

6th Workshop on Research in Computational Typology and Multilingual NLP (SIGTYP 2024)

St. Julian's, Malta
22 March 2024

ISBN: 978-1-7138-9461-2

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571

Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2024) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2025)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

<i>Syntactic dependency length shaped by strategic memory allocation</i> Weijie Xu and Richard Futrell	1
<i>GUIDE: Creating Semantic Domain Dictionaries for Low-Resource Languages</i> Jonathan Janetzki, Gerard De Melo, Joshua Nemecek and Daniel Lee Whitenack	10
<i>A New Dataset for Tonal and Segmental Dialectometry from the Yue- and Pinghua-Speaking Area</i> Ho Wang Matthew Sung, Jelena Prokic and Yiya Chen	25
<i>A Computational Model for the Assessment of Mutual Intelligibility Among Closely Related Languages</i> Jessica Nieder and Johann-Mattis List	37
<i>Predicting Mandarin and Cantonese Adult Speakers' Eye-Movement Patterns in Natural Reading</i> LI Junlin, Yu-Yin Hsu, Emmanuele Chersoni and Bo Peng	44
<i>The Typology of Ellipsis: A Corpus for Linguistic Analysis and Machine Learning Applications</i> Damir Cavar, Ludovic Mompelat and Muhammad S. Abdo	46
<i>Language Atlas of Japanese and Ryukyuan (LAJaR): A Linguistic Typology Database for Endangered Japonic Languages</i> Kanji Kato, So Miyagawa and Natsuko Nakagawa	55
<i>GTNC: A Many-To-One Dataset of Google Translations from NewsCrawl</i> Damiaan J W Reijnaers and Charlotte Pouw	58
<i>Sociolinguistically Informed Interpretability: A Case Study on Hinglish Emotion Classification</i> Kushal Tatariya, Heather Lent, Johannes Bjerva and Miryam De Lhoneux	66
<i>A Call for Consistency in Reporting Typological Diversity</i> Wessel Poelman, Esther Ploeger, Miryam De Lhoneux and Johannes Bjerva	75
<i>Are Sounds Sound for Phylogenetic Reconstruction?</i> Luise Häuser, Gerhard Jäger, Johann-Mattis List, Taraka Rama and Alexandros Stamatakis ...	78
<i>Compounds in Universal Dependencies: A Survey in Five European Languages</i> Emil Svoboda and Magda Ševčíková	88
<i>Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens</i> Nay San, Georgios Paraskevopoulos, Aryaman Arora, Xiluo He, Prabhjot Kaur, Oliver Adams and Dan Jurafsky	100
<i>ModeLing: A Novel Dataset for Testing Linguistic Reasoning in Language Models</i> Nathan Andrew Chi, Teodor Malchev, Riley Kong, Ryan Andrew Chi, Lucas Huang, Ethan A Chi, R. Thomas McCoy and Dragomir Radev	113
<i>TartuNLP @ SIGTYP 2024 Shared Task: Adapting XLM-RoBERTa for Ancient and Historical Languages</i> Aleksei Dorkin and Kairit Sirts	120
<i>Heidelberg-Boston @ SIGTYP 2024 Shared Task: Enhancing Low-Resource Language Analysis With Character-Aware Hierarchical Transformers</i> Frederick Riemenschneider and Kevin Krahn	131

<i>UDParse @ SIGTYP 2024 Shared Task : Modern Language Models for Historical Languages</i>	
Johannes Heinecke	142
<i>Allen Institute for AI @ SIGTYP 2024 Shared Task on Word Embedding Evaluation for Ancient and Historical Languages</i>	
Lester James V. Miranda	151
<i>Findings of the SIGTYP 2024 Shared Task on Word Embedding Evaluation for Ancient and Historical Languages</i>	
Oksana Dereza, Adrian Doyle, Priya Rani, Atul Ojha, Pádraic Moran and John McCrae	160