limited amounts of training data. Unfortunately, no research efforts exist that explore which type of linguistic features contribute to accurate part-of-speech tagging for the languages under investigation. This paper describes feature selection experiments with a memory-based tagger, as well as a resource-light alternative approach. Experimental results show that contextual information is often not strictly necessary to achieve a good accuracy for tagging Bantu languages and that decent results can be achieved using a very straightforward unigram approach, based on orthographic features.

**POS Annotated 50M Corpus of Tajik Language.**

*Gulshan Dovudov, Vít Suchomel and Pavel Smerk.*

Abstract
Paper presents by far the largest available computer corpus of Tajik language of the size of more than 50 million words. To obtain the texts for the corpus two different approaches were used and the paper offers a description of both of them. Then the paper describes a newly developed morphological analyzer of Tajik and presents some statistics of its application on the corpus.