LREC 2026

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

Workshop Proceedings

Editors
Antonios Anastasopoulos, Stella Markantonatou, Angela Ralli, Marcos Zampieri, Stavros Bompolas, Vivian Stamou

16 May 2026

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

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 Anastasopoulos
- 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

Workshop Program

Saturday, May 16, 2026

09:00–10:30 **Session A**
Room: Room 4
Chair: Stella Markantonatou

09:10–09:20 *A Bolu: A Structured Dataset for the Computational Analysis of Sardinian Improvisational Poetry*
Silvio Calderaro and Johanna Monti

09:20–09:30 *Saar-Voice: A Multi-Speaker Saarbrücken Dialect Speech Corpus*
Lena Sophie Oberkircher, Jesujoba Alabi, Dietrich Klakow and Jürgen Trouvain

09:30–09:40 *MD_NLP: Reconstructing an Australian English Heritage Dialect Corpus from the Mitchell-Delbridge Recordings through LLM-Assisted Speaker Attribution*
Steven Coats

09:40–09:50 *Challenges in the Detection of Dialect for Historical Languages; the Case of Old Irish Text Resources*
Adrian Doyle

10:00–10:30 ***2-minute poster presentations***

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 *Phonologically-aware Automatic Speech Recognition Evaluation of Low-Resource Languages: The Case of Basque Dialects*
Christoforos Souganidis, Asier Herranz, Ibon Saratxaga, Eva Navas and Inma Hernaez

11:10–11:20 *Systematic Normalization of Spoken Mixed-Language, Mixed-Dialect Data*
Margaret Blevins

11:20–11:30 *Handling Cross-Dialect Syntactic Variation: a Theory-Driven Web Resource*
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 *Can LLM Agents Identify Spoken Dialects like a Linguist?*
Tobias Bystrich, Lukas Hamm, Maria Hassan Akhter, Lea Fischbach, Lucie Flek and Akbar Karimi

11:45–12:30 *Invited talk (title to be announced)*
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 *Beyond Accuracy: Analyzing Dialect Confusion in Automatic Speech-Based Dialect Classification*
Lea Fischbach, Alfred Lameli and Lucie Flek

10:30–11:00 *FLEURS-Kobani: Extending FLEURS dataset for Northern Kurdish*
Daban Q. Jaff and Mohammad Mohammadamini

10:30–11:00 *Exploring the reusability of Northern Kurdish resources for Badini speech recognition*
Mohammad Mohammadamini, Aveen Jalal Mohammed, Barzan Hussein Mohammed, Dezheen H. Abdulazeez, Imad Saeed Sadeeq, Dilgash Mohammed Salih, Amera 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 Zaghouani

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

A Bolu: A Structured Dataset for the Computational Analysis of Sardinian Improvisational Poetry

Silvio Calderaro[1,2], Johanna Monti[2]
[1]Università di Pisa, Italia
[2]Università di Napoli L'Orientale, Italia
silvio.calderaro@phd.unipi.it, jmonti@unior.it

Abstract

The growing interest of Natural Language Processing (NLP) in minority languages has not yet bridged the gap in the preservation of oral linguistic heritage. In particular, extemporaneous poetry — a performative genre based on real-time improvisation, metrical-rhetorical competence — remains a largely unexplored area of computational linguistics. This methodological gap necessitates the creation of specific resources to document and analyse the structures of improvised poetry. This is the context in which A Bolu was created, the first structured corpus of extemporaneous poetry dedicated to *cantada logudorese*, a variant of the Sardinian language. The dataset comprises 2,835 stanzas for a total of 141,321 tokens. The study presents the architecture of the corpus and applies a multidimensional analysis combining descriptive statistical indices and computational linguistics techniques to map the characteristics of the poetic text. The results indicate that the production of Sardinian extemporaneous poets is characterised by recurring patterns that support Parry and Lord's theory of formulaicity. This evidence not only provides a new key to understanding oral creativity, but also offers a significant contribution to the development of NLP tools that are more inclusive and sensitive to the specificities of less widely spoken languages.

Keywords: Corpus of minority languages, Extemporaneous poetry, Oral-formulaic patterns

1. Introduction and Background

The growing interest of the Natural Language Processing (NLP) community in low-resource languages, minority varieties, and dialects reflects a broader shift toward linguistic inclusivity. For decades, computational tools and annotated resources were concentrated on a small set of high-resource languages, leaving the vast majority of the world's linguistic heritage outside the reach of modern NLP techniques. Corpus construction efforts for minority languages must typically contend with sparse digital presence, orthographic instability, and the absence of standardized annotation frameworks, challenges well attested in comparable projects for languages such as Guarani (Chiruzzo et al., 2022), Breton (Grobol and Jouitteau, 2024), and the minority languages of Italy (Ramponi, 2024). These difficulties are further compounded when the target variety is rooted in oral and performative traditions (Zumthor, 1997; Ong, 1982) that have historically resisted stabilization into annotated digital formats.

Within this landscape, Sardinian occupies a particularly complex position. Officially recognized as a minority language under Italian national law and widely regarded as the closest living descendant of Latin (Virdis, 1988). Existing resources include SardNet (Angioni et al., 2018), a lexical resource mapping Sardinian word senses onto WordNet entries, a BERT-based Part-of-Speech tagger (Carta et al., 2025), and a nascent Automatic Speech Recognition system (Chizzoni and Vietti, 2024).

The *cantada logudoresa* is a formal poetic contest in which *cantadores* improvise verses in the Logudorese dialect under strict metrical and thematic constraints, structured around the *Otada*, an octave of eight hendecasyllabic lines following an ABABABCC or ABBABACC rhyme scheme, alongside shorter forms such as the *batorina* and the closing *dispedida*. Performers must simultaneously manage metrical and rhyming complexity, thematic coherence, and real-time argumentative exchange, all under the pressure of live performance. The preservation of this tradition has historically depended on inconsistent and scattered acts of transcription, precluding any systematic computational investigation. This paper introduces **A Bolu** (Calderaro and Monti, 2026), the first structured digital corpus of the *cantada logudoresa*, designed to fill this gap and establish a reproducible framework for its quantitative and linguistic analysis. Our contributions:

1. **Digital Preservation and Resource Creation:** We provide a high-fidelity digital repository for a vulnerable minority language tradition, preventing the loss of undocumented or fragmented transcriptions and establishing a foundation for future NLP tasks in Sardinian.

2. **Multidimensional Data Modeling:** Unlike flat-text corpora, *A Bolu* is structured to include rich metadata—such as thematic assignments, performer identifiers, and precise execution

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 1–11
16 May 2026.

timestamps per stanza—modeled in a hierarchical format to facilitate complex relational queries.

3. **Computational Stylistics Analysis:** We demonstrate the utility of the dataset by proposing it as a benchmark for investigating "stylistic signatures" and lexical complexity, enabling quantitative research into how real-time improvisational pressures affect the linguistic and metrical choices of the *cantadores*.

The aim of this resource is to integrate Sardinian extemporaneous poetry into the field of computational linguistics, defining a reproducible framework that promotes new avenues of study and formal analysis of structured oral traditions.

2. Methodology

This section describes the methodological approach adopted for the construction of the corpus, divided into the phases of data acquisition 2.1, archive structuring 2.2, and curation of the raw material 2.3. The central objective was to transform a heritage of oral tradition—fragmentary and discontinuous by nature—into a structured digital resource, suitable for computational processing and quantitative linguistic analysis.

2.1. Data Acquisition and Source Selection

The primary data for this study were collected programmatically from làcanas.it (Redazione Làcanas, 2014), an online newspaper dedicated to Sardinian culture, news, and identity, with an archive of improvised poems in the Sardinian language. The scraping process targeted the poems on the site to maintain the same origin as the source. The corpus was constructed according to strict criteria of linguistic and typological homogeneity to ensure the reliability of subsequent quantitative comparisons. The sources were selected based on three fundamental parameters:

1. **Transcription Quality and Metadata Richness:**produced by the official editorial staff were included in the corpus. This restriction was implemented to ensure consistency, reliability, and philological accuracy across the dataset. A primary selection criterion was the availability of essential contextual metadata, including the performance setting, the designated themes (i.e., thematic debates), and, crucially, the recorded execution time of each stanza for each poet.
2. **Generic Consistency:** To avoid stylistic bias, the dataset exclusively comprises performances belonging to the same poetic genre, specifically*(cantada logudoresa)*. This ensures that the metrical constraints and the thematic development remain constant across the entire sample.
3. **Linguistic Variety:** The selection was restricted to a single linguistic variety of the Sardinian language (*Logudorese*). This choice eliminates lexical variation due to dialectal shifts, allowing the analysis to focus strictly on the individual poets' lexical complexity and rhyming strategies.

Despite the application of these selection criteria, the corpus inevitably reflects the intrinsic challenges associated with documenting oral traditions. It should be noted that several performances within the dataset are incomplete: in some cases, individual stanzas are partially or entirely absent from the original transcriptions. These lacunae—often resulting from recording interruptions or archival deterioration—were systematically identified during the data-cleaning phase in order to prevent distortions in the statistical distribution of the linguistic metrics. Furthermore, the texts included in the corpus do not represent complete poetry contests. Rather, the materials primarily consist of the central phase of the competition, which accounts for the majority of the recorded performances. Consequently, these contest-based transcriptions frequently omit either the opening segment (*esordiu)* or the final closing exchange (*dispèdida*) of the debate.

2.2. Data Structure and Corpus Architecture

The collected performances were encoded in a structured, hierarchical format using `JSON` (JavaScript Object Notation). This data model was designed to represent the multi-layered organisation of extemporaneous poetic debates, preserving the internal segmentation of each performance while maintaining explicit links between individual stanzas and their associated metadata. Such a structure ensures both formal consistency and computational accessibility, thereby facilitating corpus querying, annotation, and quantitative analysis.

The corpus is organized into two primary levels:

- **Global Metadata:** This top-level object captures the contextual framework of the entire performance. It includes the *title* of the debate, the *source URL* for traceability, an *introductory summary* of the event, and the *central theme* (thematic dispute) assigned to the poets.
- **Transcription Units (Stanzas):** The `transcription` field contains an ordered array of stanza-level objects, each corresponding to a single metrical unit: namely an octave

(*Otada*), quatrain (*Batorina*), couplet (*Duina*), or closing/free-form stanza (*Dispèdida*). Each object encodes a discrete poetic turn and is associated with a set of structured attributes: a unique numerical `id` (ensuring sequential traceability), the `poet` identifier, the `metrics` label specifying the metrical form, the recorded `time` of execution, and an ordered list of `verse` strings representing the individual lines of the stanza.

An example of the data representation for a single stanza and its associated metadata is provided below:

```
{
  "metadata": {
    "title": "Sozu e Masala in Ballao:
    ↪  tempus e omine",
    "intro_theme": "Cando cantaian
    ↪  manu-manu Peppe Sozu e Marieddu
    ↪  Masala sa resultada de sa gara
    ↪  fut bella e assigurada in
    ↪  partenzia. [...]",
    "core_theme": "Disputa tra
    ↪  l'Eternita inarrestabile e
    ↪  l'Ingegno Creativo",
    "source":
    ↪  "https://www.lacanas.it/..."
  },
  "transcription": [
    {
      "id": 1,
      "poet": "Sozu",
      "metrics": "Otada",
      "time": "1'18",
      "verse": [
        "Ja non cherio in piata sa
        ↪  zente",
        "chi da s'atesa restet in
        ↪  fastizu.",
        "Como sighimos che babbu e che
        ↪  fizu",
        "tantu de allegrare s'ambiente",
        "ja chi su comitadu
        ↪  intelligente",
        "at apagadu su nostru disizu",
        "ch'in parte 'e unu tema mi
        ↪  collocat",
        "e a cantare su tempus mi
        ↪  tocat."
      ]
    }
  ]
}
```

This hierarchical architecture transforms the poetic performance from a static text into a multidimensional data object. By explicitly linking linguistic output to specific constraints—such as the thematic assignment, the metric of the stanzas and the execution time—the dataset allows for a granular investigation. Furthermore, this structured format ensures the corpus is fully interoperable with modern NLP pipelines, establishing a reproducible framework for the quantitative study of Sardinian oral traditions.

2.3. Data Processing and Corpus Curation

Once the raw data were extracted, a rigorous process of curation and normalization was required to transform the scraped material into a reliable research dataset. This phase involved both automated filtering and manual philological verification to address the inconsistencies inherent in a heterogeneous digital archive.

1. **Deduplication Record :** A primary challenge was the presence of duplicate performances. Many poetic debates were found to be published multiple times under slightly different titles or categorized in different sections of the source website. These redundant entries were identified and removed to ensure that the statistical analysis of lexical frequency and poet participation remained unbiased.

2. **Entity Resolution and Normalization:** To ensure that each poet's stylistic signature was correctly attributed, we performed a normalization of personal names. Variations in transcription, such as the inconsistent use of accents (e.g., *Màsala* vs. *Masala*), were reconciled to a single canonical form. This step is crucial for the subsequent calculation of individual lexical complexity and comparative stylometry.

3. **Structural Integrity and Lacunae Flagging:** Each stanza was checked automatically and manually to verify its completeness. Given the oral and often fragmented nature of the transcriptions, we adopted a symbolic tagging system within the *metrical form* metadata field to maintain the chronological sequence of the debate without compromising the linguistic statistics:

   - **Standard labels** (e.g., `otada`): Applied to complete stanzas where all verses are present.
   - **Single asterisk** (e.g., `otada*`): Indicates a partially missing stanza where only a portion of the verses was transcribed.
   - **Double asterisk** (e.g., `otada**`): Indicates a fully missing stanza, preserving its original position in the performance flow while excluding it from the textual analysis.

4. **Temporal Standardization:** The execution time for each stanza, recorded in the source as a string format (e.g., `1'00"`), was parsed and converted into a discrete numerical variable

representing total seconds. In cases where the timing was not present in the original source, or when the stanza was incomplete (as indicated by the asterisk system), a `null` value was assigned to the field.

This approach ensures that the resulting corpus is not merely a collection of texts, but a structured digital resource that preserves both the textual content and the contextual information of the poetic contests, including stanza-level metadata, performance timing, and thematic annotation.

3. Corpus statistics

The resulting dataset, **A Bolu**, to the best of our knowledge constitutes the first structured digital corpus of Sardinian extemporaneous poetry explicitly designed for computational analysis. By aggregating dispersed transcriptions and imposing a systematic structural organization, the corpus provides a solid foundation for rigorous empirical investigation not only into this minority language tradition but also into the structure and dynamics of extemporaneous poetic performance itself.

3.1. Statistical overview

The final corpus, consists of 55 digitalized poetic sessions with a specific focus on the 20th-century oral heritage. Table 1 presents a synoptic view of the corpus dimensions. The dataset comprises 2,835 stanzas distributed across 55 poems, yielding an average of 51.55 stanzas per poem. The corpus also documents its own gaps transparently: 60 stanzas lack execution time annotations, 63 are partially incomplete in their textual transcription, and 8 are entirely missing. Rather than discarding these entries, they have been retained with explicit flags to preserve the structural integrity of each poetic debate.

Feature	Value
Total number of poems (JSON files)	55
Total number of stanzas	2,835
Average stanzas per poem	51.55
Unique identified poets	8
Stanzas with missing execution time	60
Stanzas with partially incomplete	63
Stanzas with totally incomplete	8

Table 1: General statistics of the A Bolu corpus.

3.2. Poet contributions and metrical distribution

The following section examines how stanzas are distributed across poets and metrical forms within the corpus, highlighting both the dominance of certain performers and the structural variety of the poetic debate. Table 2 provides a comprehensive breakdown of stanza counts per poet, disaggregated by metrical type and transcription completeness. Two figures dominate the dataset: Sozu and Masala each appear in 32 poems and contribute 811 and 753 stanzas respectively, together accounting for over 55% of the total corpus. This concentration reflects their historical prominence within the tradition and ensures substantial material for in-depth stylistic analysis. Piras (493 stanzas, 20 poems) Mura (339 stanzas, 15 poems) and Piredda (246 stanzas, 10 poems) constitute a second tier, while the remaining three poets, Sale, Seu, and Budrone, contribute smaller but nonetheless valuable samples, particularly for contrastive purposes. Regarding metrical variety, the *Otada* is overwhelmingly the foundational unit of the corpus, comprising 2,592 of the 2,835 stanzas (93.9%). The remaining forms,*Duina* (95 stanzas, 3.4%), *Batorina* (70 stanzas, 2.5%), and *Despedida* (7 stanzas, 0.2%), appear with considerably lower frequency, consistent with their specialized roles within the structure of the poetic debate. The distribution of these secondary forms is not uniform across the poets. *Masala* employs the widest metrical repertoire, featuring 49 *Duinas*, 35 *Batorinas*, and 3 *Despedidas*. This is followed by *Sozu* (20 *Duinas*, 16 *Batorinas*, and 2 *Despedidas*), *Mura* (16 *Duinas*, 7 *Batorinas*, and 1 *Despedida*), and *Seu* (10 *Duinas*, 12 *Batorinas*, and 1 *Despedida*). In contrast, *Piras*, *Piredda*, *Sale*, and *Budrone* perform exclusively in the *Otada* form. This asymmetry is likely attributable to the uneven distribution of available data across poets in the original source. The specifically tagged entries *Otada** (63 stanzas) and *Otada*** (8 stanzas) denote partially and totally incomplete transcriptions, respectively. These lacunae are distributed unevenly across poets: Sozu accounts for the largest share of partial gaps (21 *Otada**), a figure that is proportional to his extensive presence in the corpus and likely reflects the variable quality of the source recordings. By retaining these entries with explicit markers rather than omitting them, the dataset preserves the sequential structure and turn-taking logic of each session, which is essential for any analysis of debate dynamics and interactional patterns between competing poets.

3.3. Lexical analysis

In the context of oral improvised poetry, the lexical dimension acquires particular significance: poets must generate verse in real time under strict metrical and rhyming constraints, drawing on a mental lexicon that is simultaneously broad enough to avoid repetition and sufficiently controlled to satisfy

Poet	Tot.	Otada	Otada*	Otada**	Duina	Bator.	Disped.	Poems
Sozu	811	751	21	1	20	16	2	32
Masala	753	656	6	4	49	35	3	32
Piras	493	472	18	3	0	0	0	20
Mura	339	305	10	0	16	7	1	15
Piredda	246	242	4	0	0	0	0	10
Sale	88	85	3	0	0	0	0	4
Seu	86	62	1	0	10	12	1	4
Budrone	19	19	0	0	0	0	0	1
Total	**2,835**	**2,592**	**63**	**8**	**95**	**70**	**7**	**-**

Table 2: Comprehensive breakdown of stanzas per poet, including metrical forms and transcription lacunae (*Otada** = partially incomplete; *Otada*** = totally incomplete).

the formal requirements of the tradition. Measuring lexical richness in such a setting can shed light on the cognitive and linguistic resources that differentiate one performer from another. To ensure a methodologically robust evaluation and to mitigate the well-documented sensitivity of the standard Type-Token Ratio (TTR) to corpus length (Baayen, 2001), we employ two length-independent indices: the Moving Average Type-Token Ratio (MATTR) (Covington and McFall, 2010) and the Measure of Textual Lexical Diversity (MTLD) (McCarthy and Jarvis, 2010). For the MATTR calculation, a sliding context window of 50 words was selected, as it approximates the average length of a single *ottava*, the fundamental metrical unit of the performance, thus capturing lexical density at the level of individual poetic turns.

Table 3 reports token counts, type counts, hapax legomena, and the calculated lexical diversity indices for each poet. The full corpus comprises 141,321 tokens and 12,973 types, yielding an overall TTR of 9.18%. This low figure is consistent with the lexical saturation expected in a genre-constrained oral corpus, where recurrent grammatical forms and thematic vocabulary inevitably accumulate across large text volumes. Individual sub-corpora range from 968 tokens (Budrone) to 41,164 tokens (Sozu), a difference of more than fortyfold, which makes direct TTR comparisons across poets methodologically unreliable.

As expected, raw TTR values vary considerably, from 14.47% (Sozu) to 44.01% (Budrone), a divergence that reflects corpus size rather than genuine differences in lexical competence. When examining the MATTR (window = 50), however, scores are remarkably stable across all eight performers, with values ranging from 79.98 (Budrone) to 83.26 (Mura) and a mean of $\mu = 81.59$. This convergence suggests that local lexical density – measured within windows that approximate the length of a single *ottava* – constitutes a structural property of the *Cantada a bolu* tradition, largely independent of the total volume of text produced by each poet.

The MTLD scores provide a complementary perspective on sustained lexical diversity. The global MTLD computed on the entire concatenated corpus is 95.87, with individual values ranging from 81.20 (Budrone) to 110.77 (Mura). These figures indicate a consistent capacity for lexical variation throughout extended stretches of discourse. Mura's score of 110.77 stands out as the highest in the sample, suggesting an exceptional ability to vary vocabulary over longer sequences, even when compared to poets with larger sub-corpora such as Sozu or Masala.

The hapax legomena counts offer a further dimension of analysis. Across the full corpus, 6,789 word forms occur exactly once, accounting for approximately 52.33% of the 12,973 distinct types attested. This substantial proportion of singletons suggests that the poets do not rely exclusively on a fixed repertoire of formulaic expressions but instead introduce novel vocabulary throughout their performances. Taken together, the MATTR and MTLD results indicate that the *A Bolu* corpus exhibits a high and consistent level of lexical richness across sub-corpora of considerably different sizes, lending support to the view that local lexical density is a stable feature of this oral poetic tradition.

Author	Tokens	Types	Hapax	TTR%	MATTR	MTLD
Sozu	41,164	5,956	3,295	14.47	81.89	97.13
Masala	36,305	5,779	3,411	15.92	81.80	98.82
Piras	25,573	4,435	2,712	17.34	80.56	85.65
Mura	16,507	3,302	2,053	20.00	83.26	110.77
Piredda	12,440	2,701	1,674	21.71	80.79	94.85
Sale	4,576	1,344	915	29.37	82.60	100.10
Seu	3,788	1,226	825	32.37	82.40	104.93
Budrone	968	426	301	44.01	79.98	81.20
Total	**141,321**	**12,973**	**6,789**	**9.18**	**81.59**	**95.87**

Table 3: Lexical statistics per author and corpus totals: tokens, types, hapax legomena, TTR, MATTR, and MTLD.

3.4. Temporal Dynamics and Metric Complexity

The inclusion of execution time as a granular attribute introduces an important dimension for analy-

sis in this specific domain. Table 4 summarizes the average duration (in seconds) for the three primary metrical forms identified: the *Otada*, the *Duina*, and the *Batorina*.

The data reveal a clear correlation between metrical length and temporal duration. The *Otada*, being the foundational and most complex unit of the *cantada*, requires a corpus-wide average of 58.18 seconds per unit. In contrast, the *Duina* and the *Batorina* function as more rapid improvisational units,due to the brevity of their short nature, with average execution times of 10.02 and 24.93 seconds, respectively.

An analysis of individual poet performances highlights significant variations in improvisational pace:

- **Masala** emerges as the most rapid improviser across all categories, with an average *Otada* time of 45.57 seconds, significantly below the general mean of 58.18 seconds.
- **Sozu**, despite being one of the most prolific contributors, maintains a more measured pace with an average of 68.33 seconds per *Otada*.
- **Sale** recorded the highest average time for a complete octave (83.25 seconds), suggesting a different stylistic approach to the rhythmic constraints of the performance.

It should be noted that these poetic performances are mainly sung and often accompanied by musical interludes. For this reason, it would be misleading to correlate temporal variations with lexical data, as this would not provide truly meaningful information about the creative abilities of the poets. Some poets, in fact, prefer to dwell on the vowels /a/, /e/ and /ɛ//.

For this reason, temporal characteristics should be considered as stylistic traits specific to each poet, rather than as descriptors of actual improvisational abilities.

Author	Otada (s)	Duina (s)	Batorina (s)
Sozu	68.33	11.55	26.69
Masala	45.57	8.51	20.80
Piras	53.83	–	–
Mura	65.41	12.00	31.71
Seu	66.95	10.90	30.67
Sale	83.25	–	–
Piredda	49.20	–	–
Budrone	56.89	–	–
Mean (All)	**58.18**	**10.02**	**24.93**

Table 4: Average execution time (seconds) for the primary metrical forms per author.

3.5. Co-occurrence and Formularity

To investigate recurrent multiword patterns in the corpus, we conducted an exhaustive n-gram analysis with n ranging from 3 to 8. The analysis is exploratory in nature: its aim is not to establish statistically validated categories of formulaic reuse, but to identify recurring multiword patterns and examine their functional distribution across poets and performance events. The analysis was carried out on the full transcribed corpus of 135,747 tokens and 12,808 unique word forms, comprising performances exclusively in the *ottava* meter (*otada*). Token extraction followed a conservative normalization pipeline: lowercase conversion, removal of punctuation while preserving apostrophized clitics, and whitespace normalization, with original surface forms retained separately for display purposes. Shorter sequences ($n \leq 4$) are more likely to include high-frequency grammatical collocations and are therefore less diagnostically reliable in isolation; sequences of five tokens or more, especially when recurring across distinct performance events or poets, provide stronger evidence of formulaic reuse, with higher-order sequences ($n \geq 6$) offering the most robust signal, being least likely to occur by chance.

For each candidate sequence, two association measures were computed. Pointwise Mutual Information (Church and Hanks, 1990) is defined as:

$$\text{PMI} = \log_2 \frac{P(w_1, \ldots, w_n)}{\prod_{i=1}^{n} P(w_i)}$$

where $P(w_1, \ldots, w_n)$ is the observed joint probability of the sequence and $\prod P(w_i)$ is the expected probability under lexical independence. PMI is a reliable indicator of lexical *distinctiveness* but unstable at low observed frequencies, as sequences containing rare items yield inflated values. It was therefore supplemented with the Log-Likelihood Ratio (Dunning, 1993), G^2:

$$G^2 = 2 \sum_i O_i \ln \frac{O_i}{E_i}$$

where O_i and E_i are observed and expected cell counts contrasting occurrences of the sequence against all other corpus positions. G^2 is more stable at low frequencies and penalizes sequences whose association score is driven by constituent rarity rather than genuine co-occurrence preference. Since all n-gram candidates were evaluated within a single exploratory pass over the corpus, G^2 values are not interpreted as hypothesis tests: multiple comparison corrections would be required for inferential validity, and given the exploratory nature of the analysis we adopt G^2 strictly as a *relative ranking measure* of lexical cohesion.

The verses considered are reported in Table 5, grouped into four functional categories assigned on the basis of qualitative distributional and pragmatic criteria, reflecting the different roles a verse can play within the overall structure of a performance.

Verse / Template	Poet(s)	n	Obs.	PMI	LLR (G^2)	Type
FC 1 — Deictic formulas						
de fronte a unu pópulu [ADJ] (transl. *in front of a [ADJ] people*)	Masala	5	8	27.29	286.68	Personal deixis formula
de su palcu in susu (transl. *from the stage up above*)	Mura / Masala	5	4	22.95	119.27	Shared deixis formula
FC 2 — Personal turn-management formulas						
[ADV/PRON] una cosa narrer ti cheria (transl. *[ADV/PRON] one thing I would like to tell you*)	Sozu	5	5	29.44	194.09	Personal formula (var. onset)
su chi ses nelzende pone cura (transl. *pay attention to what you are (weaving) saying*)	Piredda	6	3	42.72	171.66	Personal formula
FC 3 — Dialogic mirroring						
a morrer de piumu est un'onore (transl. *to die by lead (a bullet) is an honour*)	Piras → Sozu	8	2	46.62	125.25	Dialogic mirroring
chi si l'at mandigada sa rustia (transl. *that the blight has devoured it*)	Masala → Mura	7	2	45.50	122.15	Dialogic mirroring
FC 4 — Formulaic variants & cross-poet templates						
e cando mai chi non b'as cumpresu (transl. *and when have you ever not understood it*)	Piredda	8	2	45.80	122.98	Cross-poet template
e cando mai chi non l'as cumpresa (transl. *and when have you ever not understood it*)	Sozu	8	2	45.57	122.34	Cross-poet template
[V] est chi non l'as [V-PART] (transl. *[it] is that you have not [V-PART] it*)	Mura / Sale	5	3	16.71	63.48	Cross-poet template

Table 5: Recurring n-grams ($n = 5$–8) grouped by functional category (FC), ordered by LLR within each group. *Obs.* = observed frequency; PMI = Pointwise Mutual Information ($\log_2$); LLR = Log-Likelihood Ratio (G^2), used as a relative ranking measure; no significance thresholds are applied. Square brackets denote variable slots.

The **first group** illustrates deictic formulas with productive open slots. The 5-gram *de fronte a unu pópulu [ADJ]* ($PMI = 27.29$, $G^2 = 286.68$, obs. 8) is attested across eight distinct performances of Masala with five adjectival completions (*signore*, *devotu*, *gentile*, *'e zente*, *festosu*), and is classified as an idiolectal formula. The 5-gram *de su palcu in susu* ($PMI = 22.95$, $G^2 = 119.27$, obs. 4) is distributed across Mura and Masala in four distinct performances, with a variable left context (*da chi t'agatas*, *las improviso*, *chi as cantadu*, *da chi m'agato*) and a frozen deictic core, indicating membership in the shared traditional repertoire. These two sequences illustrate a gradient from idiolectal formulas with a productive terminal slot to cross-poet templates with a productive left context.

The two instances grouped under the **second group** represent a functionally coherent class of *personal formulas* used to manage the poet's own turn at the opening or early internal position of a strophe. Both n-grams are strictly monoauctorial — all attested occurrences are attributed to a single poet — and both carry a consistent pragmatic function across distinct performance events and distinct opponents. The first verse is the 5-gram *[ADV/PRON] una cosa narrer ti cheria* ($PMI = 29.44$, $G^2 = 194.09$, obs. 5), attested across five distinct performances of Sozu from 1958 to the early 1980s with three onset realizations (*ma*, *deo*, *solu*). The five-word core is lexically frozen while the initial position accommodates metrically interchangeable elements. The formula marks a discourse-level shift to direct personal address and is exclusive to Sozu across the corpus, constituting evidence of idiolectal formulaic stability over a long temporal span. The second verse, *su chi ses nelzende pone cura* (transl. *pay attention to what you are weaving*; $n = 6$, obs. $= 3$, $PMI = 42.72$, $G^2 = 171.66$), is exclusively associated with Piredda and distributed across three distinct performance events (1976 and twice in 1980) involving three distinct opponents, suggesting idiolectal stability over at least a four-year span.

The third group collects instances of dialogic mirroring, attested here by two examples. The 8-gram *a morrer de piumu est un'onore* ($PMI = 46.62$, $G^2 = 125.25$, obs. 2) occurs in two consecutive strophes of the same performance (see Table 6 in the Appendix): Piras deploys it as the closing verse of strophe #62 and Sozu opens his immediate response (strophe #63) with that same line, recontextualizing the formula to dismantle the argument

it had supported. Where Piras invokes dying by bullet as an honor applicable to the bandit, Sozu accepts the formula but redirects it to argue that such honor is not indiscriminate. The second case, the 7-gram *chi si l'at mandigada sa rustia* (strophes #8–#9, Masala $\rightarrow$ Mura, $PMI = 45.50$, $G^2 = 122.15$, obs. 2 see; Table 7 in the Appendix)), follows the same structural logic. Masala introduces the verse as a closing statement on the poor harvest, and Mura opens the following strophe by echoing it verbatim before expanding the causal frame, adding the *peronòspera* (downy mildew) as a further agent of destruction. Both cases instantiate verbatim n-gram repetition across a turn boundary, yet in functionally distinct ways: argumentative inversion, in which the repeated formula is turned against its original claim, and elaborative extension, in which it is accepted and expanded with additional causal content.

The **fourth functional group** collects sequences that instantiate the same pragmatic act — asserting the opponent's failure to understand — at different levels of lexical fixity. The 8-gram *e cando mai chi non b'as cumpresu*(Piredda, $PMI = 45.80$, $G^2 = 122.98$, obs. 2) and *e cando mai chi non l'as cumpresa* (Sozu, $PMI = 45.57$, $G^2 = 122.34$, obs. 2) are near-minimal pairs sharing the invariant core *cando mai chi non [PRON] as cumpress-*, differing only in clitic form and participial gender inflection. Their attribution to two distinct poets indicates a cross-poet template. The 5-gram *est chi non l'as [V-PART]* ($PMI = 16.71$, $G^2 = 63.48$, obs. 3), distributed across Mura and Sale with completions *cumpresa* and *iscritu*, represents a more abstract level: a productive syntactic frame with an open participial slot.

4. Discussion

This study aimed to address two interconnected objectives: the construction of a structured digital corpus of Sardinian extemporaneous poetry suitable for computational analysis, and the empirical investigation of formulaic behavior within this tradition.

On the first front, **A Bolu** represents, to our knowledge, the first resource of its kind for the *cantada logudoresa*. Its hierarchical architecture demonstrates that the fragmentary and discontinuous nature of oral tradition heritage can be transformed into a structured and computationally accessible dataset without sacrificing contextual richness.

Lexical and temporal analyses provide descriptive baselines that characterize the individual stylistic profiles of the eight poets. The shared lexical core that emerges from the aggregate count of types, combined with individual variation in hapax legomena and performance tenses, suggests that the poets operate within a common expressive framework while maintaining recognizable personal signatures. These findings provide the necessary empirical grounding for the more theoretically significant question of formularity: the observed n-gram recurrences are not an artifact of lexical poverty but reflect genuine formulaic preference within a rich and varied individual repertoire.

On this front, the n-gram analysis yields the most significant results. The recurrence of higher-order n grams across distinct competitions and the positional coherence of reused lines suggest the existence of a formulaic layer in the compositional process of the *cantada logudoresa*, consistent with the **Oral-Formulaic Theory** developed by Parry and Lord (Parry and Parry, 1987; Lord, 1995). Originally conceived to account for the compositional mechanisms of Homeric epic poetry, the theory posits that oral poets rely on a shared formulaic repertoire to navigate complex metrical constraints under real-time cognitive pressure. In the context of the *cantada logudoresa*, such formulas serve a dual function: they operate as mnemonic anchors and as structural building blocks, enabling the improvising poet to construct metrically and thematically coherent octaves while responding spontaneously to a rival.

The evidence further points to a layered architecture of formulaic competence, operating simultaneously at the collective and individual level. At the collective level, cross-poet templates indicate membership in a shared traditional repertoire. At the individual level, strictly monoauctorial formulas — stable across different opponents and performance events spanning years or decades — suggest that formulaic competence also has a deeply personal dimension, individually cultivated over long stretches of a poet's career, in line with the idiolectal dimension of oral-formulaic composition discussed in Foley (Foley, 1988).

An interesting result is the interactional dimension of this formularity. The reuse of a rival's line as the opening of one's own strophe suggests that formulas function not only as individual mnemonic devices, but also as competitive rhetorical strategies — a dimension that the classical Oral-Formulaic Theory, developed primarily in the context of monologic epic composition, does not explicitly account for. Two distinct modes are attested: argumentative inversion, in which the repeated formula is turned against its original claim, and elaborative extension, in which it is accepted and expanded with additional content. This finding may point to a distinctive feature of agonistic improvisational traditions more broadly.

Overall, the data are consistent with the plausibility of the oral-formulaic hypothesis applied to this tradition. However, given the exploratory nature of

the analysis, these results should be interpreted as stimuli for further investigation rather than as a formal test of the theory. Larger datasets and methods capable of identifying near-variant formulas beyond exact string matching will be needed to move from suggestive evidence to robust empirical validation.

5. Conclusion and Future Works

This study has introduced **A Bolu**, the first structured digital corpus of Sardinian extemporaneous poetry designed for computational analysis, and has presented a preliminary investigation of lexical, temporal, and formulaic dimensions of the *cantada logudoresa*. The results provide preliminary empirical support for the oral-formulaic hypothesis within this tradition, and reveal an interactional dimension of formulaic reuse that warrants further theoretical attention.

Several directions for future work follow naturally from the limitations of the present study. The most pressing priority is the expansion of the corpus itself: incorporating new poetic sessions and a broader range of poets would substantially increase the statistical power of the analysis and improve the generalizability of the findings beyond the eight performers currently represented. A larger and more diverse dataset would also allow for more reliable cross-poet comparisons, which are currently constrained by the pronounced imbalance in subcorpus sizes.

A particularly promising avenue is the integration of the audio dimension. The current corpus is exclusively text-based, which means that relevant prosodic and musical features of the performance — including melodic contour, vowel prolongation, and rhythmic patterning — remain outside the scope of the analysis. The inclusion of aligned audio recordings would not only allow for a more accurate interpretation of execution time data, but would also open the way for acoustic and phonetic analyses capable of capturing aspects of improvisational competence that are invisible at the textual level.

Finally, the development of more sophisticated metrics for evaluating stanza complexity represents a critical methodological challenge for future work. The descriptive indices used here — hapax counts, raw n-gram frequencies, TTR, MATTR and MTLD — provide a useful baseline but are insufficient to capture the full richness of the stylistic, linguistic, and cognitive phenomena at play in real-time oral composition. Length-normalized diversity metrics, network-based models of inter-poet formulaic exchange, and methods for detecting near-variant formulas beyond exact string matching would all contribute to a more nuanced understanding of how individual poets navigate the tension between metrical constraint, thematic coherence, and improvisational speed. Ultimately, such developments would position the **A Bolu** corpus as a reference resource not only for the study of Sardinian oral traditions, but for the broader computational investigation of extemporaneous poetic performance as a window into the stylistic, linguistic, and cognitive dimensions of human creativity under constraint.

Acknowledgements

The authors wish to thank the editorial staff of *Làcanas* and the publisher *Domus de Jana* for their kind availability and for granting permission to use their data in the construction of this corpus.

6. Bibliographical References

Manuela Angioni, Franco Tuveri, Maurizio Virdis, Laura Lucia Lai, and Micol Elisa Maltesi. 2018. SardaNet: A linguistic resource for Sardinian language. In *Proceedings of the 9th Global WordNet Conference*, pages 412–419, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

R Harald Baayen. 2001. *Word frequency distributions*, volume 18. Springer Science & Business Media.

Salvatore Mario Carta, Stefano Chessa, Giulia Contu, Andrea Corriga, Andrea Deidda, Gianni Fenu, Luca Frigau, Alessandro Giuliani, Luca Grassi, Marco Manolo Manca, Mirko Marras, Francesco Mola, Bastianino Mossa, Paola Mura, Marco Ortu, Leonardo Piano, Simone Pisano, Antonella Pisu, Alessandro Sebastian Podda, Livio Pompianu, Sara Seu, and Simona Giuseppina Tiddia. 2025. A BERT-based approach for Part-of-Speech tagging in the Sardinian language. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*. CEUR-WS.

Luis Chiruzzo, Santiago Góngora, Aldo Alvarez, Gustavo Giménez-Lugo, Marvin Agüero-Torales, and Yliana Rodríguez. 2022. Jojajovai: A parallel Guarani-Spanish corpus for MT benchmarking. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2098–2107, Marseille, France. European Language Resources Association.

Ilaria Chizzoni and Alessandro Vietti. 2024. Towards an ASR system for documenting endangered languages: A preliminary study on Sardinian. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, volume 3878 of *CEUR Workshop Proceedings*, Pisa, Italy. CEUR-WS.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Michael A Covington and Joe D McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

John Miles Foley. 1988. *The Theory of Oral Composition: History and Methodology*. Indiana University Press, Bloomington, IN.

Loïc Grobol and Mélanie Jouitteau. 2024. ARBRES Kenstur: A Breton-French parallel corpus rooted in field linguistics. In *Proceedings of the Fourteenth Language Resources and Evaluation Conference*, Torino, Italy. European Language Resources Association.

Albert Bates Lord. 1995. *The singer resumes the tale*. Cornell University Press.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Walter J. Ong. 1982. *Orality and Literacy: The Technologizing of the Word*. Routledge, London.

Milman Parry and Adam Parry. 1987. *The making of Homeric verse: The collected papers of Milman Parry*. Oxford University Press.

Alan Ramponi. 2024. Language varieties of Italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38. First appeared as arXiv preprint arXiv:2209.09757, September 2022.

Redazione Làcanas. 2014. Poetas a bolu. Rubrica online, *Làcanas – Rivista bilingue delle identità*. Accessed: 2025. Available at: `https://www.lacanas.it/rubrica/poetas-a-bolu/`.

Maurizio Virdis. 1988. Sardisch: Areallinguistik. In Günter Holtus, Michael Metzeltin, and Christian Schmitt, editors, *Lexikon der Romanistischen Linguistik (LRL)*, volume 4, pages 897–913. Max Niemeyer Verlag, Tübingen.

Paul Zumthor. 1997. The construction of orality. *Poetry in Speech: Orality and Homeric Discourse*, page 7.

7. Language Resource References

Language Resources

Silvio Calderaro and Johanna Monti. 2026. *A Bolu: a Structured Dataset for the Computational Analysis of Sardinian Improvisational Poetry*. PID https://doi.org/10.5281/zenodo.19264263.

Appendix A. Dialogic Mirroring: Full Stanzas and English Translations

Original Sardinian Text	
Stanza #62 (Piras)	**Stanza #63 (Sozu)**
Ma cussos una fama tenen totu e sempre de sa zente sun a galla. A tie fatu t'an sa facia gialla proite su valore nde as connotu. Nara proite su Milite Ignotu no fut bandidu e moltu l'an a balla. Est moltu che bandidu e gherradore: **a morrer de piumu est un'onore.**	**A morrer de piumu est un'onore** ma cuss'onore non deghet a totu prite disparidade no as connotu o lis pones su propiu valore chi a fiancu 'e Marras e Pintore mi cheres ponner su Milite Ignotu ch'in su Vitorianu che at sa losa: creo no siat sa matessi cosa.
English Translation	
But they all have a reputation *and are always held high by the people.* *To you, they made your face turn pale* *because you have known their value.* *Tell me why the Unknown Soldier* *was not a bandit, yet he was killed by a bullet.* *He died like a bandit and a warrior:* ***to die by lead is an honor.***	***To die by lead is an honor*** *but that honor does not suit everyone* *because you have seen no disparity* *or you give them the same value* *when, alongside Marras and Pintore,* *you want to place the Unknown Soldier* *who has his tomb in the Vittoriano:* *I believe it is not the same thing.*

Table 6: Comparison of stanzas for the *piumu* occurrence.

Original Sardinian Text	
Stanza #8 (Masala)	**Stanza #9 (Mura)**
S'annu passadu s'annada fit mala, canta sicagna e canta caristia! In tota sa Sardigna de ua ebbia si nd'at salvadu solu calchi iscala: fit cosa in generale e in-d-ogni ala **chi si l'at mandigada sa rustia.** Ma bell'e gai s'oju nd'at annotu chi su tantu 'e su 'inu ch'est etotu.	**Chi si l'at mandigada sa rustia** tue afirmadu as che poesianu. No fit s'annu passadu cuss'ebbia, bistadu est su tribagliu totu invanu e ocannu ca est fritu su 'eranu sa 'ide l'at distruta sa 'iddia: s'annu passadu sende a bide pròspera mandigada si l'at sa peronòspera.
English Translation	
Last year the season was bad, *so much drought and so much famine!* *In all of Sardinia, only a few clusters* *of grapes were saved here and there:* *it was a general thing in every part* ***that the frost has eaten it all up.*** *But even so, the eye notices* *that the amount of wine is still the same.*	***That the frost has eaten it all up*** *you have stated like a poet.* *It wasn't just like that last year,* *all the labor was in vain* *and this year, because the spring is cold,* *the frost has destroyed the vine:* *last year, while the vine was prosperous,* *the downy mildew ate it up.*

Table 7: Comparison of stanzas for the *rustia* occurrence.