Can You Hear Naples? Building and Benchmarking a Neapolitan Speech Corpus

Michael Cacioli michael.cacioli2008@gmail.com
Gilmour Academy
Algoverse AI Research

Liam Eggleston liameeggleston@gmail.com
Algoverse AI Research

Jatin Sarabu jatinsarabu13@gmail.com
Algoverse AI Research

Kevin Zhu kevin@algoverseacademy.com
Algoverse AI Research

Editors: Tatsuya Komatsu, Keisuke Imoto, Xiaoxue Gao, Nobutaka Ono, Nancy F. Chen

Abstract

This paper presents the creation and analysis of the first spoken corpus for Neapolitan, a richly historic but under-resourced Romance dialect of Southern Italy. Despite its cultural importance, Neapolitan has been largely omitted from computational resources, limiting both dialectological research and the development of equitable speech technologies. We address this gap by creating the first structured spoken resource for Neapolitan, enabling systematic evaluation of dialectal ASR performance. Each clip was manually transcribed in orthographic Neapolitan and automatically aligned using OpenAI's Whisper API, configured for standard Italian. To figure out how well Whisper transcribed the spoken Neapolitan sentences, we checked the outputs against the correct human-written texts using a few different methods. Specifically, we looked at how often the words matched (BLEU), how different the transcriptions were overall (normalized Levenshtein distance), and how closely the sets of words lined up (Jaccard similarity). We also used Word Error Rate (WER), but to make it easier to interpret, we converted it to similarity by subtracting from one (1–WER). A higher value means the transcription was more accurate. On average, this similarity measure came out very low, around 0.1306 ($\sigma = 0.1654$), meaning roughly 87 percent of the words were transcribed incorrectly. The other evaluation measures told the same story: normalized Levenshtein similarity averaged around 0.6360, and Jaccard similarity was just 0.1078. Today's auto- matic speech recognition tools have significant trouble in handling dialects like Neapolitan. This paper makes three crucial steps: (1) developed an easy-to-follow process anyone can use to build similar datasets for other dialects, (2) released the first openly accessible Neapolitan speech corpus, and (3) demonstrated just how critical it is to build ASR systems specifically trained on dialects, supporting not just computational linguistic research but also efforts to preserve these unique languages.

1. Introduction

The development of spoken language corpora is crucial for advancing computational linguistic tools and for preserving linguistic diversity, especially when it comes to under-resourced

Figure 1: Geographic distribution of contemporary Neapolitan speakers across southern Italy, with concentration in Naples and surrounding provinces.

languages and dialects. This is highlighted by various studies Godard et al. (2018); Ćavar et al. (2016); Yang et al. (2025c,e) that emphasize the need for such resources. Among these languages and dialects, Neapolitan stands out as a prominent yet linguistically underserved dialect.

Neapolitan (Napoletano) is an Italo-Romance variety spoken in and around Naples whose divergence from Standard Italian is systematic, not merely accentual, spanning phonology and vocabulary. Because of phenomena such as frequent final-vowel deletion and dialect-specific lexemes, Italian-trained ASR models often fail to transfer cleanly, an issue we analyze in detail in Section 6.1. The dialect presents a unique set of challenges and op-

Figure 2: Three-stage pipeline for building and assessing the Neapolitan spoken corpus: (1) domain-specific text selection, (2) native speaker recording and formatting, and (3) automatic speech recognition (ASR) evaluation using Whisper transcription with quantitative analysis.

portunities for linguistic research, largely due to its rich historical and cultural significance. However, despite this significance, Neapolitan remains underrepresented in computational studies and corpus development efforts.

Our project aims to fill this gap by constructing and analyzing a spoken corpus of Neapolitan. This initiative is intended to facilitate further linguistic and computational exploration of the dialect. We utilize contemporary methodologies, drawing from recent literature on corpus creation and dialect identification Yang et al. (2025d,b), as well as machine learning approaches applied to spoken language Ardila et al. (2020). The project is inspired by foundational efforts like the VoLIP corpus. Additionally, it incorporates recent advances such as self-supervised learning techniques for dialect classification Alvarez et al. (2025), and developments in Italian audio-to-text transcription models. Our research aims to capture the linguistic richness of Neapolitan by implementing systematic data collection, precise annotation, and innovative analytical methods Hamlaoui et al. (2018); Yang et al. (2025a).

In this paper, we detail the process of compiling the corpus. We outline the methodologies employed for aligning and annotating audio recordings, making use of advanced Italian audio-to-text models. Furthermore, we demonstrate the application of interpretable classifiers in identifying distinguishing lexical and phonetic features of the Neapolitan dialect. Our contributions to this field not only increase the resources available for Italian dialectology but also provide a replicable framework. This framework can assist researchers working with similarly under-documented dialects on a global scale.

The importance of constructing such a corpus cannot be overstated, as it serves as a foundational resource that supports a wide array of linguistic studies. Through it, researchers can delve deeper into the syntactic, semantic, and phonological aspects of the Neapolitan dialect, uncovering patterns and structures that are unique to it. By doing so, they also contribute to the broader field of dialectology and language preservation, ensuring that these linguistic treasures are not lost to time.

Through our efforts, we aim not only to preserve the Neapolitan dialect but also to highlight its value and influence in the broader landscape of Italian culture and linguistics. This work stands as a testament to the power of collaborative and interdisciplinary research, drawing on expertise from linguists, computer scientists, and cultural historians to create a resource that is both academically rigorous and practically useful.

2. Related Work

Research on spoken corpora for under-resourced languages emphasizes the need for systematic documentation frameworks. Foundational Italian efforts, such as Voghera and Cutugno (2006), established standardized approaches for studying spoken language, offering methodological guidance applicable to Neapolitan. The VoLIP corpus Alfano et al. (2014) expanded this work by linking audio to orthographic transcriptions and capturing structured diaphasic and diamesic variation—an approach that informs the organization of new dialect corpora.

In the realm of dialect identification, the SUKI team's approach in the VarDial Evaluation Campaign 2022 demonstrated effective methods for distinguishing between closely related language varieties, including Italian dialects Aepli et al. (2022). Their findings highlight the potential of machine learning techniques in handling dialectal variations. This exemplifies how advanced computational methods can address the complexity of distinguishing subtle linguistic features within a closely knit linguistic family.

Bentum et al. (2024) focused on the creation and automatic alignment of a historical Dutch dialect speech corpus. Their methodologies for aligning audio recordings with transcriptions can inform similar efforts for Neapolitan, especially when dealing with non-standard dialectal variations. The application of these alignment techniques ensures accurate representation and analysis of recorded speech, crucial for maintaining the integrity of dialect studies.

Furthermore, La Quatra et al. (2024) employed self-supervised learning models to analyze and classify Italian regional language varieties based on speech data. Such models offer a promising avenue for automated analysis, enabling large-scale examination of speech data with minimal manual intervention. Xie et al. (2024) presented methods for identifying distinguishing lexical features of dialects using interpretable classifiers. Applying such techniques to Neapolitan can enhance the understanding of its unique lexical characteristics.

3. Dataset Collection

Source	Clips	Avg. Duration (s)	Total Duration (s)
Plays	63	4.59	289.34
Poetry	49	4.01	196.61
Blogs	29	8.89	257.77
Total	**141**	**5.27**	**743.72**

Table 1: Distribution of Neapolitan speech clips by source Neapolitan text domain.

The text for this dataset was collected and compiled manually under the guidance of a native Neapolitan speaker. These specific domains were selected to represent a wide array of Neapolitan style, structure, and vocabulary, reflecting both traditional and contemporary uses of Neapolitan. Text was sourced from publicly available Neapolitan literature and blogs. The selection followed three criteria: (1) each source needed to contain linguistically authentic Neapolitan rather than Italian-influenced orthographies; (2) the text needed to include dialect-specific phonological features such as final-vowel deletion, consonant assimilation, and regionally marked lexemes; and (3) a native speaker verified that each candidate sentence reflected natural, culturally representative usage. These constraints ensured that the dataset captured both stylistic breadth and genuine dialectal structure. Plays and poetry allowed the analysis of expressive and culturally-rich language, while blogs provided informal, community-driven language.

In developing this dataset, careful consideration was given to the selection of source materials, ensuring that a broad spectrum of the Neapolitan language was represented. This was crucial in capturing the essence of Neapolitan expressions, idioms, and the subtle nuances that form the heart of its linguistic identity. By including a variety of text types, from formal literature to casual blogs, the dataset offers a comprehensive view of how Neapolitan can be used in different contexts, each bringing its own flavor and depth to the language.

This dataset consisted of recorded audio clips, all done by a native Neapolitan speaker, ensuring authentic intonation and pronunciation. Recordings were made in a quiet, indoor environment to minimize background noise, using a consistent speaking pace and volume. All clips were recorded on an iPhone 13 using Apple's built-in Voice Memos application. The speaker was a male native Neapolitan speaker with the Central Neapolitan phonological profile typical of the Naples metropolitan region. The recordings phonetically exhibit hallmark Neapolitan features, including systematic apocope and lexical items unattested in Standard Italian. These characteristics make the dataset suitable for evaluating dialect-sensitive ASR behavior. The resulting audio files were saved in the M4A (MPEG-4 Audio) format, which balances high audio quality with efficient file size. Each clip represented one spoken Neapolitan sentence, making the dataset suitable for alignment with the corresponding text in downstream computational tasks.

The decision to use the iPhone 13 and its Voice Memos application was driven by the need to maintain consistency in audio quality across all recordings. This choice ensured that the audio captured was clear and free from distortions, a critical factor for anyone relying on the dataset for learning or analysis purposes. Furthermore, by saving the files in the M4A format, the audio retains its clarity while being manageable in terms of storage, making it easier for users to download and utilize the data in various computational applications.

4. Whisper API

To evaluate the transcription capabilities of conventional speech-to-text systems in underrepresented languages, we used OpenAI's Whisper API. Whisper is a general-purpose automatic speech recognition (ASR) system trained on a large multilingual and multitask supervised dataset. It supports numerous languages explicitly; however, Neapolitan is not

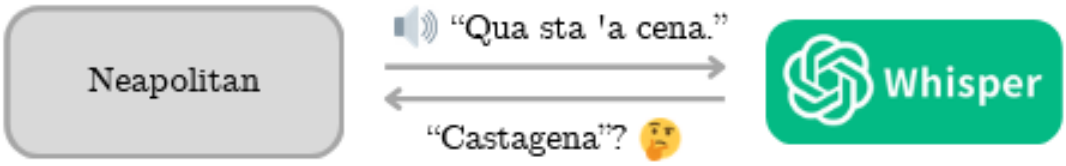

Figure 3: Illustration of a transcription mismatch between a spoken Neapolitan sentence ("Qua sta 'a cena.") and OpenAI's Whisper automatic speech recognition (ASR) system, which incorrectly transcribes it as "Castagena."

among the supported options. Consequently, all transcriptions were performed using the `language="it"` parameter, which corresponds to Standard Italian.

This choice reflects the common workaround adopted when dealing with dialects or minority languages that lack dedicated automatic speech recognition (ASR) models. Given the linguistic proximity between Neapolitan and Italian, using the Italian model offers a practical, if imperfect, proxy. However, as our results show, this approach leads to significant inaccuracies, demonstrating the limitations of Whisper in handling low-resource, non-standard language varieties like Neapolitan.

These transcription mismatches often stem from subtle phonetic and lexical divergences that the Italian model was never designed to accommodate. For example, Neapolitan contains a wide array of vowel reductions, consonant elisions, and regionally variable expressions that differ considerably from Standard Italian. Even when a Neapolitan word closely resembles its Italian counterpart, the pronunciation and stress pattern may cause Whisper's Italian model to infer a completely unrelated word. This not only affects intelligibility but also disrupts sentence meaning, leading to outputs that are syntactically correct in Italian but semantically nonsensical in Neapolitan context.

In practice, this creates major challenges for researchers and developers attempting to build ASR systems for dialect-rich regions. The absence of training data specifically tailored to Neapolitan means that even high-performing multilingual models like Whisper cannot generalize effectively across these linguistic boundaries. The resulting transcripts may appear superficially accurate yet conceal numerous underlying errors that distort the cultural and linguistic authenticity of the speech. Moreover, these inaccuracies highlight an ongoing issue in language technology: the systematic underrepresentation of dialects and minority varieties in large-scale training corpora.

From a practical standpoint, using the Italian model remains a necessary compromise. It allows for an initial, rough transcription that can later be manually corrected or used as a baseline for developing more accurate dialect-aware systems. This process, however, is time-consuming and underscores the urgent need for dedicated Neapolitan data resources. Building such corpora, as this project demonstrates, is not only vital for technical progress but also for preserving the linguistic identity of communities whose speech patterns have historically been marginalized by standardization. We emphasize that Whisper was not

used to create or curate the corpus itself; it was used only after corpus construction as an external evaluation tool to characterize cross-dialect ASR performance.

5. Results

We evaluated Whisper's transcription performance on a set of 141 spoken Neapolitan audio clips, each aligned with a reference transcription created by a native speaker. Because Whisper does not support Neapolitan directly, we used the Italian setting, which we assumed would perform best among the available options.

To assess transcription quality, we relied on four common metrics. First, we used Word Error Rate (WER). To simplify interpretation, we report both the standard Word Error Rate (WER) and its complement (1 – WER). The mean WER was 86.94 % ($\sigma = 16.54$ %), corresponding to 1 – WER = 0.1306 ($\sigma = 0.1654$). In other words, roughly 87 % of words were transcribed incorrectly on average. In other words, Whisper rarely had success with transcribing the sentences accurately, although a few outliers scored significantly higher.

Similarly to WER, BLEU scores, which reflect phrase-level overlap, were very low. The average BLEU was 0.0436 ($\sigma = 0.0961$), and most clips hovered near zero. There were occasional examples with higher BLEU, often when the sentence resembled standard Italian more than usual.

We also considered normalized Levenshtein similarity, a character-level metric that captures how many small edits would be needed to match the transcription to the reference. This score was higher on average—0.6360 ($\sigma = 0.1375$)—suggesting that while the output was usually wrong, it often sounded close

Finally, Jaccard similarity, which compares the sets of unique words in each sentence, showed the lowest performance overall. The average was 0.1078 ($\sigma = 0.1294$), reinforcing the idea that most predictions had little actual word overlap with the reference.

It's clear that Whisper struggles to generalize to Neapolitan. Even when using a related language setting, the model failed to produce reliable transcriptions across all four metrics. These findings point to the urgent need for speech recognition systems trained specifically on dialectal and low-resource varieties, especially when high fidelity is required for research or preservation work. Because the corpus contains a single speaker, these results should be interpreted as a speaker-specific diagnostic rather than a population-level estimate of Neapolitan ASR difficulty. We therefore refrain from generalizing beyond this dataset and frame our results as evidence of a systematic trend that warrants larger multi-speaker, multi-dialect follow-up work.

6. Analysis and Discussion

Our evaluation of Whisper API's Italian model on Neapolitan dialect audio reveals significant limitations in transcription accuracy. Although Whisper is trained on standard Italian, it struggled to correctly capture many dialect-specific words and pronunciations unique to Neapolitan. The transcriptions often contained numerous errors, including frequent misrecognitions, phoneme substitutions, and incorrect word insertions or omissions. These issues indicate that Whisper's model does not sufficiently generalize to Neapolitan.

Clip	Domain	Verbatim
002.m4a	Play	E chesto capisce tu: 'e denare!
003.m4a	Play	E cu' 'e denare t'he accattato tutto chello ca he voluto!
004.m4a	Play	Ma Filumena Marturano ha fatto correre essa a te! E currive senza ca te n'addunave.
005.m4a	Play	E ancora he 'a correre, ancora he 'a iettà 'o sango a capi comme se campa e se prucede 'a galantomo!

Table 2: Representative example utterances from the Neapolitan-Spoken-Corpus, including clip identifier, source domain, and verbatim transcription.

The lack of a Neapolitan-specific Whisper API model is evidence of a lack of computational support for vulnerable languages.

In analyzing the transcriptions, we observed that even when the spoken Neapolitan sentences shared clear lexical or phonetic similarities with standard Italian, Whisper's predictions often defaulted to Italian cognates or unrelated Italian words. This pattern highlights how large-scale models may rely too heavily on language priors from dominant varieties, effectively erasing or distorting the linguistic identity of minority dialects. For example, vowel reduction and final consonant deletion—core phonological features of Neapolitan—frequently led the model to hallucinate entire Italian words that were never spoken. Such errors make the resulting text appear superficially fluent but semantically inaccurate, creating a misleading impression of correctness.

This misalignment between perceived and actual accuracy reflects a broader challenge within multilingual automatic speech recognition: the assumption that related languages are mutually intelligible at the computational level. While Italian and Neapolitan share etymological roots, their phonotactics, stress patterns, and vocabulary differ enough that using Italian-trained models introduces systematic bias. In practice, this means that even when the model produced partial matches—such as correctly identifying one or two words in a sentence—the overall message often became unintelligible. This failure is particularly concerning in applied or archival contexts, where accurate transcription is critical for preserving oral material and linguistic nuance.

This performance gap highlights the inherent challenges faced by automatic speech recognition systems when processing regional dialects with distinctive phonetic and lexical features. Our findings emphasize the importance of developing automatic speech recognition (ASR) models trained on or adapted for dialectal speech to improve transcription quality and usability.

6.1. Corpus Coverage and Linguistic Divergence

A single-speaker, domain-diverse corpus is sufficient to reveal systematic ASR failure while preserving a controlled benchmark for replication. As detailed in Section 3, the Neapolitan-Spoken-Corpus (NSC) consists of 141 sentence-level recordings across three domains—traditional plays, regional poetry, and community blogs—captured under identical acoustic conditions on an iPhone 13. This configuration ensures consistent signal quality while maintaining

Neapolitan (NSC)	Standard Italian
Si ero ancora giovan, avesse scalat chella mundagn'.	*Se fossi stato giovane, avrei scalato quella montagna.*
Aggio juto a Rroma l'at ajere.	*Sono andato a Roma ieri.*
Si fusse stato in Norvegia, avesse potuto veré l'laurora boreal.	*Se fossi stato in Norvegia, avrei potuto vedere l'aurora boreale.*

Table 3: Parallel Neapolitan–Italian examples illustrating systematic divergence in the two languages.

stylistic breadth within a single voice. Each recording contributes to a dataset designed for linguistic transparency and replicability, allowing researchers to analyze ASR behavior without the noise introduced by heterogeneous recording setups.

The linguistic divergence between Neapolitan and Italian provides a clear explanation for Whisper's poor performance. Neapolitan systematically departs from Standard Italian in phonology and morphology, exhibiting features such as final-vowel deletion (e.g., *mangia* $\rightarrow$ *mangi'*), consonant assimilation (*vado a* $\rightarrow$ *vagg' a*), and characteristic stress shifts that alter phonemic boundaries. These differences interact with the Italian model's learned priors, causing predictable substitution of Italian cognates and the misalignment of otherwise accurate acoustic tokens. In practice, these mismatches manifest as fluent but semantically inaccurate transcriptions that obscure the dialect's linguistic identity.

The NSC thus fulfills a dual role: (1) it demonstrates how even high-capacity multilingual ASR systems falter when encountering dialectal input, and (2) it establishes a reproducible resource that future research can extend to multi-speaker, multi-domain settings. This design choice reflects a common trade-off in early corpus construction: prioritizing reproducibility over speaker diversity. While this limits population-level generalization, it enables precise diagnostic evaluation of model bias. Future expansions will incorporate balanced demographic and situational variation.

6.2. Limitations

As noted in Section 6.1, the NSC's design centers on a single speaker recorded under uniform conditions. While this approach ensures acoustic consistency, it inherently limits dialectal and demographic variation. The absence of multiple speakers reduces generalizability for population-level inference, meaning all quantitative findings should be viewed as speaker-specific diagnostics rather than dialect-wide estimates. Nevertheless, this design facilitates controlled testing of cross-dialectal transfer—a critical precursor to multi-speaker expansion. Beyond technical considerations, these constraints also shape how the dataset can be interpreted within broader sociolinguistic contexts. The limited speaker representation means that subtle intra-dialectal differences—such as generational variation, prosodic rhythm, and localized intonation contours—remain unexplored. These features are integral to understanding how Neapolitan functions as both a linguistic and cultural system. By highlighting their absence, this study underscores the importance of integrating sociopho-

netic diversity in future work, ensuring that computational modeling of dialects accounts not only for lexical or acoustic accuracy but also for the expressive and identity-bearing aspects of speech.

Additionally, Neapolitan lacks a fully standardized modern written form. This is due to the language's suppression following the unification of Italy in the 19th century. During which, standard Italian was forced upon all regions of Italy, discouraging the use of regional dialects such as Neapolitan. This limited the variation and availability of high-quality sources that could be used in the creation of this dataset. The inconsistent orthography complicated the transcription process, as different linguistic sources use different conventions for representing certain sounds, stress patterns, and elisions. This inconsistency can introduce small but meaningful differences across samples, affecting how ASR evaluation metrics interpret correctness. While this variability reflects the authentic nature of Neapolitan as a living dialect, it also makes it difficult to establish a "ground truth" transcription standard suitable for machine learning evaluation.

Beyond speaker uniformity, two practical constraints limit generalization: the modest dataset size (141 clips, $\approx$ 12 min total) and the absence of spontaneous conversational speech. These are typical of early-stage dialect corpora and provide controlled, replicable baselines. Future expansions will incorporate natural dialogue, multi-speaker variation, and richer acoustic contexts to improve coverage.

We emphasize two constraints that limit generalization: (1) the corpus contains 141 clips from a single native speaker, and therefore our results describe ASR behavior for this speaker and these domains rather than cross-speaker population-level performance; and (2) Neapolitan orthography is not standardized.

Another key limitation arises from the reliance on Whisper's Italian model as a baseline for evaluation. Since Whisper was not trained on Neapolitan, its transcription errors may reflect both model-specific weaknesses and linguistic mismatches. Consequently, the dataset's evaluation results, while revealing, may underestimate the potential performance of future dialect-specific models. This introduces an interpretive limitation—distinguishing between model failure and dataset difficulty is challenging without comparative baselines across multiple architectures. In this way, the current findings highlight the need for future work to expand both dataset diversity and comparative model evaluation.

Beyond dataset-specific factors, these findings also expose a broader methodological issue in speech technology evaluation. ASR models trained on standardized languages often inherit cultural and linguistic assumptions that bias performance across dialects. For low-resource settings like Neapolitan, this means that even technical metrics such as WER can reflect sociolinguistic marginalization as much as algorithmic accuracy. Addressing this challenge requires designing benchmarks that explicitly account for linguistic diversity, not just acoustic variability. Such benchmarks would enable fairer comparisons across models and encourage the development of evaluation frameworks that better capture dialectal authenticity and communicative intent.

6.3. Qualitative Error Analysis

A closer inspection of the Whisper outputs reveals four recurrent error types: phonetic hallucination, lexical substitution, stress misplacement, and syntactic bias. In phonetic hal-

lucination, the model "hears" Italian cognates where none exist. This is a major issue. For instance, the Neapolitan *'sta 'a cena'* ("here's the dinner") becomes *"Castagena."* Lexical substitutions occur when region-specific words are replaced by semantically unrelated Italian ones. Stress misplacement often turns partially correct phoneme recognition into full semantic loss.

Syntactic bias emerges when Italian word order is forced onto Neapolitan clauses. Together, these patterns show that the model's training bias is linguistic rather than acoustic, reinforcing the need for dialect-specific modeling. This also highlights how cross-linguistic interference can lead ASR systems to prioritize dominant language structures over local syntactic norms. In practice, biases like these not only distort the intended meaning, but also erase markers of dialectal identity. These are elements that are central to linguistic authenticity. Addressing these issues will require training pipelines that explicitly represent dialectal syntax rather than assuming transfer from standardized varieties.

6.4. Ethical Considerations

All participants involved in the dataset creation, including the speaker and annotators, gave informed consent. No personal or sensitive content was included in the data. The dataset is intended solely for academic research and will be publicly released under a Creative Commons Attribution-NonCommercial (CC BY-NC 4.0) license to ensure responsible, non-commercial use while requiring proper attribution. Informed consent was obtained in written form, and this work was reviewed and deemed exempt from formal IRB approval under institutional guidelines for minimal-risk linguistic data collection.

7. Native Speaker Perspective on Language Revitalization

The issue of the endangerment of Neapolitan is one that's hugely affecting the rich culture of Naples, Italy, and beyond. To learn from members of the community, we consulted a native Neapolitan speaker. This speaker confirmed the severity of the issue.

The native speaker we spoke with cited experiences as a child when "even the speaking of the Neapolitan language, let alone the writing of it, would land you detention, if not worse." This stands out as a main reason why the language now faces endangerment. This lack of writing in the language led to the loss of an agreed-upon Neapolitan writing system. While the phonetics and alphabet of Neapolitan remain unanimous, the exact orthography varies.

Our speaker acknowledged the use of our audio dataset as being, "extremely innovative." Despite the inconsistencies in the writing of Neapolitan, the way it is spoken has remained consistently agreed upon. It's for this reason that our Neapolitan audio corpus is a major advancement. The lack of an agreed-upon modern system of Neapolitan writing following Italy's unification leaves oral communication as the last way of carrying on the cultural significance of this language. It was for this reason that we chose to prioritize the authentic speech of a native Neapolitan speaker.

8. Conclusion and Future Work

We present a curated Neapolitan speech dataset[1] encompassing diverse Neapolitan registers—literary, poetic, and informal—captured under consistent acoustic conditions. The dataset captures diverse linguistic registers—from literary to informal—and is designed to support research in low-resource automatic speech recognition (ASR), dialect modeling, and language preservation.

Future work will expand the dataset along three dimensions: (1) increasing speaker diversity to include variation in age, gender, and regional accent; (2) broadening domain coverage to include spontaneous conversation and oral storytelling; and (3) providing high-quality transcriptions, phonetic alignments, and optional code-switching annotations. The results from this experiment underscore the need for future work focused on specialized dialect-aware automatic speech recognition (ASR) development.

By releasing this resource, we aim to encourage further computational work on Neapolitan and similar endangered or marginalized language varieties.

References

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, Gyeongju, Republic of Korea, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.vardial-1.13.

Iolanda Alfano, Francesco Cutugno, Aurelio De Rosa, Claudio Iacobini, Renata Savy, and Miriam Voghera. Volip: A corpus of spoken italian and a virtuous example of reuse of linguistic resources. In *Proceedings of the Olomouc Linguistics Colloquium (OLINCO)*, pages 23–33, 2014.

Jesus Alvarez, Daua Karajeanes, Ashley Prado, John Ruttan, Ivory Yang, Sean O'Brien, Vasu Sharma, and Kevin Zhu. Advancing uto-aztecan language technologies: A case study on the endangered comanche language. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 27–37, 2025.

Rosana Ardila, Megan Branson, Kelly Davis, Mark Kohler, Reuben Meyer, Michael Henretty, Michael Morais, Lindsey Saunders, Francis Tyers, and Peter Warden. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4218–4222, Marseille, France, 2020. European Language Resources Association (ELRA).

Martijn Bentum, Antal van den Bosch, and Martijn van der Meulen. Corpus creation and automatic alignment of historical Dutch dialect speech. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italy, 2024. European Language Resources Association (ELRA). URL https://aclanthology.org/2024.lrec-main.357.

1. https://huggingface.co/datasets/anonymous-nsc-author/Neapolitan-Spoken-Corpus/tree/main

Pierre Godard, Gilles Adda, Martine Adda-Decker, Hélène Bonneau-Maynard, Germain Nyangi Kouarata, Gabriel Mba, Markus Mueller, Annie Rialland, Laurent Besacier, and François Yvon. A very low resource language speech corpus for computational language documentation experiments. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1286–1293, Miyazaki, Japan, 2018. European Language Resources Association (ELRA).

Fatima Hamlaoui, Emmanuel-Moselly Makasso, Markus Mueller, Jonas Engelmann, Gilles Adda, Alex Waibel, and Sebastian Stüker. Bulbasaa: A bilingual basaá–french speech corpus for the evaluation of language documentation tools. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 2018. European Language Resources Association (ELRA).

Moreno La Quatra, Alessio Cignarella, and Sara Tonelli. Speech analysis of language varieties in Italy. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italy, 2024. European Language Resources Association (ELRA). URL https://aclanthology.org/2024.lrec-main.1317.

Miriam Voghera and Francesco Cutugno. An observatory on spoken italian linguistic resources and descriptive standards. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*. European Language Resources Association (ELRA), 2006.

Roy Xie, Orevaoghene Ahia, Yulia Tsvetkov, and Antonios Anastasopoulos. Extracting lexical features from dialects via interpretable dialect classifiers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 54–69, Mexico City, Mexico, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.naacl-short.5.

Ivory Yang, Xiaobo Guo, Yuxin Wang, Hefan Zhang, Yaning Jia, William Dinauer, and Soroush Vosoughi. Recontextualizing revitalization: A mixed media approach to reviving the nüshu language. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025a.

Ivory Yang, Weicheng Ma, Carlos Guerrero Alvarez, William Dinauer, and Soroush Vosoughi. What is it? towards a generalizable native american language identification system. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 105–111, 2025b.

Ivory Yang, Weicheng Ma, and Soroush Vosoughi. Nüshurescue: Reviving the endangered nüshu language with ai. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7020–7034, 2025c.

Ivory Yang, Weicheng Ma, Chunhui Zhang, and Soroush Vosoughi. Is it Navajo? accurate language detection for endangered athabaskan languages. In *Proceedings of the 2025*

Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 277–284, 2025d. URL https://aclanthology.org/2025.naacl-short.24/.

Ivory Yang, Chunhui Zhang, Yuxin Wang, Zhongyu Ouyang, and Soroush Vosoughi. Visibility as survival: Generalizing nlp for native alaskan language identification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6965–6979, 2025e.

Damir Ćavar, Małgorzata E. Ćavar, and Hilaria Cruz. Endangered language documentation: Bootstrapping a chatino speech corpus, forced aligner, asr. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, 2016. European Language Resources Association (ELRA).

Data Availability and Citation Notice

The Neapolitan-Spoken-Corpus (NSC) is released under the Creative Commons Attribution–NonCommercial 4.0 International (CC BY-NC 4.0) license. Researchers may use the data for non-commercial research purposes with proper citation of this paper and dataset DOI. Future versions of the dataset will include multi-speaker extensions and spontaneous speech additions, following the same license and documentation format.

Appendix A: Extended Dataset Metadata

This appendix provides expanded documentation of the Neapolitan-Spoken-Corpus (NSC) beyond what is included in the main paper, offering researchers additional clarity regarding dataset design, phonological structure, and speaker characteristics.

A.1 Speaker Demographics

The corpus was recorded by a male native speaker of Neapolitan raised in central Naples. His speech reflects the Central Neapolitan variety typical of the metropolitan region.

A.2 Phonological Profile of Recordings

The recordings exhibit hallmark Neapolitan phonological properties:

- **Final-vowel deletion (apocope):** *mangia* → *mangi'*.
- **Consonant assimilation:** *vado a* → *vagg' a*.
- **Vowel reduction:** heavy centralization or deletion in unstressed syllables.
- **Dialect-specific lexemes:** *vastaso*, *scetato*, *zumpa*.

A.3 Recording Conditions

All clips were recorded:

- on an iPhone 13 (Voice Memos),
- indoors in a quiet room,
- with stable speaking pace and volume,
- saved as M4A (AAC, 256 kbps).

A.4 Orthographic Conventions

Because Neapolitan orthography is not standardized, we applied:

- apostrophes to mark consonant/vowel deletion,
- doubled consonants to signal phonetic length,
- minimal Italianized spellings unless historically attested.

Appendix B: Extended Corpus Construction Pipeline

B.1 Text Selection and Verification

Candidate sentences were filtered using three criteria:

1. linguistic authenticity (no Italianized spellings),
2. presence of dialectal phonology (e.g., apocope, assimilation),
3. native-speaker confirmation of naturalness.

B.2 Audio Recording Protocol

The speaker completed three sessions:

- domain-specific reading (plays, poetry, blogs),
- discarded clips with reading mistakes or noise,
- trimmed silence at boundaries.

B.3 File Normalization

Before Whisper evaluation:

- peak-normalized audio to –1.0 dB,
- trimmed silences at –35 dB,
- uniform naming: `NNN.m4a`.

B.4 Metadata Fields

Each clip has a JSON metadata entry with:

- clip ID,
- text domain,
- transcription,
- duration,

Appendix D: Whisper Error Cases and Error Typology

Example 1
Ref: Qua sta 'a cena.
Whisper: Castagena.
Type: phonetic hallucination via Italian prior.

Example 2
Ref: Aggio juto a Rroma ajere.
Whisper: Oggi sono andato a Roma ieri.
Type: syntactic overcorrection.

Example 3
Ref: Nun 'o voglio veré.
Whisper: Non lo voglio vedere.
Type: automatic Italianization.

Appendix E: Future Multi-Speaker Expansion Plan

Planned extensions include:

- balanced gender/age variation,
- spontaneous conversational and narrative speech,
- phonological minimal-pair sets,
- expanded domains such as oral histories,
- dialect-specific phonetic lexicon development.

Appendix F: Extended Figures and Tables

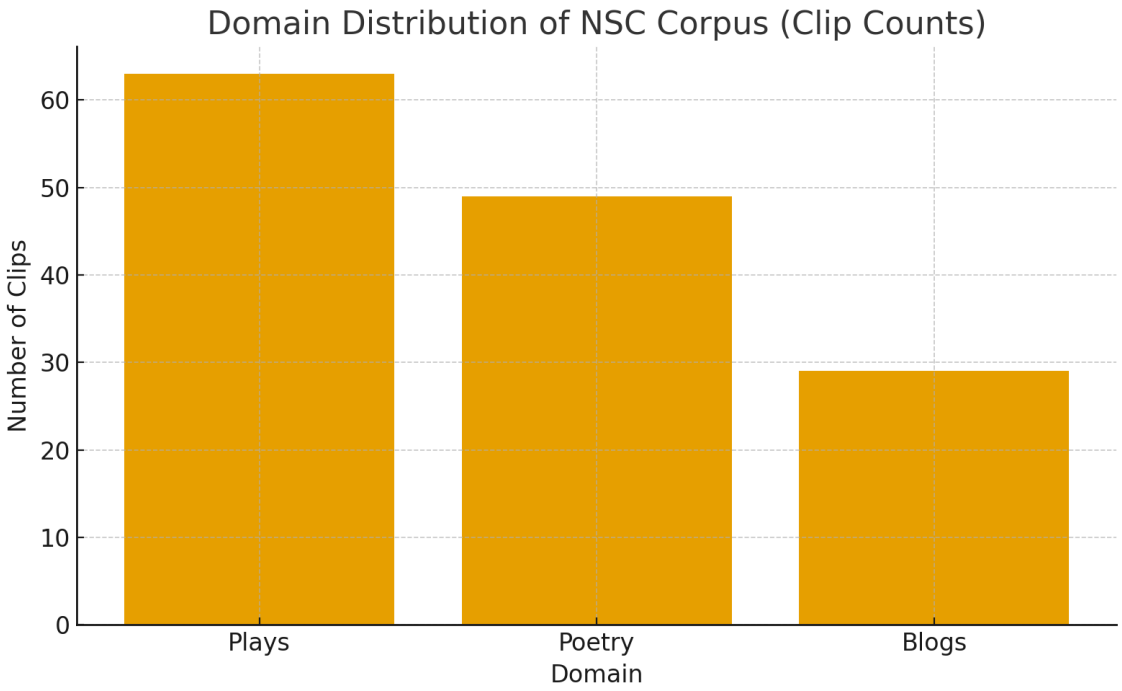

Figure 4: Distribution of text-source domains in the corpus.

Error Type	**Description / Example**
Phonetic Hallucination	Whisper inserts nonexistent Italian cognates; e.g., *'sta 'a cena → Castagena*.
Syntactic Overcorrection	Whisper rewrites to Standard Italian syntax; e.g., *Aggio juto → Oggi sono andato*.
Automatic Italianization	Substitution to nearest Italian equivalent; *veré → vedere*.
Stress Misplacement	Vowel reduction misinterpreted as deletion, altering meaning.

Table 4: Expanded Whisper error typology (full-size extended version).