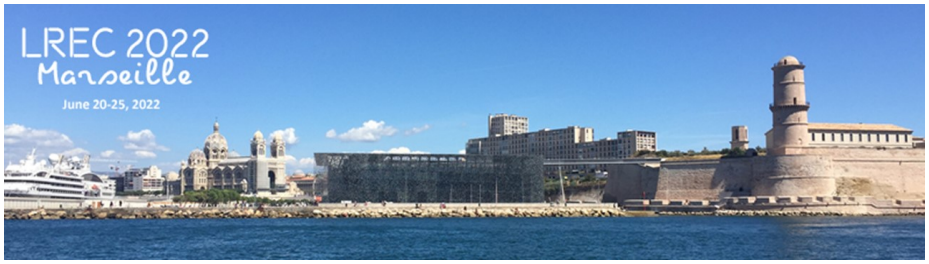

LREC 2022 Workshop
Language Resources and Evaluation Conference
20-25 June 2022

Workshop on Processing Language Variation: Digital Armenian (DigitAm)

Editors:
Victoria Khurshudyan, Nadi Tomeh, Damien Nouvel, Anaid Donabedian, Chahan Vidal-Gorene

Proceedings of the LREC 2022 workshop on Processing Language Variation: Digital Armenian (DigitAm)

Edited by:
Victoria Khurshudyan, Nadi Tomeh, Damien Nouvel, Anaid Donabedian, Chahan Vidal-Gorene

ISBN: 978-2-493814-04-3
EAN: 9782493814043

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

Preface

This volume includes the proceedings of the workshop Processing Language Variation: Digital Armenian held in Marseille, France, June 20, 2022. It is organized by the team of DALiH project: Digitizing Armenian Linguistic Heritage (DALiH)[1]: Armenian Multi-variational Corpus and Data Processing, more particularly by the three research centres: Structure et Dynamique des Langues (SeDyL)/INALCO, Laboratoire d'Informatique de Paris-Nord (LIPN) /Université Sorbonne Paris Nord and Équipe de Recherche Textes, Informatique, Multilinguisme (ERTIM)/INALCO. The workshop is in line with the international conference Digital Armenian first held in Paris, INALCO, in 2019.

The workshop welcomed papers on exploring the problems connected with language variation processing through interoperability of NLP and linguistic resources and tools in particular (but not limited to) for multi-variational under-resourced languages, multi-variational corpora designing and functionality, the evaluation of language scalar variation and the degree of interoperability relevance, language variety identification and distance measuring etc.

A significant gap exists for the availability of NLP resources for different languages with a few languages having quasi-complete NLP coverage and many others being under-resourced (or no-resourced at all). Besides, the under-resourced languages can often have variation either at synchronic (dialects, oral vernacular varieties) or diachronic level (ancient variants of a target language) for which resources can be completely absent especially if no written tradition exists for a target variety. The workshop will focus on processing and reutilisation of NLP resources for under-resourced languages with variation in general, with a particular attention to the Armenian language data.

Current state-of-the-art NLP approaches open up remarkable perspectives not only to exploit the available NLP resources of the well-resourced languages for the under-resourced ones, but also to recycle the existing resources of a target language for its varieties (multi-variational resources) instead of processing target language/variety-based new NLP resources from scratch.

The existing resources are often heterogeneous in terms of accessibility, formatting, linguistic background and they are usually specialized in only one type of a tool/resource (scanned text and/or plain-text databases, dictionaries, annotation models/tools, annotated corpora and datasets etc.). Therefore, one of the important issues is to work out approaches and standards of harmonization and interoperability of the existing data and resources.

Overall, six papers were selected for the workshop. Two papers focus on different aspects of Classical and Middle Armenian linguistic data processing (Analyse Automatique de l'Ancien Arménien. Évaluation d'une méthode hybride « dictionnaire » et « réseau de neurones » sur un Extrait de l'Adversus Haereses d'Irénée de Lyon by Kepeklian and Kindt; and Describing Language Variation in the Colophons of Armenian Manuscripts by Van Elverdinghe and Kindt) and one paper explores the variational identification for Classical Armenian and two modern standards (Dialects Identification of Armenian Language by Avetisyan). Modern Armenian standards are targeted in the paper presenting a morphological transducer for Modern Western Armenian (A Free/Open-Source Morphological Transducer for Western Armenian by Dolatian et al.), and another on Eastern Armenian National Corpus (Eastern Armenian National Corpus: State of the Art and Perspectives by Khurshudyan et al.), Finally, one paper explores the possibilities of Automatic Speech Recognition model (ASR) model processing for modern Armenian varieties (Towards a Unified ASR System for the Armenian Standards by Chakmakjian and Wang).

Workshop Organizers

[1]The project DALiH is funded by French National Research Agency ANR-21-CE38-0006.

Organizers

Victoria Khurshudyan – Inalco, Sedyl, CNRS
Anaid Donabedian – Inalco, Sedyl, CNRS
Chahan Vidal-Gorene – École Nationale des Chartes-PSL
Nadi Tomeh – LIPN, Université Sorbonne Paris Nord
Damien Nouvel – INALCO, ERTIM

Program Committee:

Victoria Khurshudyan, Inalco, Sedyl, CNRS, IRD
Anaid Donabedian, Inalco, Sedyl, CNRS, IRD
Chahan Vidal-Gorene, École Nationale des Chartes-PSL
Nadi Tomeh, LIPN, Université Sorbonne Paris Nord
Damien Nouvel, Inalco, ERTIM
Emmanuel Cartier, LIPN, Université Sorbonne Paris Nord
Thierry Charnois, LIPN, Université Sorbonne Paris Nord
Ilaine Wang, Inalco, ERTIM
Vladimir Plungian, Vinogradov Russian Language Institute, Russian Academy of Sciences
Timofey Arkhangelskiy, University of Hamburg

Table of Contents

Conference Program

2:00pm–2:20pm Opening with a presentation on the project of Digitizing Armenian Linguistic Heritage (DALiH): Armenian Multivariational Corpus and Data Processing
Victoria Khurshudyan

Session 1

2:20pm–2:40pm *A Free/Open-Source Morphological Transducer for Western Armenian*
Hossep Dolatian, Daniel Swanson and Jonathan Washington

2:40pm–3:00pm *Dialects Identification of Armenian Language*
Karen Avetisyan

3:00pm–3:20pm *Analyse Automatique de l'Ancien Arménien. Évaluation d'une méthode hybride « dictionnaire » et « réseau de neurones » sur un Extrait de l'Adversus Haereses d'Irénée de Lyon*
Bastien Kindt and Gabriel Kepeklian

3:20pm–3:40pm *Describing Language Variation in the Colophons of Armenian Manuscripts*
Bastien Kindt and Emmanuel Van Elverdinghe

3:40pm–4:00pm *Eastern Armenian National Corpus: State of the Art and Perspectives*
Victoria Khurshudyan, Timofey Arkhangelskiy, Misha Daniel, Vladimir Plungian, Dmitri Levonian, Alex Polyakov and Sergei Rubakov

Session 2

4:30pm–4:50pm *Towards a Unified ASR System for the Armenian Standards*
Samuel Chakmakjian and Ilaine Wang

4:50pm–5:50pm ***Round table***

5:50pm–6:00pm Closing remarks
Viktoria Khurshudyan