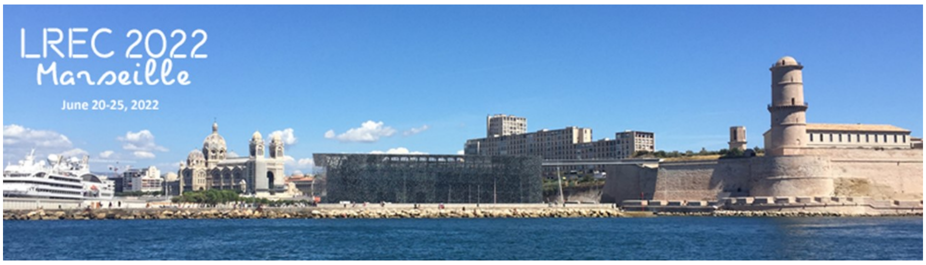

LREC 2022 Workshop
Language Resources and Evaluation Conference
20-25 June 2022

2nd Workshop on Novel Incentives in Data Collection from People: models, implementations, challenges and results (NIDCP 2022)

PROCEEDINGS

Editors:
James Fiumara, Christopher Cieri, Mark Liberman, Chris Callison-Burch

2nd Workshop on Novel Incentives in Data Collection from People: models, implementations, challenges and results (NIDCP 2022)

Edited by: James Fiumara, Christopher Cieri, Mark Liberman, Chris Callison-Burch

ISBN: 978-2-493814-05-0
EAN: 9782493814050

For more information:
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

Introduction

The many research communities that rely upon Language Resources (LR) have benefitted from massive contributions from data centers, government agencies and research groups around the world. Nevertheless, research potential remains largely untapped because the LRs that fuel development fall far short of need as measured by volume, data type, and language coverage. Searches for data sets regularly go unfulfilled even for the dozen languages with the greatest populations and gross linguistic products.

Notwithstanding advances in data collection and processing, the supply of LRs continues to lag behind need in part because of the limited incentive models employed. Throughout the history of LR development, the commonest incentives offered to people in exchange for their contributions of raw language data and judgements were monetary. Perhaps this tendency is based on convenience or perhaps it reflects a belief concerning the ethics of data contribution. In any case, that bias has limited the LR user communities' ability to collect data for example: in the absence of ready funding, in situations where funding cannot easily be transferred and, from groups, such as indigenous communities, with other motivations. The focus on monetary incentives has also limited opportunities to understand how other incentives might attract different workforces, what kinds of workflows might be optimal for such workforces and how their contributions could be integrated into research and technology development efforts.

Social media in contrast has employed a wider range of incentives including: access to information and entertainment; possibilities for self-expression, sharing and publicizing intellectual or creative work; chances to vent frustrations or convey thoughts sometimes anonymously; forums for socializing; situations in which to develop competence that may lead to new prospects; competition, status, prestige, and recognition; payment or discounts in real and virtual worlds; access to services and infrastructure based on contributions; opportunities to contribute to a greater cause or good.

Within HLT communities there have been a few projects that employ these incentives. SPICE provided contributors with access to a speech recognition system that was built from their own contributions. Let's Go improved access to public transit. Herme offered the unusual experience of interacting with a tiny, cute robot. Crowd Curio offered experiential learning of e.g. historical linguistic behaviors. "On Everyone's Mind and Lips" mapped the linguistic landscape of Austria. LanguageARC offers citizen linguists opportunities to contribute to research on timely issues such as bias in public discourse, documenting under-resourced languages and building normative models that can be used in the study of neuro-divergence and neurodegenerative disease.

However, outside our fields, and sometimes outside our reach, are efforts that employ variable incentives to a much greater effect creating massive LRs. LibriVox offer contributors the chance to create audio recordings of classic works of literature, develop their skills as reader and voice actors, work within a community of similarly minded volunteers and enable access to the blind, illiterate and others for whom existing versions were inaccessible. On the other hand, researchers cannot always rely on contributions from social media providers whose products are not always well matched to our research questions or who may be unable or unwilling to share their holdings in the ways that our research programs need.

Given the perpetual need for larger and more diverse LRs, the success of novel incentives in other fields that collect data from human contributors and the early successes and growth of interest among LR creators, this workshop will continue the discussion from the 2016 LREC Workshop on Novel Incentives in Data Collection and the 2020 LREC Workshop on Citizen Linguistics and Language Resource Development.

Organizers

Chris Callison-Burch, University of Pennsylvania (USA)
Christopher Cieri, Linguistic Data Consortium, University of Pennsylvania (USA)
James Fiumara, Linguistic Data Consortium, University of Pennsylvania (USA)
Mark Liberman, Linguistic Data Consortium, University of Pennsylvania (USA)

Program Committee:

Nicoletta Calzolari, CNR-ILC (Italy)
Jon Chamberlain, University of Essex (United Kingdom)
Yiya Chen, Leiden University (Netherlands)
Christopher Cieri, Linguistic Data Consortium, University of Pennsylvania (USA)
Maxine Eskenazi, Carnegie Mellon University (USA)
James Fiumara, Linguistic Data Consortium, University of Pennsylvania (USA)
Karën Fort, Sorbonne Université (France)
Bruno Guillaume, INRIA (France)
Matthew Lease, University of Texas at Austin (USA)
Mark Liberman, Linguistic Data Consortium, University of Pennsylvania (USA)
Massimo Poesio, Queen Mary University of London (United Kingdom)
Odette Scharenborg, Delft University of Technology (Netherlands)
Jennifer Tracey, Linguistic Data Consortium, University of Pennsylvania (USA)
Jonathan Wright, Linguistic Data Consortium, University of Pennsylvania (USA)
Jiahong Yuan, Baidu (China)

Table of Contents

Workshop Program

9:00–9:20 *The NIEUW Project: Developing Language Resources through Novel Incentives*
James Fiumara, Christopher Cieri, Mark Liberman, Chris Callison-Burch, Jonathan Wright and Robert Parker

9:20–9:40 *Use of a Citizen Science Platform for the Creation of a Language Resource to Study Bias in Language Models for French: A Case Study*
Karën Fort, Aurélie Névéol, Yoann Dupont and Julien Bezançon

9:40–10:00 *Fearless Steps APOLLO: Advanced Naturalistic Corpora Development*
John H.L. Hansen, Aditya Joglekar, Szu-Jui Chen, Meena Chandra Shekar and Chelzy Belitz

10:00–10:20 *Creating Mexican Spanish Language Resources through the Social Service Program*
Carlos Daniel Hernandez Mena and Ivan Vladimir Meza Ruiz

11:00–11:20 *Fictionary-Based Games for Language Resource Creation*
Steinunn Rut Friðriksdóttir and Hafsteinn Einarsson

11:20–11:40 *Using Mixed Incentives to Document Xi'an Guanzhong*
Juhong Zhan, Yue Jiang, Christopher Cieri, Mark Liberman, Jiahong Yuan, Yiya Chen and Odette Scharenborg

11:40–12:00 *Crowdsourced Participants' Accuracy at Identifying the Social Class of Speakers from South East England*
Amanda Cole

12:00–12:20 *About the Applicability of Combining Implicit Crowdsourcing and Language Learning for the Collection of NLP Datasets*
Verena Lyding, Lionel Nicolas and Alexander König

12:20–12:40 *The Influence of Intrinsic and Extrinsic Motivation on the Creation of Language Resources in a Citizen Linguistics Project about Lexicography*
Barbara Heinisch