



LREC 2026

**Proceedings of Workshop on Dialects in NLP — A
Resource Perspective (DialRes) @ LREC 2026**

Workshop Proceedings

Editors

**Antonios Anasastasopoulos, Stella Markantonatou,
Angela Ralli, Marcos Zampieri, Stavros Bompolas,
Vivian Stamou**

16 May 2026

©ELRA Language Resources Association (ELRA), 2026
These proceedings are licensed under a Creative Commons Attribution-
NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-60-9

Preface

The proceedings contains the 34 papers accepted for presentation at the Workshop on Dialects in NLP — A Resource Perspective (DialRes-LREC2026, <https://dialres.github.io/dialres/index.html>), co-located with the Fifteenth Biennial Language Resources and Evaluation Conference (LREC). The conference was held at the Palau de Congressos de Palma in Palma de Mallorca, Spain, from 11 to 16 May 2026, with the DialRes workshop taking place on 16 May 2026.

For this first edition of DialRes, we received 43 submissions. We are deeply grateful to the dialect studies and NLP communities for embracing DialRes as a venue for presenting their work. This strong response highlights the growing interest in dialectal research and confirms the need for a dedicated forum focusing on the development and use of dialectal resources.

The accepted papers cover both living and historical dialects from nearly all continents. Contributions on historical dialects include studies on English, German, Heptanesian Greek, Old Irish, and Transcarpathian varieties. Research on living dialects spans a wide range of linguistic contexts, including Arabic, Aromanian, Bangla, Basque, Formosan languages, German, Italian, Kurdish, Modern Greek, and Slavic varieties (such as Pomak and Ukrainian), as well as Wancho.

The contributions address a broad spectrum of topics, ranging from the collection and development of oral and written dialectal resources—such as corpora, treebanks, benchmarks, and specialized databases—to their application in dialectometry and dialect classification. Many papers employ state-of-the-art methodologies, including Automatic Speech Recognition (ASR), neural parsing techniques, and Large Language Models (LLMs).

We would also like to express our gratitude to the Archimedes Research Unit of the Athena Research Center and the UniDive COST Action for endorsing the workshop. In particular, the Archimedes Research Unit generously supported the participation of three of the organizers.

Last but not least, we extend our thanks to the members of the DialRes Program Committee for their contributions to the successful organization of the workshop.

The DialRes Workshop Organizers:

Antonis Anastasopoulos
Stella Markantonatou
Angela Ralli
Marcos Zampieri
Stavros Bompolas
Vivian Stamou

Organizing Committee

- Antonios Anasastasopoulos
- Stella Markantonatou
- Angela Ralli
- Marcos Zampieri
- Stavros Bompolas
- Vivian Stamou

Program Committee

- Mohamed Aghzal (George Mason University, USA)
- Syeda Sabrina Akter (George Mason University, USA)
- Esra Ayten (Istanbul Medeniyet University, Türkiye)
- Verginica Barbu-Mititelu (Romanian Academy – Research Institute for Artificial Intelligence, Romania)
- Johnatan Bonilla (Humboldt University of Berlin, Germany)
- Cristina Bosco (University of Turin, Italy)
- Stergios Chatzikyriakidis (University of Crete, Greece)
- Antonis Dimakis (National and Kapodistrian University of Athens & Archimedes Center, Greece)
- Fahim Faisal (George Mason University, USA)
- Voula Giouli (Aristotle University of Thessaloniki, Greece)
- Mirjana Ilić (University of Niš, Serbia)
- Brian Joseph (Ohio State University, USA)
- Petros Karatsareas (University of Westminster, UK)
- Panagiotis Krimpas (Democritus University of Thrace, Greece)
- J. Elizabeth Liebl (George Mason University, USA)
- Johann-Mattis List (University of Passau, Germany)
- Nikola Ljubešić (Jožef Stefan Institute, Slovenia)
- Irina Lobzhanidze (Ilia State University, Georgia)
- Chutong Meng (George Mason University, USA)
- Nikolett Mus (Hungarian Research Centre for Linguistics, Hungary)
- Petya Osenova (Sofia University "St. Kliment Ohridski", Bulgaria)

- Joshua Otten (George Mason University, USA)
- Georgios Paraskevopoulos (ILSP/Athena Research Center, Greece)
- Siyao (Logan) Peng (LMU Munich, Germany)
- Alistair Plum (University of Luxembourg, Luxembourg)
- Prokopis Prokopidis (ILSP/Athena Research Center, Greece)
- Christoph Purschke (University of Luxembourg, Luxembourg)
- Josep Quer (ICREA – Universitat Pompeu Fabra, Spain)
- Yves Scherrer (University of Oslo, Norway)
- Krešimir Sojat (University of Zagreb, Croatia)
- Aarohi Srivastava (University of Notre Dame, USA)
- Robert Sviben (Institute for the Croatian Language, Croatia)
- Chihiro Taguchi (University of Notre Dame, USA)
- Zeerak Talat (Simon Fraser University, Canada)
- Alexandros Tantos (Aristotle University of Thessaloniki, Greece)
- Belu Ticona (George Mason University, USA)
- Chara Tsoukala (ILSP/Athena Research Center, Greece)
- Socrates Vakirtzian (University of Crete, Greece)
- Alexandra Vella (University of Malta, Malta)
- Shuly Wintner (University of Haifa, Israel)
- Murong Yue (George Mason University, USA)

Table of Contents

<i>A Bolu: A Structured Dataset for the Computational Analysis of Sardinian Improvisational Poetry</i> Silvio Calderaro and Johanna Monti	1
<i>Saar-Voice: A Multi-Speaker Saarbrücken Dialect Speech Corpus</i> Lena Sophie Oberkircher, Jesujoba Alabi, Dietrich Klakow and Jürgen Trouvain	12
<i>MD_NLP: Reconstructing an Australian English Heritage Dialect Corpus from the Mitchell-Delbridge Recordings through LLM-Assisted Speaker Attribution</i> Steven Coats	24
<i>Challenges in the Detection of Dialect for Historical Languages; the Case of Old Irish Text Resources</i> Adrian Doyle	33
<i>Phonologically-aware Automatic Speech Recognition Evaluation of Low-Resource Languages: The Case of Basque Dialects</i> Christoforos Souganidis, Asier Herranz, Ibon Saratxaga, Eva Navas and Inma Hernaez	48
<i>Systematic Normalization of Spoken Mixed-Language, Mixed-Dialect Data</i> Margaret Blevins	58
<i>Handling Cross-Dialect Syntactic Variation: a Theory-Driven Web Resource</i> Emanuela Li Destri, Marco Longhin, Gaia Sorge, Sofia Ferroni, Giovanni Battista Matteazzi, Andrea Artioli, Lorenzo Carletti, Federico Motta, Giuseppe Longobardi and Cristina Guardiano	70
<i>Can LLM Agents Identify Spoken Dialects like a Linguist?</i> Tobias Bystrich, Lukas Hamm, Maria Hassan Akhter, Lea Fischbach, Lucie Flek and Akbar Karimi	83
<i>Beyond Accuracy: Analyzing Dialect Confusion in Automatic Speech-Based Dialect Classification</i> Lea Fischbach, Alfred Lameli and Lucie Flek	93
<i>FLEURS-Kobani: Extending FLEURS dataset for Northern Kurdish</i> Daban Q. Jaff and Mohammad Mohammadamini	104
<i>Exploring the reusability of Northern Kurdish resources for Badini speech recognition</i> Mohammad Mohammadamini, Aveen Jalal Mohammed, Barzan Hussein Mohammed, Dezh- heen H. Abdulazeez, Imad Saeed Sadeeq, Dilgash Mohammed Salih, Amara Ismail Melhum and Abuobaida Abdullah Dheyab	110
<i>Wancho Dialectometry: Community-created data and the Living Dictionaries project</i> Kellen Parker van Dam	116
<i>Dialectometry and Evaluation of the ePark Corpus for Low-Resource Formosan Language Di- alects</i> Henry Gagnier	124

<i>A Dialectal Corpus for Ukrainian: Collection, Classification, and Standardization</i> Yuliia Frund and Sina Ahmadi	135
<i>German Dialects Across Situations, Generations, and Regions: The REDE corpus as an Oral Resource for NLP</i> Hanna Fischer and Alfred Lameli	144
<i>A Catalog of Basque Dialectal Resources: Online Collections and Standard-to-Dialectal Adaptations</i> Jaione Bengoetxea, Itziar Gonzalez-Dios and Rodrigo Agerri	153
<i>WoVis: Interactive Visualization of Word Embeddings for Semantic Change in Historical and Dialectal Language Resources</i> Filip Miletić, Maximilian Henkel, Rene Cutura, Sophie Sadler, Quynh Quang Ngo, Michael Sedlmair and Sabine Schulte im Walde	165
<i>Speaker Normalization via Voice Conversion Reveals a Human-Machine Dissociation in Dialect Classification</i> Caroline Kleen, Lea Fischbach, Akbar Karimi, Lucie Flek and Alfred Lameli	177
<i>South Tyrolean Dialect-to-Standard Speech Translation: A Resource</i> Greta H. Franzini and Luca Ducceschi	188
<i>TransVar – the Corpus for Variation and Change Study of the Historical Transcarpathian lects</i> Iliia Afanasev	195
<i>The Generator-Eraser Paradox: Community Guidelines for Responsible LLM-Assisted Dialect Resource Creation</i> Wajdi Zaghouani	209
<i>The Texas German Dialect Project Corpus as a Diachronic Resource for Investigating Language Contact</i> Thomas Schmidt, Margaret M. Blevins, Hans C. Boas and Glenn Gilbert	221
<i>Pontic Greek in the Caucasus: an online corpus</i> Svetlana Berikashvili and Stavros Skopeteas	230
<i>Meaning Over Morphology: A Multi-Metric Benchmark of LLMs for Bangla Dialect Translation</i> Soumik Deb Niloy, Subhey Sadi Rahman, Mahbub E Sobhani, Md. Golam Rabiul Alam, Farig Yousuf Sadeque and Md. Rezuwan Hassan	238
<i>Sociolinguistic aspects of crowdsourcing for a vocal corpus of Alsatian</i> Pascale Erhart, Lucile Hamm, Sam Bigeard, Carole Werner, Malek Yaich and Slim Ouni	256
<i>HeptaTAX: A Neuro-Symbolic Pipeline and Benchmark for Classifying 16th-Century Heptanesian Notarial Acts</i> Stergios Chatzikyriakidis, Eleni Karantzola and Vasiliki Makri	265
<i>Towards Semantic Access and Interoperability in Digital Dialectal Atlases. A Case Study</i> Paola Marongiu and Simonetta Montemagni	274

A CLDF-Compliant Lexical Database for Modern Greek Dialects: Resource Design and Dialectometric Analysis

Stavros Bompolas, Natalia Chousou-Polydouri, Manuela Genitsaridi, Danae Karatzanou, Georgios Kostopoulos, Elena Anagnostopoulou and Dimitra Melissaropoulou 287

A Speech Resource for the Pontic Greek Dialect: Transcription Choices and Baseline ASR Evaluation

Rodanna Konstantinidou, Chara Tsoukala, Vivian Stamou, Voula Giouli and Stella Markantonatou 300

First Steps in ASR for Cypriot Greek: Challenges and Insights

Vivian Stamou, Spyros Armostis, Antigoni Klimi, Georgios Paraskevopoulos, Vassilis Katsouros and Antonios Anastasopoulos 308

Structural Divergence under Shared Language-Level Specification: Griko in Universal Dependencies

Stavros Bompolas, Emanuela Pinna, Josep Quer, Marika Lekakou and Stella Markantonatou 315

Digital Preservation of Aromanian Through Knowledge Management and Automatic Speech Recognition Evaluation

Marija Pendevska and Hristina Nastevska 327

A Novel Typology of Mutually Intelligible Words: The Case of Slavic Languages

Edward Klyshinsky and Yulia Badryzlova 337

Transfer Learning for an Endangered Slavic Variety: Dependency Parsing in Pomak Across Contact-Shaped Dialects

Sercan Karakas 345

Workshop Program

Saturday, May 16, 2026

09:00–10:30	Session A Room: Room 4 Chair: Stella Markantonatou
09:10–09:20	<i>A Bolu: A Structured Dataset for the Computational Analysis of Sardinian Improvisational Poetry</i> Silvio Calderaro and Johanna Monti
09:20–09:30	<i>Saar-Voice: A Multi-Speaker Saarbrücken Dialect Speech Corpus</i> Lena Sophie Oberkircher, Jesujoba Alabi, Dietrich Klakow and Jürgen Trouvain
09:30–09:40	<i>MD_NLP: Reconstructing an Australian English Heritage Dialect Corpus from the Mitchell-Delbridge Recordings through LLM-Assisted Speaker Attribution</i> Steven Coats
09:40–09:50	<i>Challenges in the Detection of Dialect for Historical Languages; the Case of Old Irish Text Resources</i> Adrian Doyle
10:00–10:30	<i>2-minute poster presentations</i>
10:30–11:00	Coffee break
10:30–11:00	Poster Session Chairs: Stavros Bompolas, Vivian Stamou

Saturday, May 16, 2026 (continued)

11:00–13:00	Session B Room: Room 4 Chair: Antonios Anastasopoulos
11:00–11:10	<i>Phonologically-aware Automatic Speech Recognition Evaluation of Low-Resource Languages: The Case of Basque Dialects</i> Christoforos Souganidis, Asier Herranz, Ibon Saratxaga, Eva Navas and Inma Hernaez
11:10–11:20	<i>Systematic Normalization of Spoken Mixed-Language, Mixed-Dialect Data</i> Margaret Blevins
11:20–11:30	<i>Handling Cross-Dialect Syntactic Variation: a Theory-Driven Web Resource</i> Emanuela Li Destri, Marco Longhin, Gaia Sorge, Sofia Ferroni, Giovanni Battista Matteazzi, Andrea Artioli, Lorenzo Carletti, Federico Motta, Giuseppe Longobardi and Cristina Guardiano
11:30–11:40	<i>Can LLM Agents Identify Spoken Dialects like a Linguist?</i> Tobias Bystrich, Lukas Hamm, Maria Hassan Akhter, Lea Fischbach, Lucie Flek and Akbar Karimi
11:45–12:30	<i>Invited talk (title to be announced)</i> Prof. Barbara Plank, LMU Munich, Visiting Prof ITU Copenhagen
12:30–13:00	Community discussion
10:30–11:00	List of Posters
10:30–11:00	<i>Beyond Accuracy: Analyzing Dialect Confusion in Automatic Speech-Based Dialect Classification</i> Lea Fischbach, Alfred Lameli and Lucie Flek
10:30–11:00	<i>FLEURS-Kobani: Extending FLEURS dataset for Northern Kurdish</i> Daban Q. Jaff and Mohammad Mohammadamini
10:30–11:00	<i>Exploring the reusability of Northern Kurdish resources for Badini speech recognition</i> Mohammad Mohammadamini, Aveen Jalal Mohammed, Barzan Hussein Mohammed, Dezheen H. Abdulazeez, Imad Saeed Sadeeq, Dilgash Mohammed Salih, Amara Ismail Melhum and Abuobaida Abdullah Dheyab

Saturday, May 16, 2026 (continued)

- 10:30–11:00 *Wancho Dialectometry: Community-created data and the Living Dictionaries project*
Kellen Parker van Dam
- 10:30–11:00 *Dialectometry and Evaluation of the ePark Corpus for Low-Resource Formosan Language Dialects*
Henry Gagnier
- 10:30–11:00 *A Dialectal Corpus for Ukrainian: Collection, Classification, and Standardization*
Yuliia Frund and Sina Ahmadi
- 10:30–11:00 *German Dialects Across Situations, Generations, and Regions: The REDE corpus as an Oral Resource for NLP*
Hanna Fischer and Alfred Lameli
- 10:30–11:00 *A Catalog of Basque Dialectal Resources: Online Collections and Standard-to-Dialectal Adaptations*
Jaione Bengoetxea, Itziar Gonzalez-Dios and Rodrigo Agerri
- 10:30–11:00 *WoVis: Interactive Visualization of Word Embeddings for Semantic Change in Historical and Dialectal Language Resources*
Filip Miletić, Maximilian Henkel, Rene Cutura, Sophie Sadler, Quynh Quang Ngo, Michael Sedlmair and Sabine Schulte im Walde
- 10:30–11:00 *Speaker Normalization via Voice Conversion Reveals a Human-Machine Dissociation in Dialect Classification*
Caroline Kleen, Lea Fischbach, Akbar Karimi, Lucie Flek and Alfred Lameli
- 10:30–11:00 *South Tyrolean Dialect-to-Standard Speech Translation: A Resource*
Greta H. Franzini and Luca Ducceschi
- 10:30–11:00 *TransVar – the Corpus for Variation and Change Study of the Historical Transcarpathian lects*
Ilia Afanasev
- 10:30–11:00 *The Generator-Eraser Paradox: Community Guidelines for Responsible LLM-Assisted Dialect Resource Creation*
Wajdi Zaghouni
- 10:30–11:00 *The Texas German Dialect Project Corpus as a Diachronic Resource for Investigating Language Contact*
Thomas Schmidt, Margaret M. Blevins, Hans C. Boas and Glenn Gilbert
- 10:30–11:00 *Pontic Greek in the Caucasus: an online corpus*
Svetlana Berikashvili and Stavros Skopeteas

Saturday, May 16, 2026 (continued)

- 10:30–11:00 *Meaning Over Morphology: A Multi-Metric Benchmark of LLMs for Bangla Dialect Translation*
Soumik Deb Niloy, Subhey Sadi Rahman, Mahbub E Sobhani, Md. Golam Rabiul Alam, Farig Yousuf Sadeque and Md. Rezuwan Hassan
- 10:30–11:00 *Sociolinguistic aspects of crowdsourcing for a vocal corpus of Alsatian*
Pascale Erhart, Lucile Hamm, Sam Bigeard, Carole Werner, Malek Yaich and Slim Ouni
- 10:30–11:00 *HeptaTAX: A Neuro-Symbolic Pipeline and Benchmark for Classifying 16th-Century Heptanesian Notarial Acts*
Stergios Chatzikyriakidis, Eleni Karantzola and Vasiliki Makri
- 10:30–11:00 *Towards Semantic Access and Interoperability in Digital Dialectal Atlases. A Case Study*
Paola Marongiu and Simonetta Montemagni
- 10:30–11:00 *A CLDF-Compliant Lexical Database for Modern Greek Dialects: Resource Design and Dialectometric Analysis*
Stavros Bompolas, Natalia Chousou-Polydouri, Manuela Genitsaridi, Danae Karatzanou, Georgios Kostopoulos, Elena Anagnostopoulou and Dimitra Melissaropoulou
- 10:30–11:00 *A Speech Resource for the Pontic Greek Dialect: Transcription Choices and Baseline ASR Evaluation*
Rodanna Konstantinidou, Chara Tsoukala, Vivian Stamou, Voula Giouli and Stella Markantonatou
- 10:30–11:00 *First Steps in ASR for Cypriot Greek: Challenges and Insights*
Vivian Stamou, Spyros Armostis, Antigoni Klimi, Georgios Paraskevopoulos, Vassilis Katsouros and Antonios Anastasopoulos
- 10:30–11:00 *Structural Divergence under Shared Language-Level Specification: Griko in Universal Dependencies*
Stavros Bompolas, Emanuela Pinna, Josep Quer, Marika Lekakou and Stella Markantonatou
- 10:30–11:00 *Digital Preservation of Aromanian Through Knowledge Management and Automatic Speech Recognition Evaluation*
Marija Pendevska and Hristina Nastevska
- 10:30–11:00 *A Novel Typology of Mutually Intelligible Words: The Case of Slavic Languages*
Edward Klyshinsky and Yulia Badryzlova
- 10:30–11:00 *Transfer Learning for an Endangered Slavic Variety: Dependency Parsing in Pomak Across Contact-Shaped Dialects*
Sercan Karakas