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

Marija Pendevska, Hristina Nastevska
Ss. Cyril and Methodius University of Skopje
Boulevard Goce Delchev 9, Skopje 1000, North Macedonia
{m.pendevska, hristinanastevska}@gmail.com

Abstract

This paper presents a knowledge management framework for the digital preservation of Aromanian, an endangered Eastern Romance language spoken across the Balkans, combined with the first systematic evaluation of automatic speech recognition (ASR) models on Aromanian dialectal speech. The proposed three-module framework encompasses localization of existing resources, distribution through digital platforms, and creation of new linguistic content through technological innovation. Empirical analysis of knowledge management practices among 176 respondents validates the framework design, revealing that information quality (28.3%) and personal utility (24.1%) are the most valued criteria for knowledge sharing, while digital information seeking (27.1%) is the dominant behaviour. Within this framework, we evaluate models from the OpenAI Whisper family across three sizes (medium, large-v2, large-v3) and multiple language settings on two Aromanian varieties: Gramosteanj and Crushova. All configurations yield word error rates (WER) above 88%, with character error rates (CER) as low as 34% under optimal conditions, indicating partial phonotactic capture despite word-level failure. The Latin language setting with large-v3 consistently achieves the best results. Romanization of non-Latin script output substantially reduces CER, confirming script mismatch as a major error source. These findings underscore the limitations of current pretrained ASR models for endangered languages and the need for dedicated resources and adaptation strategies within a broader language preservation framework.

Keywords: Aromanian, endangered language preservation, automatic speech recognition, knowledge management, Whisper, dialectal variation, digital preservation

1. Introduction

The Aromanian language represents a critical case in the preservation of endangered languages. With only a few thousand predominantly adult speakers remaining, Aromanian faces imminent extinction despite official recognition as a minority language in North Macedonia (Kahl, 2008). This precarious situation reflects broader global challenges: approximately 40% of the world's 6,000–7,000 languages are considered endangered (Moseley, 2010), and speakers of regional and underrepresented varieties remain largely excluded from modern speech and language technologies.

The need for systematic digital preservation of Aromanian extends beyond traditional documentation approaches that focus primarily on linguistic analysis. As Crystal (2000) emphasizes, language revitalization requires not only documentation but also the capacity for active use in contemporary contexts. This study addresses this gap by proposing a comprehensive framework that treats language preservation as a knowledge management challenge, encompassing collection, storage, distribution, accessibility, and continuous creation of knowledge (Pendevska, 2019).

Recent advances in automatic speech recognition (ASR), driven by large-scale multilingual models such as Whisper (Radford et al., 2022), XLS-R (Babu et al., 2021), and Omnilingual ASR (Omnilingual ASR Team et al., 2025), have achieved strong performance on high-resource standard languages. However, these improvements do not readily extend to low-resource or dialectally diverse languages, where recognition accuracy degrades substantially (Blaschke et al., 2025). Systematic evaluation on endangered languages is essential to understand model behaviour and guide preservation strategies.

This paper addresses the intersection of two critical needs: (i) the development of a systematic framework for digital preservation of Aromanian using knowledge management (KM) principles (Pendevska, 2019), and (ii) the first empirical evaluation of ASR technology on Aromanian speech. We propose a three-module framework encompassing localization, distribution, and creation of linguistic resources, and within this framework evaluate Whisper models on two Aromanian varieties—Gramosteanj and Crushova—across multiple language settings. Our contributions are fourfold: we establish the first ASR benchmark for Aromanian; we conduct a systematic evaluation across model sizes and language configurations; we analyze the impact of script mismatch and romanization; and we situate these findings within a replicable KM framework for endangered language communities, validated through empirical analysis of knowledge management practices.

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 327–336
16 May 2026.

The significance of this research extends beyond the Aromanian language. The proposed framework offers a replicable model for endangered language preservation that can be adapted for similar linguistic communities worldwide. By integrating traditional linguistic documentation methods with modern digital technologies and KM practices, this study contributes to both language preservation efforts and knowledge management theory.

2. Background and Theoretical Framework

2.1. The Aromanian Language and Its Status

Aromanian (also known as Vlach or Macedo-Romanian; ISO 639-3: rup) is an Eastern Romance language spoken in fragmented communities across the southern Balkans, with the largest concentrations in Greece, North Macedonia, Albania, and Romania. Due to prolonged multilingual contact, Aromanian has absorbed substantial lexical, phonological, and morphosyntactic influence from Greek, Albanian, and South Slavic languages (Capidan, 1932; Friedman, 2012). An educational programme exists in North Macedonia, where the language has been taught as an elective subject in grades 3–9 since 2007.

Early scholarly documentation by Weigand (1895) and Capidan (1932) established the foundational linguistic record for Aromanian, distinguishing northern dialects (Farsheroti and other Albanian groups) from southern dialects (Gramosteanj and groups around Mount Olympus), with the Farsheroti representing a distinct variety. Marković (2007) further refined this classification, identifying that northern dialects (Farsheroti, Moskopoleans, and Muzachins) lack the characters ã and â found in southern dialects (Gramosteanj and Pindeani). Contemporary scholars including Nastev (1988) and Kahl (2008) have continued this tradition in areal linguistic studies and comparative analyses. Figure 1 illustrates the geographic distribution of the Eastern Romance language varieties in the southwestern Balkans, highlighting the fragmented nature of Aromanian-speaking communities.

2.2. Standardization and Orthographic Challenges

The systematization of Aromanian writing began approximately one century ago with the Romanian schools movement in Macedonia. The initial orthographic approach adopted the Romanian alphabet but faced challenges in representing phonemes unique to Aromanian. The language remained unstandardized until the 1997 Bitola Standardization

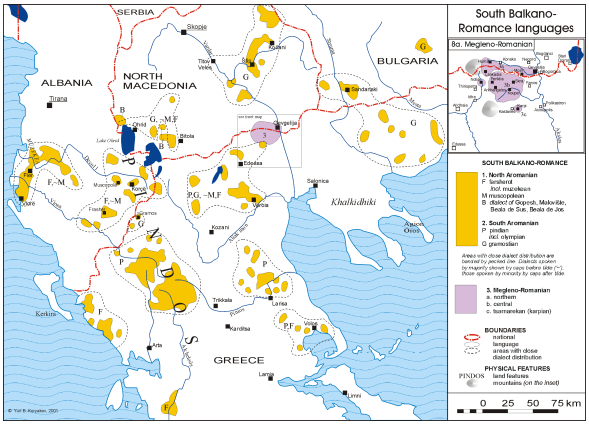

Figure 1: Distribution and dialects of the Eastern Romance languages in the southwestern Balkans. Source: Wikimedia Commons (CC BY-SA 3.0).

Symposium, which established foundational conventions by adopting a 35-character alphabet designed for compatibility with digital systems (Cunia, 2000). This represents a compromise between phonemic accuracy and technological accessibility, accommodating dialectal variation while maintaining orthographic unity (Marković, 2007).

However, the orthographic question remains sensitive and unresolved across the broader Aromanian-speaking world. The Bitola standard has been adopted by most Aromanian writers in North Macedonia, Serbia, Albania, Bulgaria, and Romania, but it is not officially recognized in Greece, where Aromanian has no formal status and alternative orthographic approaches have been proposed. Notably, Beis and Dasoulas (2017) present a Latin-based writing system developed through the teaching activities of the Society of Aromanians of Athens, designed to capture Aromanian sounds that do not exist in Greek or Romanian. Their proposal illustrates that even recent approaches from the Greek side converge on Latin-based orthography, while diverging from the Bitola conventions in specific character choices. This ongoing debate reflects the broader sociolinguistic complexity of standardizing a language spoken across multiple national contexts with differing language policies.

In North Macedonia, an east–west dialectal opposition exists, divided by the course of the Vardar river. The two varieties examined in this study—Gramosteanj (originating from the Gramos mountain region, spoken dominantly in the eastern part of North Macedonia) and Crushova (from Kruševo, western part of North Macedonia)—represent geographically and linguistically distinct communities within this division. The Gramosteanj dialect belongs to the southern variety group, while the Crushova dialect is a northern variety (Marković, 2007; Nastev, 1988). Both are written in the Latin script with community-specific orthographic con-

ventions. This late and incomplete standardization, combined with urbanization and globalization pressures, has contributed to the language's current endangered status.

Contemporary Aromanian communities are dispersed globally, with speakers found across the Balkan Peninsula (North Macedonia, Greece, Albania, Serbia, Romania) and diaspora communities in the USA and Australia. This geographic dispersion presents challenges but also opportunities for digital preservation, as digital platforms can connect speakers beyond traditional boundaries while also capturing dialectal variations.

2.3. Knowledge Management Theory for Language Preservation

Knowledge management theory provides a structured approach to linguistic preservation extending beyond traditional archival methods. The Knowledge Management Cycle (KMC) model (Evans et al., 2014) offers a seven-phase framework—identify, preserve, share, use, learn, improve, and create new knowledge—that aligns with the needs of endangered language communities, which must simultaneously preserve existing knowledge and create content for contemporary use.

Pendevska (2019) demonstrated how KM principles can be systematically applied to innovation processes, examining knowledge management practices at the individual, group, and organizational levels. This work provides the theoretical foundation for the framework proposed in this study, establishing that systematic approaches to knowledge flow organization are essential for effective preservation and distribution of knowledge. Mittelmann et al. (2022) further emphasize that effective KM requires systematic organization and distribution networks ensuring knowledge reaches relevant stakeholders with appropriate quality. For endangered languages, this means creating platforms that serve both community members wishing to maintain cultural identity and external stakeholders interested in studying linguistic diversity.

The relentless advance of digital technologies has transformed the possibilities for language study. Text-to-speech synthesis, automated language processing, and online learning platforms now enable endangered languages to participate in digital communication ecosystems (Rehm and Uszkoreit, 2012). The development of speech corpora and linguistic resources for low-resource languages has become increasingly feasible through advances in machine learning and natural language processing (Besacier et al., 2014). Digital technologies offer opportunities for creating interactive learning environments and preserving linguistic and cultural knowledge (Bird and Simons, 2002), while advances in ASR and NLP create new possibilities for endangered language participation in digital ecosystems (Joshi et al., 2020).

2.4. ASR for Endangered and Dialectal Speech

ASR for dialectal and non-standard speech remains a persistent challenge. Performance degradation has been documented for regional varieties across diverse language families, including Arabic (Nasr et al., 2023), Japanese (Takahashi et al., 2024), German (Blaschke et al., 2025; Sicard et al., 2023), and Indian languages (Alumäe et al., 2023).

Particularly relevant is recent work on Modern Greek dialects. Vakirtzian et al. (2025) introduced the first ASR benchmark for Greek varieties and showed that zero-shot Whisper performance degrades sharply with dialectal distance from Standard Modern Greek (SMG), with WER exceeding 100% for contact-influenced varieties. Tsoukala et al. (2026) expanded this benchmark, confirming a strong interaction between dialectal distance and ASR difficulty: southern dialects achieved WERs of 60–70%, while the contact-influenced Cappadocian yielded WER near 97%. Fine-tuning substantially reduced error rates across all varieties. These findings are directly relevant to Aromanian, which as a non-Greek Romance language with extensive Balkan contact features and no representation in ASR training data, presents an extreme out-of-distribution case.

3. Knowledge Management Framework for Aromanian

3.1. Three-Module Framework

We propose a three-module framework grounded in KM principles for the systematic preservation and revitalization of Aromanian (Pendevska, 2019). The framework is organized around the KMC model (Evans et al., 2014), with each module addressing specific functional requirements while maintaining integration across the full knowledge management lifecycle:

Module 1: Language Localization focuses on identifying, collecting, and digitally preserving existing linguistic and cultural resources. This encompasses historical documentation (Weigand, 1895; Capidan, 1932), television archives spanning nearly 30 years of Aromanian-language broadcasting in North Macedonia, and family sources. Content collection from family archives employs structured questionnaires covering personal data, family background, Aromanian language proficiency and identity status, and cultural practices. The dialectal mapping methodology includes audio record-

ings across different speaker communities, creating geographically representative documentation. Additionally, community engagement in populating the Mozilla Common Voice database (Mozilla Foundation, 2025) provides the foundation for building a contemporary speech corpus. Technical requirements include multiple file formats (television archives, personal recordings, visual documentation, slides, PDF) with intuitive navigation suitable for older users with limited digital experience.

Module 2: Language Learning and Distribution enables creation of accessible platforms for language education and community engagement. The digital platform provides tutorial lessons using synchronous and asynchronous media, incorporating active Aromanian teachers who contribute their pedagogical expertise for broader distribution. A key achievement is the Aromanian riddle generator presenting 100 riddles from Prof. Dr. Nastev's collection (Nastev, 1988) in Aromanian, Macedonian, and English (Pendevska et al., 2024), representing the first digital educational tool for Aromanian language learning. Distribution strategies emphasize community engagement through user-generated content creation, with integration of radio broadcasting, podcasts, and YouTube channels providing Aromanian content for wider audiences.

Module 3: Language Development and New Knowledge Creation addresses technological innovation. A significant milestone was the integration of Aromanian into the Mozilla Common Voice platform in May 2025, marking the first inclusion of an endangered and unstandardized language in this major open-source speech collection initiative (Mozilla Foundation, 2025). Speech synthesis development leverages collected data to create Aromanian-specific text-to-speech voices, facilitating language learning among younger digital-native generations and enabling integration of augmentative and alternative communication (AAC) devices worldwide. Language corpus processing enables automated linguistic analysis and supports machine learning applications for language learning and translation. Within this module, we evaluate the capabilities and limitations of current ASR technology for Aromanian, as reported in the following sections.

3.2. Empirical Validation Through KM Survey

Primary data collection included a survey examining KM practices among 176 respondents, conducted in 2017 (Pendevska, 2019). The research instrument addressed information gathering behaviours, evaluation criteria, and sharing practices at the individual, group, and organizational levels. The empirical analysis informs and validates the framework design through the findings reported below.

Information source	**%**
Internet / search engines / multimedia	27.1
Internal documents	18.8
Colleagues / experts	17.6
Training / seminars	14.1
Other sources	22.4

Table 1: Information sources used by respondents (N=176), adapted from Pendevska (2019).

Evaluation criterion	**%**
Information quality	28.3
Personal utility	24.1
Source reliability	19.3
Timeliness	15.2
Other criteria	13.1

Table 2: Most valued criteria for evaluating information (N=176), adapted from Pendevska (2019).

Table 1 presents results on information-seeking behaviour. Internet and search engine use (27.1%) is the dominant channel, followed by internal documents (18.8%), indicating strong readiness for digital language platforms. This finding suggests that digital approaches to language learning will align with established user behaviours.

Table 2 reports information evaluation criteria. Information quality (28.3%) and personal utility (24.1%) emerge as the most valued criteria, supporting the framework's emphasis on high-quality, relevant digital content rather than simply digitizing existing materials without curation.

The survey reveals strong efficiency expectations: a majority of respondents (65.5%) expect to find personal information within five minutes, while shared organizational information requires somewhat longer (46.7% find it within 5–15 minutes). These expectations indicate that digital language platforms must prioritize intuitive navigation and efficient search capabilities.

Table 3 presents information sharing practices. Text remains the primary format (29.6%), followed by tabular data (26.2%) and images/presentations (22.6%), while multimedia recordings are less common (7.5%), suggesting potential for growth in audiovisual language content. E-mail is the most used sharing channel (34.9%), followed by meetings (18.1%).

These findings validate each module: the prevalence of digital information seeking (27.1%) and internal document consultation (18.8%) support Module 1's focus on searchable digital archives; the preference for e-mail communication (34.9%) and meeting-based sharing (18.1%) validate Module 2's

Sharing format	%
Text	29.6
Tabular data	26.2
Images / presentations	22.6
Multimedia recordings	7.5
Other formats	14.1

Table 3: Information sharing formats used by respondents (N=176), adapted from Pendevska (2019).

Dialect	Original	Processed
Crushova	55:42	39:45
Gramosteanj	42:17	42:14

Table 4: Audio duration (mm:ss) before and after preprocessing.

provision of both asynchronous and synchronous learning opportunities; and the prioritization of information quality (28.3%) and personal utility (24.1%) support Module 3's focus on creating relevant, high-quality technological tools.

4. ASR Evaluation

4.1. Speech Data

Speech data for the ASR evaluation were collected from native speakers of the Gramosteanj and Crushova varieties reading dialectal texts written following the orthographic conventions discussed in Section 2.2. For standardization, all audio recordings were converted to single-channel (mono) WAV files with a sampling rate of 16 kHz.

Initially, the audio and corresponding texts were automatically aligned using the CTC Forced Aligner (CTC Forced Aligner Team, 2024). Following this automated step, the alignment boundaries were manually inspected and corrected in Praat (Boersma and Weenink, 2024) to ensure accuracy. The resulting TextGrids and audio files were then used to segment the recordings into individual utterance-level files for the evaluation phase. During this process, non-speech material such as long pauses and other untranscribed portions was removed to isolate clean speech segments. As a result, the effective audio duration was reduced for some datasets, as reflected in Table 4. Ground-truth transcriptions follow the orthographic conventions of each speaker community, reflecting the Bitola-based Latin-script standard with community-specific adaptations.

The evaluation data, including audio segments and aligned transcriptions, are publicly available on OSF. Publishing this data as the first open ASR evaluation set for Aromanian supports reproducibility and enables future benchmarking by the research community.

4.2. Models and Language Settings

We evaluate three Whisper models (Radford et al., 2022): medium, large-v2 (trained on ∼680k hours), and large-v3 (trained on ∼1M hours of weakly labeled audio plus ∼4M hours of pseudo-labeled data). All evaluations are zero-shot, without Aromanian-specific adaptation.

Since Aromanian is not supported by Whisper, we test multiple target language settings selected for linguistic proximity and contact history: Albanian (sq), Latin (la), Macedonian (mk), Greek (el), Bulgarian (bg), Romanian (ro), and automatic detection. These settings were chosen to cover genealogically related languages (Latin, Romanian), contact languages (Greek, Albanian, Macedonian, Bulgarian), and the model's own language identification capability. Romanian and Bulgarian were additionally motivated by the observation that automatic language detection frequently classified Aromanian utterances as Romanian or Bulgarian. The same set of models and language settings was evaluated for both varieties.

4.3. Evaluation Metrics and Romanization

Performance is measured using two standard metrics: Word Error Rate (WER) and Character Error Rate (CER). WER compares the machine-generated transcription to the ground-truth reference word by word, calculated as the total number of word substitutions, deletions, and insertions divided by the total number of reference words. CER operates at the character level, making it particularly useful for morphologically rich languages and for evaluating fine-grained accuracy. In cases where the target or detected language utilizes a non-Latin script (e.g., Greek, Macedonian, Bulgarian), the model's output was transliterated to the Latin script using the universal romanizer uroman (Hermjakob et al., 2018), ensuring a phonetically grounded comparison between the model's transcription and the Latin-script ground truth.

5. Results

5.1. Gramosteanj

Table 5 reports results for the Gramosteanj dialect. The best overall performance is achieved by large-v3 with Latin, yielding the lowest WER (91.60%) and CER (34.45%)—the only setting with WER below 100%.

Albanian consistently produces the lowest CER among non-Latin-output settings (42–44%), sug-

gesting phonological overlap between Aromanian and Albanian is partially captured at the character level. Romanian yields competitive results, with large-v3 achieving WER of 99.30% and CER of 43.24% (39.27% after romanization), notable given the genealogical proximity between the two languages. For Cyrillic-output languages, standard CER exceeds 90% due to script mismatch; after romanization, CER drops substantially (e.g., Macedonian from 90.78% to 41.90% and Bulgarian from 90.72% to 46.17% with large-v3).

5.2. Crushova

Table 6 reports results for the Crushova dialect. The pattern mirrors Gramosteanj: **large-v3 with Latin** achieves the best performance (WER 88.67%, CER 35.57%)—again the only sub-100% WER configuration.

Romanian yields competitive CER with large-v3 (49.26%), notable given that Romanian is genealogically close major language. However, the model's Romanian decoder still cannot reliably reconstruct Aromanian words (WER 100.99%). Bulgarian and Macedonian show similar CER patterns, with standard CER above 90% dropping to approximately 45–54% after romanization.

5.3. Cross-Dialect Comparison

The two dialects exhibit remarkably similar performance profiles. Crushova achieves slightly lower best-case WER (88.67% vs. 91.60%) and comparable best-case CER (35.57% vs. 34.45%). This consistency across two geographically and sociolinguistically distinct communities suggests that the observed performance levels reflect fundamental model limitations with respect to Aromanian rather than dialect-specific properties.

6. Discussion

6.1. Comparison with Greek Dialect ASR

Our results can be contextualized by comparison with Greek dialectal ASR benchmarks. Vakirtzian et al. (2025) reported zero-shot Whisper large-v3 WER of 58.42% on Eastern Cretan and 32.71% on Messenian—varieties close to SMG—but WER exceeding 100% for contact-influenced Aivaliot and Griko. Tsoukala et al. (2026) showed WERs of 61.85% (Aperathiot), 70.24% (Cretan), 80.87% (Lesbian), and 96.65% (Cappadocian).

Aromanian, with best zero-shot WER of 88–92%, falls between Lesbian and Cappadocian in difficulty. This is consistent with Aromanian's position as a language entirely unrepresented in training data and exhibiting substantial contact-induced lexical borrowing. The parallel with Cappadocian—both are contact-influenced varieties representing extreme out-of-distribution cases—is particularly instructive.

6.2. Language Setting and Script Effects

The Latin setting consistently outperforms all other configurations, likely because it eliminates script mismatch and shares phonological features with Aromanian as a fellow Romance language. Albanian, despite producing Latin-script output, yields higher WER, suggesting lexical and morphological differences outweigh script compatibility. The romanization experiments demonstrate that a large portion of CER for Cyrillic-output languages reflects script mismatch: after romanization via uroman, Macedonian CER drops by approximately 50 percentage points, reaching levels comparable to Albanian. This parallels the Griko findings of Vakirtzian et al. (2025).

6.3. WER–CER Divergence

Across all configurations, WER substantially exceeds CER, most pronounced for the best settings (e.g., large-v3 Latin: 91.60% WER vs. 34.45% CER for Gramosteanj). This indicates that while models fail to reconstruct correct word forms, they capture meaningful phonotactic and subword structure. The lower CER for Aromanian under optimal settings (34–36%) compared to Cappadocian (59.72%) may reflect the benefit of native Latin-script alignment.

6.4. Implications for the Preservation Framework

The ASR evaluation results have direct implications for the KM framework (Pendevska, 2019). Within Module 3, the findings demonstrate that current off-the-shelf ASR models cannot produce usable Aromanian transcriptions without dedicated adaptation. This underscores the importance of community-driven data collection initiatives such as the Mozilla Common Voice integration, which will provide the training data necessary for future model fine-tuning. The high CER under optimal conditions (34–36%) suggests that fine-tuning on even modest amounts of dialectal data—following the approach shown effective for Greek dialects (Tsoukala et al., 2026)—could yield substantial improvements, potentially enabling practical ASR applications for documentation and education within the Module 2 distribution platform.

The empirical KM findings further reinforce these implications. The dominance of digital information seeking (27.1%) and the strong efficiency expectations (65.5% expect results within five minutes) indicate that digital language preservation

Model	Lang	WER (%)	CER (%)	WER rom. (%)	CER rom. (%)
medium	sq	100.80	42.15	—	—
medium	la	120.86	80.71	—	—
medium	mk	108.94	93.69	103.79	47.42
medium	el	111.14	98.85	109.59	68.44
medium	bg	109.90	92.40	104.62	46.97
medium	ro	100.78	46.16	99.90	43.02
medium	auto	106.83	71.84	103.79	45.85
large-v2	sq	109.20	43.90	—	—
large-v2	la	97.87	41.87	—	—
large-v2	mk	106.31	91.35	99.37	42.36
large-v2	el	109.67	97.96	108.03	64.39
large-v2	bg	109.43	92.47	103.46	46.00
large-v2	ro	102.36	45.70	100.94	42.10
large-v2	auto	105.35	58.87	103.31	43.92
large-v3	sq	112.09	44.26	—	—
large-v3	la	**91.60**	**34.45**	—	—
large-v3	mk	106.15	90.78	100.36	41.90
large-v3	el	106.45	94.36	104.73	60.11
large-v3	bg	107.39	90.72	102.34	46.17
large-v3	ro	99.30	43.24	97.72	39.27
large-v3	auto	107.32	68.98	104.19	44.83

Table 5: ASR results on the Gramosteanj dialect. Romanized (rom.) metrics are reported where transliteration to Latin script was applied. Best results in bold.

Model	Lang	WER (%)	CER (%)	WER rom. (%)	CER rom. (%)
medium	sq	101.78	45.83	—	—
medium	la	105.79	72.83	—	—
medium	mk	110.25	93.55	106.35	54.05
medium	el	113.29	100.83	112.18	82.25
medium	bg	111.26	93.25	107.11	54.50
medium	ro	114.12	63.02	113.02	60.31
medium	auto	104.73	62.02	102.32	47.00
large-v2	sq	108.57	46.44	—	—
large-v2	la	97.14	42.77	—	—
large-v2	mk	109.15	92.08	104.24	47.92
large-v2	el	113.14	101.35	112.13	77.34
large-v2	bg	111.26	92.84	106.77	51.98
large-v2	ro	109.69	58.39	108.39	55.26
large-v2	auto	108.40	57.82	107.24	50.00
large-v3	sq	112.43	46.48	—	—
large-v3	la	**88.67**	**35.57**	—	—
large-v3	mk	108.61	91.07	103.45	45.42
large-v3	el	110.29	97.17	109.39	74.05
large-v3	bg	110.62	91.25	106.35	54.04
large-v3	ro	100.99	49.26	99.55	46.22
large-v3	auto	107.74	63.12	105.27	47.97

Table 6: ASR results on the Crushova dialect. Romanized (rom.) metrics are reported where transliteration to Latin script was applied. Best results in bold.

platforms must prioritize intuitive navigation and efficient search capabilities. The emphasis on information quality (28.3%) as the top evaluation criterion confirms that automated transcription tools, once sufficiently accurate, would be highly valued by the target user community. The relatively low current usage of multimedia recordings for sharing (7.5%) suggests significant potential for growth as speech technologies improve.

7. Conclusion

This paper presents an integrated approach to Aromanian language preservation, combining a knowledge management framework with the first systematic ASR evaluation for this endangered language. The three-module KM framework—encompassing localization, distribution, and creation (Pendevska, 2019)—provides a replicable model for endangered language communities, validated through empirical analysis of KM practices (N=176) and supported by concrete achievements including Mozilla Common Voice integration and the first digital Aromanian riddle generator (Pendevska et al., 2024).

The ASR evaluation demonstrates that current multilingual models fail to produce acceptable Aromanian transcriptions in zero-shot settings (WER $>$ 88%), but CER as low as 34–36% under optimal conditions indicate meaningful subword capture. The Latin language setting with Whisper large-v3 consistently achieves the best results, and romanization of non-Latin output substantially improves CER. These findings position Aromanian alongside contact-influenced Greek varieties as one of the most challenging cases for pretrained multilingual models.

The empirical KM findings—particularly the dominance of digital information seeking, the prioritization of information quality, and the strong efficiency expectations—provide actionable design guidelines for digital language preservation platforms. Future work should explore fine-tuning on community-collected data, evaluate additional model architectures, and conduct longitudinal studies tracking the framework's effectiveness in achieving language revitalization. The Aromanian case demonstrates that with systematic KM approaches and realistic understanding of technological capabilities, endangered languages can begin transitioning from passive documentation to active participation in digital ecosystems.

8. Limitations

All ASR evaluations are zero-shot; fine-tuning experiments could not be performed due to limited data and the absence of standard train/dev/test splits. The absence of a standardized Aromanian orthography means ground-truth transcriptions reflect individual transcriber conventions, potentially inflating error rates. Only Whisper models were evaluated; other architectures (XLS-R, Omnilingual ASR) may yield complementary insights. The aggregate WER and CER metrics do not reveal which specific linguistic phenomena contribute most to errors; qualitative error analysis is left for future work. Additionally, the survey data informing the KM framework was collected in 2017 and examines general KM practices rather than language-specific behaviours, potentially limiting direct applicability to the language preservation context. The geographic and demographic scope of the survey may not fully represent global Aromanian communities, particularly diaspora communities with different technological access and cultural preservation priorities.

9. Data Availability

The oral speech data used for the ASR evaluation in this study are publicly available via the Open Science Framework (OSF) repository at the following link: `https://osf.io/z2n7r/overview`.

The repository contains the processed audio segments and corresponding aligned transcriptions used for evaluation, ensuring full reproducibility of the reported results

Ethics Statement

All speech data were collected with informed consent. The survey was conducted with voluntary participation. The goal is to support linguistic documentation and technological inclusion, not deployment in real-world applications. The reported ASR models exhibit high error rates and should not be used for practical applications involving Aromanian speakers.

Acknowledgments

First and foremost, I am humbled and give my deepest respect to Ms. Ljubica Georgievska for her dedicated work on keeping the Aromanian language alive and contemporary.

10. Bibliographical References

Tanel Alumäe, Jiaming Kong, and Daniil Robnikov. 2023. Dialect Adaptation and Data Augmentation for Low-Resource ASR: TalTech Systems for the MADASR 2023 Challenge.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale.

Stamatis Beis and Fanis Dasoulas. 2017. Proposal for the writing system of the Vlach language. *Glossologia*, 25:51–69.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Steven Bird and Gary Simons. 2002. Seven dimensions of portability for language documentation and description.

Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, and Barbara Plank. 2025. A Multi-Dialectal Dataset for German Dialect ASR and Dialect-to-Standard Speech Translation. In *Interspeech 2025*, pages 913–917. ISCA.

Theodor Capidan. 1932. *Aromânii: Dialectul aromân*. Imprimeria Naţională, Bucharest.

David Crystal. 2000. *Language Death*. Cambridge University Press.

Tiberius Cunia. 2000. On the standardization of the Aromanian system of writing. *The Farsarotul*.

M. Evans, K. Dalkir, and C. Bidian. 2014. A holistic view of the knowledge life cycle: The knowledge management cycle (KMC) model. *Electronic Journal of Knowledge Management*, 12(2):85–97.

Victor A. Friedman. 2012. Aromanian in the Balkan linguistic league: A comparative approach. volume 14, pages 51–66.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Thede Kahl. 2008. Does the Aromanian have a chance of survival? Some thoughts about the loss of language and language preservation. *The Romance Balkans*, pages 123–140.

Marjan Marković. 2007. *Govorot na Aromancite Frašeroti vo Ohridsko-Struškiot region (vo Balkanski kontekst)*. MANU, Skopje. Doctoral dissertation defended in 2000.

Angelika Mittelmann, Gabriele Vollmar, et al. 2022. Wissensmanagement-kompetenzkatalog. Version 2.0. Gesellschaft für Wissensmanagement e.V. (GfWM).

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. UNESCO Publishing, Paris.

Seham Nasr, Rehab Duwairi, and Muhannad Quwaider. 2023. End-to-end speech recognition for arabic dialects. *Arabian Journal for Science and Engineering*, 48.

Božidar Nastev. 1988. *Aromanski studii, prilozi kon balkanistikata*. MANU, Skopje.

Omnilingual ASR Team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. 2025. Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages.

Marija Pendevska. 2019. *The Impact of Knowledge Management on Innovation in Enterprises in the Republic of Macedonia*. Ph.D. thesis, Faculty of Economics, Ss. Cyril and Methodius University in Skopje, Skopje.

Marija Pendevska, Branislav Gerazov, and Branko Prlja. 2024. Digital preservation of Aromanian through a knowledge management framework: The first digital riddle generator. In *Proceedings of the 2nd International Conference Dedicated to 100 Years Prof. Dr. Božidar Nastev*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision.

Georg Rehm and Hans Uszkoreit. 2012. Language technology 2012: Current state and opportunities. In *The META-NET Strategic Research Agenda for Multilingual Europe 2020*, pages 27–31.

Clement Sicard, Kajetan Pyszkowski, and Victor Gillioz. 2023. Spaiche: Extending State-of-the-Art ASR Models to Swiss German Dialects.

Naoki Takahashi, Shogo Miwa, Yuta Kamiya, Takumi Toyama, Raufun Nahar, and Atsuhiko Kai. 2024. Comparison of Large Pre-trained Models and Adaptation Methods for Japanese Dialects ASR. In *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*, pages 811–814.

Chara Tsoukala, Stavros Bompolas, Antigoni Margariti, Konstantina Panagiotou, Maria Elisavet

Plaiti, Nefeli Tzanakaki, Petros Karatsareas, Angela Ralli, Antonios Anastasopoulos, and Stella Markantonatou. 2026. Extending ASR evaluation resources for Modern Greek dialects. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 210–222, Rabat, Morocco. Association for Computational Linguistics.

Socrates Vakirtzian, Vivian Stamou, Yannis Kazos, and Stella Markantonatou. 2025. Dialectal treebanks and their relation with the standard variety: The case of East Cretan and Standard Modern Greek. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 776–784, Tallinn, Estonia. University of Tartu Library.

Gustav Weigand. 1895. *Die Aromunen: Ethnographisch-Philologisch-Historische Untersuchung über das Volk der sogenannten Makedo-Romanen oder Zinzaren*. J.A. Barth, Leipzig.

11. Language Resource References

Boersma, Paul and Weenink, David. 2024. *Praat: Doing Phonetics by Computer*. University of Amsterdam. PID https://www.praat.org/.

CTC Forced Aligner Team. 2024. *CTC Forced Aligner*. Open Source. PID https://github.com/huggingface/ctc-forced-aligner.

Hermjakob, Ulf and May, Jonathan and Knight, Kevin. 2018. *Out-of-the-Box Universal Romanization Tool Uroman*. USC Information Sciences Institute. PID https://github.com/isi-nlp/uroman.

Mozilla Foundation. 2025. *Mozilla Common Voice*. Mozilla Foundation. PID https://commonvoice.mozilla.org/.