NAACL-HLT 2021

Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)

Proceedings of the First Workshop

June 11, 2021
Preface

This area is in all probability unmatched, anywhere in the world, in its linguistic multiplicity and diversity. A couple of thousand languages and dialects, at present divided into 17 large families and 38 small ones, with several hundred unclassified single languages, are on record. In one small portion of the area, in Mexico just north of the Isthmus of Tehuantepec, one finds a diversity of linguistic type hard to match on an entire continent in the Old World.

—McQuown (1955)
Workshop Organizers:
Manuel Mager
Arturo Oncevay
Annette Rios
Ivan Vladimir Meza Ruiz
Alexis Palmer
Graham Neubig
Katharina Kann

Shared Task Organizers:
Manuel Mager
Arturo Oncevay
Abteen Ebrahimi
John Ortega
Annette Rios
Angela Fan
Ximena Gutierrez-Vasques
Luis Chiruzzo
Gustavo A. Giménez-Lugo
Ricardo Ramos
Ivan Vladimir Meza Ruiz
Rolando Coto-Solano
Alexis Palmer
Elisabeth Mager
Vishrav Chaudhary
Graham Neubig
Ngoc Thang Vu
Katharina Kann

Program Committee:
Abhilasha Ravichander
Abteen Ebrahimi
Adam Wiemerslage
Alexandra Birch
Alfonso Medina-Urrea
Annette Rios
Antonios Anastasopoulos
Arturo Oncevay
Barry Haddow
Candace Ross
Cynthia Montaño
Daan van Esch
Ekaterina Vylomova
Emily M. Bender
Emily Prud’hommeaux
Eneko Agirre
Fernando Alva-Manchego
Francis Tyers
Gerardo Sierra Martínez
Ivan Vulić
# Table of Contents

**qxoRef 1.0: A coreference corpus and mention-pair baseline for coreference resolution in Conchucos Quechua**  
Elizabeth Pankratz ................................................................. 1

**A corpus of K’iche’ annotated for morphosyntactic structure**  
Francis Tyers and Robert Henderson ........................................ 10

**Investigating variation in written forms of Nahuatl using character-based language models**  
Robert Pugh and Francis Tyers .................................................. 21

**Apurinã Universal Dependencies Treebank**  
Jack Rueter, Marília Fernanda Pereira de Freitas, Sidney Da Silva Facundes, Mika Hämäläinen and Niko Partanen ................................................................. 28

**Automatic Interlinear Glossing for Otomi language**  
Diego Barriga Martínez, Victor Mijangos and Ximena Gutierrez-Vasques ......................... 34

**A survey of part-of-speech tagging approaches applied to K’iche’**  
Francis Tyers and Nick Howell .................................................. 44

**Highland Puebla Nahuatl Speech Translation Corpus for Endangered Language Documentation**  
Jiatong Shi, Jonathan D. Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan and Shinji Watanabe 53

**End-to-End Automatic Speech Recognition: Its Impact on the Workflow in Documenting Yoloxóchitl Mixtec**  
Jonathan D. Amith, Jiatong Shi and Rey Castillo García ............................ 64

**A finite-state morphological analyser for Paraguayan Guaraní**  
Anastasia Kuznetsova and Francis Tyers .......................................... 81

**Morphological Segmentation for Seneca**  
Zoey Liu, Robert Jimerson and Emily Prud’hommeaux ............................. 90

**Representation of Yine [Arawak] Morphology by Finite State Transducer Formalism**  
Adriano Ingunza Torres, John Miller, Arturo Oncevay and Roberto Zariquey Biondi .......... 102

**Leveraging English Word Embeddings for Semi-Automatic Semantic Classification in Nêhiyawêwin (Plains Cree)**  
Atticus Harrigan and Antti Arppe .............................................. 113

**Restoring the Sister: Reconstructing a Lexicon from Sister Languages using Neural Machine Translation**  
Remo Nitschke ............................................................................. 122

**Expanding Universal Dependencies for Polysynthetic Languages: A Case of St. Lawrence Island Yupik**  
Hyunji Park, Lane Schwartz and Francis Tyers ................................. 131

**The More Detail, the Better? – Investigating the Effects of Semantic Ontology Specificity on Vector Semantic Classification with a Plains Cree / nêhiyawêwin Dictionary**  
Daniel Dacanay, Atticus Harrigan, Arok Wolvengrey and Antti Arppe 143
Experiments on a Guarani Corpus of News and Social Media
Santiago Góngora, Nicolás Giossa and Luis Chiruzzo .................................................. 153

Towards a First Automatic Unsupervised Morphological Segmentation for Inuinnaqtun
Ngoc Tan Le and Fatiha Sadat ................................................................. 159

Toward Creation of Ancash Lexical Resources from OCR
Johanna CORDOVA and Damien Nouvel ...................................................... 163

Ayuuk-Spanish Neural Machine Translator
Delfino Zacarías Márquez and Ivan Vladimir Meza Ruiz ........................................... 168

Explicit Tone Transcription Improves ASR Performance in Extremely Low-Resource Languages: A Case Study in Bribri
Rolando Coto-Solano .................................................................................. 173

Towards a morphological transducer and orthography converter for Western Tlacolula Valley Zapotec
Jonathan Washington, Felipe Lopez and Brook Lillehaugen ....................................... 185

Peru is Multilingual, Its Machine Translation Should Be Too?
Arturo Oncevay .......................................................................................... 194

Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas

Open Machine Translation for Low Resource South American Languages (AmericasNLP 2021 Shared Task Contribution)
Shantipriya Parida, Subhadarshi Panda, Amulya Dash, Esau Villatoro-Tello, A. Seza Doğruöz, Rosa M. Ortega-Mendoza, Amadeo Hernández, Yashvardhan Sharma and Petr Motlicek .................................................. 218

NRC-CNRC Machine Translation Systems for the 2021 AmericasNLP Shared Task
Rebecca Knowles, Darlene Stewart, Samuel Larkin and Patrick Littell ................................. 224

Francis Zheng, Machel Reid, Edison Marrese-Taylor and Yutaka Matsuo ........................... 234

The REPU CS’ Spanish–Quechua Submission to the AmericasNLP 2021 Shared Task on Open Machine Translation
Oscar Moreno ............................................................................................. 241

Moses and the Character-Based Random Babbling Baseline: CoAStaL at AmericasNLP 2021 Shared Task
Marcel Bollmann, Rahul Aralikatte, Héctor Murrieta Bello, Daniel Hershcovich, Miryam de Lhonneux and Anders Søgaard .......................................................................... 248

The Helsinki submission to the AmericasNLP shared task
Raúl Vázquez, Yves Scherrer, Sami Virpioja and Jörg Tiedemann ........................................ 255

IndT5: A Text-to-Text Transformer for 10 Indigenous Languages
El Moatez Billah Nagoudi, Wei-Rui Chen, Muhammad Abdul-Mageed and Hasan Cavusoglu . 265
Workshop Program

June 11, 2021

qxoRef 1.0: A coreference corpus and mention-pair baseline for coreference resolution in Conchucos Quechua
Elizabeth Pankratz

A corpus of K’iche’ annotated for morphosyntactic structure
Francis Tyers and Robert Henderson

Investigating variation in written forms of Nahuatl using character-based language models
Robert Pugh and Francis Tyers

Apurinã Universal Dependencies Treebank
Jack Rueter, Marília Fernanda Pereira de Freitas, Sidney Da Silva Facundes, Mika Hämiäläinen and Niko Partanen

Automatic Interlinear Glossing for Otomi language
Diego Barriga Martínez, Victor Mijangos and Ximena Gutierrez-Vasques

A survey of part-of-speech tagging approaches applied to K’iche’
Francis Tyers and Nick Howell

Highland Puebla Nahuatl Speech Translation Corpus for Endangered Language Documentation
Jiatong Shi, Jonathan D. Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan and Shinji Watanabe

End-to-End Automatic Speech Recognition: Its Impact on the Workflow in Documenting Yoloxóchitl Mixtec
Jonathan D. Amith, Jiatong Shi and Rey Castillo García

A finite-state morphological analyzer for Paraguayan Guaraní
Anastasia Kuznetsova and Francis Tyers

Morphological Segmentation for Seneca
Zoey Liu, Robert Jimerson and Emily Prud’hommeaux

Representation of Yine [Arawak] Morphology by Finite State Transducer Formalism
Adriano Ingunza Torres, John Miller, Arturo Oncevay and Roberto Zariquey Biondi
Leveraging English Word Embeddings for Semi-Automatic Semantic Classification in Nêhiyawêwin (Plains Cree)  
Atticus Harrigan and Antti Arppe

Restoring the Sister: Reconstructing a Lexicon from Sister Languages using Neural Machine Translation  
Remo Nitschke

Expanding Universal Dependencies for Polysynthetic Languages: A Case of St. Lawrence Island Yupik  
Hyunji Park, Lane Schwartz and Francis Tyers

The More Detail, the Better? – Investigating the Effects of Semantic Ontology Specificity on Vector Semantic Classification with a Plains Cree / nêhiyawêwin Dictionary  
Daniel Dacanay, Atticus Harrigan, Arok Wolvengrey and Antti Arppe

Experiments on a Guarani Corpus of News and Social Media  
Santiago Góngora, Nicolás Giossa and Luis Chiruzzo

Towards a First Automatic Unsupervised Morphological Segmentation for Inuinnaqtun  
Ngoc Tan Le and Fatiha Sadat

Toward Creation of Ancash Lexical Resources from OCR  
Johanna CORDOVA and Damien Nouvel

Ayuuk-Spanish Neural Machine Translator  
Delfino Zacarías Márquez and Ivan Vladimir Meza Ruiz

Explicit Tone Transcription Improves ASR Performance in Extremely Low-Resource Languages: A Case Study in Bribri  
Rolando Coto-Solano

Towards a morphological transducer and orthography converter for Western Tlacolula Valley Zapotec  
Jonathan Washington, Felipe Lopez and Brook Lillehaugen

Peru is Multilingual, Its Machine Translation Should Be Too?  
Arturo Oncevay

Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas  
Open Machine Translation for Low Resource South American Languages (AmericasNLP 2021 Shared Task Contribution)
Shantipriya Parida, Subhadarshi Panda, Amulya Dash, Esau Villatoro-Tello, A. Seza Doğruöz, Rosa M. Ortega-Mendoza, Amadeo Hernández, Yashvardhan Sharma and Petr Motlicek

NRC-CNRC Machine Translation Systems for the 2021 AmericasNLP Shared Task
Rebecca Knowles, Darlene Stewart, Samuel Larkin and Patrick Littell

Francis Zheng, Machel Reid, Edison Marrese-Taylor and Yutaka Matsuo

The REPU CS’ Spanish–Quechua Submission to the AmericasNLP 2021 Shared Task on Open Machine Translation
Oscar Moreno

Moses and the Character-Based Random Babbling Baseline: CoASiL at AmericasNLP 2021 Shared Task
Marcel Bollmann, Rahul Aralikatte, Héctor Murrieta Bello, Daniel Hershcovich, Miriam de Lhoneux and Anders Søgaard

The Helsinki submission to the AmericasNLP shared task
Raúl Vázquez, Yves Scherrer, Sami Virpioja and Jörg Tiedemann

IndT5: A Text-to-Text Transformer for 10 Indigenous Languages
El Moatez Billah Nagoudi, Wei-Rui Chen, Muhammad Abdul-Mageed and Hasan Cavusoglu

IGT2P: From Interlinear Glossed Texts to Paradigms
Sarah Moeller, Ling Liu, Changbing Yang, Katharina Kann and Mans Hulden

An FST Morphological Analyzer for the Gitksan language
Clarissa Forbes, Garrett Nicolai and Miikka Silfverberg

Tackling the Low-resource Challenge for Canonical Segmentation
Manuel Mager, Özlem Çetinoğlu and Katharina Kann

Fortification of Neural Morphological Segmentation Models for Polysynthetic Minimal-Resource Languages
Katharina Kann, Manuel Mager, Ivan Vladimir Meza Ruiz and Hinrich Schütze

Broadening Text Resources for K’iche’
Dominique O’Donnell and Constantine Lignos