Introduction

Created in April 2017, SIGUL (http://www.elra.info/en/sig/sigul/) is a joint Special Interest Group of the European Language Resources Association (ELRA) and of the International Speech Communication Association (ISCA). SIGUL intends to bring together a number of professionals involved in the development of language resources and technologies for under-resourced languages. Its main objective is to build a community that not only supports linguistic diversity through technology and ICT but also commits to increase the lesser-resourced languages (regional, minority, or endangered) chances to survive the digital world through language and speech technology.

Before the creation of SIGUL, two workshops addressed language technologies for low resource languages: there have been 6 editions of SLTU (Spoken Language Technologies for Under-resourced languages) which started in 2008; and 3 editions of CCURL (Collaboration and Computing for Under-Resourced Languages) which started in 2014. For 2020, and as a satellite event of LREC, SIGUL board decided to organize the 1st Joint Workshop of SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) (SLTU-CCURL 2020).

We solicited papers related to all areas of natural language processing, speech and computational linguistics, as well as those at the intersection with digital humanities and documentary linguistics, provided that they address less-resourced languages. One goal of this workshop was to offer a venue where researchers in different disciplines and from varied backgrounds can fruitfully explore new areas of intellectual and practical development while honoring their common interest of sustaining less-resourced languages.

Our programme committee comprised 60 experts in natural language processing and spoken language processing from 19 countries. Each of the 64 submitted papers was reviewed by 3 committee members. We finally accepted 54 papers for the proceedings. We would like to express our sincere thanks to all members of this committee (who worked hard despite the difficult conditions associated with the pandemic) and authors for their great work in making this event a scientifically recognised international Workshop. We would also like to extend our thanks to all our sponsors: Google as platinum sponsor; ELRA, ISCA and ACL/SIGEL for endorsing this event.

Unfortunately, as a consequence of the COVID-19 pandemic, LREC 2020 has been canceled and - as a satellite event of the Conference - SLTU-CCURL 2020 has been canceled as well. We nevertheless hope that you will find these workshop proceedings relevant and stimulating for your own research. We are looking forward to see you soon for future events organised by SIGUL.

SLTU-CCURL-2020 Workshop co-chairs:

Dorothee Beermann (NTNU, Norway)
Laurent Besacier (LIG – Université Grenoble Alpes, France)
Sakriani Sakti (NAIST, Japan)
Claudia Soria (CNR-ILC, Italy)
Organizers

Dorothee Beermann (NTNU, Norway)
Laurent Besacier (LIG – Université Grenoble Alpes, France)
Sakriani Sakti (NAIST, Japan)
Claudia Soria (CNR-ILC, Italy)

Program Committee:

Adrian Doyle (University of Galway, Ireland)
Alexey Karpov (SPIIRAS, Russian Federation)
Alexis Palmer (University of North Texas, USA)
Amir Aharoni (Wikimedia Foundation)
Andras Kornai (Hungarian Academy of Sciences, Hungary)
Angelo Mario Del Grosso (CNR-ILC, Italy)
Antti Arppe (University of Alberta, Canada)
Atticus Harrigan (University of Alberta, Canada)
Charl Van Heerden (Saigen, South Africa)
Daan Van Esch (Google)
Dafydd Gibbon (Bielefeld University, Germany)
Delyth Prys (Bangor University, UK)
Dewi Bryn Jones (Bangor University, UK)
Dorothee Beermann (NTNU, Norway)
Emily Le Chen (University of Illinois, USA)
Federico Boschetti (CNR-ILC, Italy)
Francis Tyers (Indiana University, USA)
Gerard Bailly (GIPSA Lab, CNRS)
Gilles Adda (LIMSI/IMMI CNRS, France)
Heysem Kaya (Utrecht University, The Netherlands)
Hyunjí “Hayley” Park (University of Illinois at Urbana-Champaign, USA)
Irina Kipyatkova (SPIIRAS, Russia)
Jeff Good (University at Buffalo, USA)
Jelske Dijkstra (Fryske Akademy, The Netherlands)
John Judge (ADAPT DCU, Ireland)
John Philip McCrae (National University of Ireland Galway, Ireland)
Jonas Fromseier Mortensen (Google)
Jordan Lachler (University of Alberta, Canada)
Joseph Mariani (LIMSI-CNRS, France)
Katherine Schmirler (University of Alberta, Canada)
Kepa Sarasola (University of the Basque Country, Spain)
Kevin Scannell (Saint Louis University, Missouri, USA)
Klara Ceberio (Elhuyar, Spain)
Lane Schwartz (University of Illinois at Urbana-Champaign, USA)
Lars Hellan (NTNU, Norway)
Lars Steinert (University of Bremen, Germany)
Laurent Besacier (LIG-IMAG, France)
Maite Melero (Barcelona Supercomputing Center, Spain)
Marcely Zanon Boito (LIG-IMAG, France)
Mathieu Mangeot-Nagata (LIG-IMAG, France)
Matt Coler (University of Groningen, The Netherlands)
Mohammad A. M. Abushariah (The University of Jordan, Jordan)
Nick Thieberger (University of Melbourne / ARC Centre of Excellence for the Dynamics of Language, Australia)
Omar Farooq (AMU, India)
Pierric Sans (Google)
Pradip K Das (IIT, India)
Richard Littauer (University of Saarland, Germany)
Sahar Ghannay (LIMSI, CNRS, France)
Sakriani Sakti (NAIST, Japan)
Satoshi Nakamura (NAIST, Japan)
Sebastian Stüker (KIT, Germany)
Shyam S Agrawal (KIIT, India)
Sjur Moshagen (UiT The Arctic University of Norway, Norway)
Solomon Teferra Abate (Addis Ababa University, Ethiopia)
Steven Bird (Charles Darwin University, Australia)
Tanja Schultz (Uni-Bremen, Germany)
Thang Vu (Uni-Stuttgart, Germany)
Teresa Lynn (ADAPT Centre, Ireland)
Trond Trosterud (Tromsø University, Norway)
Win Pa Pa (UCS Yangon, Myanmar)
# Table of Contents

**Neural Models for Predicting Celtic Mutations**  
Kevin Scannell ................................................................. 1

Alexander Gutkin .......................................................... 9

**Open-Source High Quality Speech Datasets for Basque, Catalan and Galician**  
Oddur Kjartansson, Alexander Gutkin, Alena Butryna, Isin Demirsahin and Clara Rivera .... 21

**Two LRL & Distractor Corpora from Web Information Retrieval and a Small Case Study in Language Identification without Training Corpora**  
Armin Hoenen, Cemre Koc and Marc Rahn .................................. 28

**Morphological Disambiguation of South Sámi with FSTs and Neural Networks**  
Mika Hämäläinen and Linda Wiechetek .................................. 36

**Effects of Language Relatedness for Cross-lingual Transfer Learning in Character-Based Language Models**  
Mittul Singh, Peter Smit, Sami Virpioja and Mikko Kurimo ................................. 41

**Multilingual Graphemic Hybrid ASR with Massive Data Augmentation**  
Chunxi Liu, Qiaochu Zhang, Xiaohui Zhang, Kritika Singh, Yatharth Saraf and Geoffrey Zweig 46

**Neural Text-to-Speech Synthesis for an Under-Resourced Language in a Diglossic Environment: the Case of Gascon Occitan**  
Ander Corral, Igor Leturia, Aure Séguier, Michäel Barret, Benaset Dazéas, Philippe Boula de Mareüil and Nicolas Quint ................................................................. 53

**Transfer Learning for Less-Resourced Semitic Languages Speech Recognition: the Case of Amharic**  
Yonas Woldemariam ............................................................ 61

**Semi-supervised Acoustic Modelling for Five-lingual Code-switched ASR using Automatically-segmented Soap Opera Speech**  
Nick Wilkinson, Astik Biswas, Emre Yilmaz, Febe De Wet, Ewald Van der westhuizen and Thomas Niesler ................................................................. 70

**Investigating Language Impact in Bilingual Approaches for Computational Language Documentation**  
Marcely Zanon Boito, Aline Villavicencio and Laurent Besacier ................. 79

**Design and evaluation of a smartphone keyboard for Plains Cree syllabics**  
Eddie Santos and Atticus Harrigan ........................................ 88

**MultiSeg: Parallel Data and Subword Information for Learning Bilingual Embeddings in Low Resource Scenarios**  
Efsun Sarioglu Kayi, Vishal Anand and Smaranda Muresan .......................... 97

**Poio Text Prediction: Lessons on the Development and Sustainability of LTs for Endangered Languages**  
Gema Zamora Fernández, Vera Ferreira and Pedro Manha ........................... 106

**Text Corpora and the Challenge of Newly Written Languages**  
Alice Millour and Karën Fort ................................................ 111
Adapting a Welsh Terminology Tool to Develop a Cornish Dictionary
Delyth Prys ................................................................. 235

Multiple Segmentations of Thai Sentences for Neural Machine Translation
Alberto Poncelas, Wichaya Pidchamook, Chao-Hong Liu, James Hadley and Andy Way .... 240

Automatic Extraction of Verb Paradigms in Regional Languages: the case of the Linguistic Crescent varieties
elena knyazeva, Gilles Adda, Philippe Boula de Mareüil, Maximilien Guérin and Nicolas Quint245

FST Morphology for the Endangered Skolt Sami Language
Jack Rueter and Mika Hamalainen .................................................. 250

Voted-Perceptron Approach for Kazakh Morphological Disambiguation
Gulmira Tolegen, Alymzhan Toleu and Rustam Mussabayev ........................................... 258

DNN-Based Multilingual Automatic Speech Recognition for Wolaytta using Oromo Speech
Martha Yifiru Tachbelie, Solomon Tefera Abate and Tanja Schultz .................................... 265

Building Language Models for Morphological Rich Low-Resource Languages using Data from Related Donor Languages: the Case of Uyghur
Ayimunishagu Abulimiti and Tanja Schultz .................................................. 271

Basic Language Resources for 31 Languages (Plus English): The LORELEI Representative and Incident Language Packs
Jennifer Tracey and Stephanie Strassel ................................................................. 277

On the Exploration of English to Urdu Machine Translation
Sadaf Abdul Rauf, Syeda Abida, Noor-e- Hira, Syeda Zahra, Dania Parvez, Javeria Bashir and Qurat-ul-ain Majid ................................................................. 285

Developing a Twi (Asante) Dictionary from Akan Interlinear Glossed Texts
Dorothee Beermann, Lars Hellan, Pavel Mihaylov and Anna Struck .................................... 294

Adapting Language Specific Components of Cross-Media Analysis Frameworks to Less-Resourced Languages: the Case of Amharic
Yonas Woldemariam and Adam Dahlgren ................................................................. 298

Phonemic Transcription of Low-Resource Languages: To What Extent can Preprocessing be Automated?
Guillaume Wisniewski, Séverine Guillaume and Alexis Michaud ........................................ 306

Manual Speech Synthesis Data Acquisition - From Script Design to Recording Speech
Atli Sigurgeirsson, Gunnar Örnölfsson and Jón Guðnason .................................................. 316

Owóksape - An Online Language Learning Platform for Lakota
Jan Ullrich, Elliot Thornton, Peter Vieira, Logan Swango and Marek Kupiec ......................... 321

A Corpus of the Sorani Kurdish Folkloric Lyrics
Sina Ahmadi, Hossein Hassani and Kamaladdin Abedi .......................................................... 330

Improving the Language Model for Low-Resource ASR with Online Text Corpora
Nils Hjortnaes, Timofey Arkhangelskiy, Niko Partanen, Michael Rießler and Francis Tyers ... 336
A Summary of the First Workshop on Language Technology for Language Documentation and Revitalization

"A Passage to India": Pre-trained Word Embeddings for Indian Languages
Saurav Kumar, Saunack Kumar, Diptesh Kanojia and Pushpak Bhattacharyya ............... 352

A Counselling Corpus in Cantonese
John Lee, Tianyuan Cai, Wenxiu Xie and Lam Xing ................................................. 358

Speech Transcription Challenges for Resource Constrained Indigenous Language Cree
Vishwa Gupta and Gilles Boulianne ................................................................. 362

Turkish Emotion Voice Database (TurEV-DB)
Salih Firat Canpolat, Zuhal Ormanoğlu and Deniz Zeyrek ................................. 368