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

Anonymous authors
Paper under double-blind review

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 compiling 141 sentence-level audio recordings across three domains—traditional plays, regional poetry, and community blogs—captured by a native speaker under controlled acoustic conditions. 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 (σ = 0.1654), meaning roughly 87% 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 automatic 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 essential for the advancement of computational linguistic tools and for preserving linguistic diversity, particularly for under-resourced languages and dialects (Godard et al. 2018; M. Ćavar, D. Ćavar, and Cruz 2016; Yang, Ma, and Vosoughi 2025). Among these, Neapolitan—a prominent yet linguistically underserved dialect spoken primarily in Naples and its surrounding areas in Italy—presents unique challenges and opportunities for linguistic research. Despite its historical and cultural significance, Neapolitan remains underrepresented in computational studies and corpus development.

Our project addresses this gap by constructing and analyzing a spoken corpus of Neapolitan, facilitating further linguistic and computational exploration. We leverage contemporary methodologies from recent literature on corpus creation, dialect identification (Yang, Ma, Zhang, et al. 2025; Yang, Ma, C. G. Alvarez, et al. 2025), and machine learning applied to spoken language (Ardila et al. 2020). Drawing inspiration from foundational efforts such as the VoLIP corpus, recent advances like self-supervised learning techniques for dialect classification (J. Alvarez et al. 2025), and developments in Italian audio-to-text transcription models, our research seeks to capture the linguistic richness of Neapolitan through systematic data collection, precise annotation, and innovative analytical methods (Hamlaoui et al. 2018).

Figure 1: 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.

In this paper, we describe the process of corpus compilation, outline methodologies for aligning and annotating audio recordings using advanced Italian audio-to-text models, and demonstrate the application of interpretable classifiers to identify distinguishing lexical and phonetic features of the Neapolitan dialect. Our contributions not only enrich resources available for Italian dialectology but also provide a replicable framework that can aid researchers working with similarly under-documented dialects globally.

2 Related Work

The development of spoken corpora for under-resourced languages and dialects has garnered increasing attention in recent years. Several initiatives have focused on the Italian linguistic landscape, providing valuable insights for the creation of a Neapolitan spoken corpus.

(Voghera and Cutugno 2006) introduced the national project ”Parlare italiano: osservatorio degli usi linguistici,” aiming to collect theoretical and applied results on spoken language and to implement standardized methods for its study. This work underscores the importance of structured approaches to spoken language documentation.

Building upon this, (Alfano et al. 2014) presented the VoLIP corpus, which associates audio signals with orthographic transcriptions from the LIP Corpus. Designed to represent diaphasic, diatopic, and diamesic variation, the corpus comprises approximately 60 hours of recordings, facilitating the compilation of a frequency lexicon for spoken Italian.

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.

(Bentum, Bosch, and Meulen 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.

Furthermore, (La Quatra, Cignarella, and Tonelli 2024) employed self-supervised learning models to analyze and classify Italian regional language varieties based on speech data. Their work provides insights into distinguishing features of regional dialects, which can aid in the analysis of Neapolitan speech patterns.

Lastly, (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)
Neapolitan plays	63	4.59	289.34
Neapolitan poetry	49	4.01	196.61
Neapolitan blogs	29	8.89	257.77
Total	**141**	**5.83**	**743.72**

Table 1: Distribution of Neapolitan speech clips by source 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. Plays and poetry allowed the analysis of expressive and culturally-rich language, while blogs provided informal, community-driven 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 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.

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

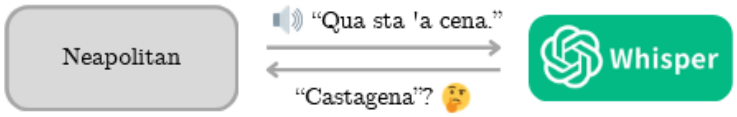

Figure 2: Illustration of a transcription mismatch between a spoken Neapolitan sentence ("Qua sta 'a cena.") and OpenAI's Whisper automatic speech recognition (ASR) system, which incorrectly transcribes it as "Castagena."

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.

Metric	Mean	Stdev	Min	Max
WER (1 - wer)	0.1306	0.1654	0.0000	0.9091
Levenshtein (normalized)	0.6360	0.1375	0.0870	0.9804
BLEU	0.0436	0.0961	0.0000	0.8932
Jaccard	0.1078	0.1294	0.0000	0.8333

Table 2: Similarity metrics comparing Whisper transcriptions to ground truth Neapolitan transcriptions. Lower values indicate worse 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 (`\texttt{language="it"}`), 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 1 minus WER (1–WER), where higher scores indicate better alignment with the reference. The mean 1–WER score was 0.1306 (σ = 0.1654), which implies that, on average, roughly 87% of words were incorrect. 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 (σ = 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 (σ = 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 (σ = 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.

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

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 Limitations

Despite the effectiveness and contributions of this Neapolitan audio dataset, several limitations affected both its scope and usability. The corpus currently contains only 141 sentence-level audio clips recorded by a single native speaker. This limits the dialectal variation within Neapolitan, such as age, gender, and regional accent differences.

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.

6.2 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.

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 Future Work and Conclusion

We present a curated Neapolitan speech dataset[1] consisting of 141 audio clips across three domains, all recorded by a native speaker. 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.

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

References

Aepli, Noëmi et al. (2022). "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: Association for Computational Linguistics. URL: https://aclanthology.org/2022.vardial-1.13.

Alfano, Iolanda et al. (May 2014). "VOLIP: a corpus of spoken Italian and a virtuous example of reuse of linguistic resources". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 3897–3901. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/906_Paper.pdf.

Alvarez, Jesus et al. (2025). "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)*, pp. 27–37.

Ardila, Rosana et al. (May 2020). "Common Voice: A Massively-Multilingual Speech Corpus". English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 4218–4222. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.520.

Bentum, Martijn, Antal van den Bosch, and Martijn van der Meulen (2024). "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: European Language Resources Association (ELRA). URL: https://aclanthology.org/2024.lrec-main.357.

Ćavar, Malgorzata, Damir Ćavar, and Hilaria Cruz (May 2016). "Endangered Language Documentation: Bootstrapping a Chatino Speech Corpus, Forced Aligner, ASR". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 4004–4011. URL: https://aclanthology.org/L16-1632.

Godard, Pierre et al. (May 2018). "A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://aclanthology.org/L18-1531.

Hamlaoui, Fatima et al. (May 2018). "BULBasaa: A Bilingual Basaa-French Speech Corpus for the Evaluation of Language Documentation Tools". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://aclanthology.org/L18-1533.

La Quatra, Moreno, Alessio Cignarella, and Sara Tonelli (2024). "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: European Language Resources Association (ELRA). URL: https://aclanthology.org/2024.lrec-main.1317.

Voghera, Miriam and Francesco Cutugno (May 2006). "An observatory on Spoken Italian linguistic resources and descriptive standards." In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2006/pdf/720_pdf.pdf.

Xie, Roy et al. (2024). "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)*. Mexico City, Mexico: Association for Computational Linguistics, pp. 54–69. URL: https://aclanthology.org/2024.naacl-short.5.

Yang, Ivory, Weicheng Ma, Carlos Guerrero Alvarez, et al. (2025). "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)*, pp. 105–111.

Yang, Ivory, Weicheng Ma, and Soroush Vosoughi (2025). "NüshuRescue: Reviving the Endangered Nüshu Language with AI". In: *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7020–7034.

Yang, Ivory, Weicheng Ma, Chunhui Zhang, et al. (2025). "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)*, pp. 277–284. URL: https://aclanthology.org/2025.naacl-short.24/.