DIGITAL HERITAGE (2025)
S. Campana, D. Ferdani, H. Graf, G. Guidi, Z. Hegarty, S. Pescarin, and F. Remondino (Editors)

Morphosyntactic Variation in Italian and Romansh Dialects: The Manzini & Savoia (2005) Corpus Within Project CHANGES

Greta Mazzaggio[1], Carlo Zoli[2], Neri Binazzi[1], Luca Andrea Ludovico[3], Mael Vittorio Vena[3], Maria Rita Manzini[1] and Leonardo Maria Savoia[1]

[1]University of Florence, Department of Literature and Philosophy, Italy
[2]Smallcodes, Italy
[3]University of Milan, Department of Computer Science, Italy

Abstract

This paper presents the digitization and online dissemination of the Manzini & Savoia (2005) corpus, one of the most comprehensive resources on morphosyntactic variation in Italian and Romansh dialects. Developed within Project CHANGES (Cultural Heritage Active Innovation for Sustainable Society), the initiative responds to the urgent need for systematic documentation and open-access preservation of linguistic diversity. The new platform integrates a relational PostgreSQL database, a Strapi-based backend, and an interactive web interface, offering multiple modes of exploration–including map-based navigation, morphosyntactic query, and access to original fieldwork notebooks. The entire dataset (64,472 examples with IPA transcription and metadata) is openly available on Zenodo for independent research and reuse. The project also explores experimental applications of Large Language Models (LLMs) for automatic annotation, demonstrating the potential for computational approaches in dialectology. This work provides a replicable model for sustainable digital archiving and fosters interdisciplinary research across linguistic, computational, and cultural heritage domains.

CCS Concepts

• ***Human-centered computing*** → ***Web-based interaction;*** • ***Information systems*** → *Relational database model;* • ***Applied computing*** → ***Language translation;*** • ***Social and professional topics*** → *Cultural characteristics;*

1. Introduction

The digital documentation and preservation of linguistic and cultural heritage is increasingly recognized as a scientific and social priority, particularly in the face of accelerating processes of language shift and loss. Compared to other domains of cultural heritage, the linguistic sector has been slower to fully leverage the opportunities offered by digital technologies, resulting in persistent gaps in accessibility, visibility, and long-term sustainability of linguistic resources [MB24].

Within Italy, linguistic diversity is particularly vulnerable. Recent decades have witnessed a marked decline in the everyday use of dialects, especially within the family sphere, as reported in sociolinguistic surveys and national statistics [MLV*23]. This ongoing reduction in the transmission of local varieties constitutes not only a loss of communicative practices, but a significant erosion of intangible cultural heritage–impacting traditional knowledge, oral history, and community identity.

While recent cultural and academic initiatives have modestly increased public awareness of the value of dialects, the risk of irreversible loss remains high for many local varieties. It is therefore essential to undertake systematic documentation, open-access archiving, and innovative digital valorization of at-risk dialectal data.

In response to these challenges, Project CHANGES (Cultural Heritage Active Innovation for Sustainable Society) was launched within the framework of the *Piano Nazionale di Ripresa e Resilienza* (PNRR; National Recovery and Resilience Plan). The project aims to promote innovative, sustainable strategies for the preservation, enhancement, and open dissemination of cultural and linguistic resources. As part of CHANGES, our initiative addresses the urgent need for systematic documentation, open-access archiving, and digital valorization of dialectal data at risk of disappearance. The work is financially supported by the *Ministry of University and Research*, with funding provided by the European Union through the *NextGenerationEU program*.

This paper presents an updated version of the digitization project of the Manzini and Savoia corpus [MS05], initially introduced in Mazzaggio et al. [MLV*23] and Mazzaggio & Binazzi [MB24]. Since the corpus is no longer available in print and physical copies are no longer produced or commercially accessible, digitalization is essential to ensure its preservation and ongoing availability for future research.

The new digital corpus presents substantial technological en-

DOI: 10.2312/dh.20253066

hancements, providing a significantly improved platform for accessing, exploring, and analyzing dialectal linguistic data. The corpus includes data from 459 Italian dialect varieties plus 9 from Corsica and 19 from Switzerland, collected through field research and transcribed using the International Phonetic Alphabet (IPA), with the aim of making this valuable heritage accessible and analyzable for scholars and communities.

The digital platform that we developed integrates a relational PostgreSQL database and an interactive web interface, allowing users to explore dialectal data through multiple analytical tools [MLV*23, MB24]. Specifically, the platform provides:

- *Syntactic analysis* – Users can query the morphosyntactic structures of different dialects, tracing variations and patterns in verb morphology, negation, agreement, and cliticization;
- *Geographical exploration* – Data can be navigated based on regional and provincial classifications;
- *Dialect variety search* – The corpus allows users to search by locality name, linking dialects to their specific town or village of origin;
- *Interactive mapping* – An integrated geospatial visualization tool displays dialect locations via selectable areas on an interactive map, enhancing the spatial analysis of linguistic variation;
- *Original fieldwork notebooks* – Savoia's original handwritten notebooks, where he manually annotated dialectal data during fieldwork, have been scanned and included in the platform. These notebooks contain additional linguistic data not present in the original corpus.

As part of the digitalization process, we ensured that the corpus data remains *Open Access* and reusable for further analysis. To achieve this, a raw dataset has been published on Zenodo [SMM*25]. Future developments include the integration of Large Language Models (LLMs) for automated annotation (*POS-tagging*) and linguistic pattern recognition (*stemming*), further expanding the potential for computational analysis of dialectal data.

This initiative not only preserves Italian linguistic diversity but also establishes a flexible model for developing linguistic corpora and atlases that can be applied to other languages, offering a replicable methodology on a global scale.

The remainder of this paper is organized as follows. Section 2 presents the structure, sources, and distinctive features of the digitized Manzini & Savoia corpus, including both published data and newly digitized fieldwork materials. Section 3 details the technical architecture and functionalities of the web platform. Section 4 examines the integration of georeferencing and interactive mapping for spatial exploration of dialectal data. Section 5 discusses experimental applications involving Large Language Models (LLMs) and computational methods for annotation. Section 6 explores the relationships and potential synergies between this work and other research areas within Project CHANGES. Finally, Section 7 summarizes the main results and outlines future perspectives.

Figure 1: *Leonardo Maria Savoia during a fieldwork session with local informants.*

2. The Manzini & Savoia Digital Corpus: Structure and Features

2.1. Strategies for Data Gathering

The data included in the Manzini and Savoia corpus [MS05] were collected through extensive fieldwork conducted over several decades by Leonardo Maria Savoia (Figure 1). The fieldwork was grounded in a theoretical framework inspired by generative grammar, aiming to model speakers' linguistic competence rather than to map dialect boundaries in a traditional dialectological sense.

Informants were selected based on their native proficiency in the local variety and, when possible, their metalinguistic awareness. The data collection process was based on an ideal dialect questionnaire, progressively refined through field experience. Informants were asked to provide dialectal equivalents of Italian sentences and to comment on their acceptability. These interactions often led to spontaneous variants, paraphrases, or further elaborations, enriching the dataset beyond the initial elicitation targets.

Unlike traditional dialect surveys relying on tape recordings, the methodology here privileged direct, real-time transcription of utterances. Savoia manually recorded data in field notebooks, annotating phonological and syntactic nuances as they emerged in dialogue. Phonetic transcription followed the International Phonetic Alphabet (IPA), with broad criteria to balance linguistic accuracy and readability.

The corpus includes data from 459 Italian dialect varieties plus 9 from Corsica and 19 from Switzerland. The transcription of stress followed consistent conventions: only primary stress was marked, mostly on content words (verbs, nouns, adjectives, tonic pronouns), whereas stress on clitics or function words was generally omitted. Variability in phonetic realization was acknowledged and, when possible, normalized without compromising the empirical richness of the material.

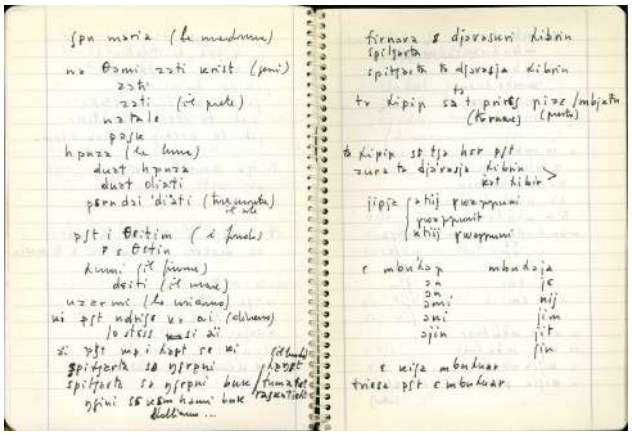

Figure 2: *Example of original fieldwork material: Savoia's handwritten notebook.*

2.2. Original Materials

The foundational materials of the corpus are Leonardo Maria Savoia's handwritten notebooks (Figure 2), which document decades of fieldwork across the Italian-speaking area. These notebooks were used to transcribe, in real time, the linguistic judgments and utterances provided by informants during elicitation sessions. They contain not only the dialectal data themselves, but also extensive marginalia, contextual notes, and informants' metalinguistic reflections–elements that enrich the linguistic analysis and provide insight into fieldwork dynamics.

The 2005 publication [MS05] represents a selection of this material. However, a significant portion of the original field notes–both preceding and following the publication–remained unpublished and are now being digitized as part of the current project. This includes new varieties and data collected after 2005, as well as expanded documentation for previously attested varieties.

The digital corpus will therefore include:

- the **core dataset** derived from the 2005 volumes, available as an Open Access downloadable dataset on Zenodo [SMM*25];
- the **digitized notebooks** in high-resolution PDF format, providing access to the complete handwritten source material, including data not featured in the published corpus.

This dual approach ensures transparency and replicability of the data while significantly extending the empirical base for future research.

2.3. Organization in the Printed and Digital Corpus

In the printed volumes [MS05], the data are organized according to morphosyntactic phenomena, rather than by geographic region. Each chapter is dedicated to a specific area of grammatical variation–such as agreement, cliticization, negation, or auxiliary selection–and illustrates variation through a comparative presentation of dialectal examples drawn from multiple varieties.

This typologically oriented structure reflects the authors' theoretical commitment to generative grammar and parameter theory [Cho95]: rather than mapping dialect geography per se, the corpus is designed to test hypotheses about linguistic universals and microvariation. Geographic information is included for each data point, allowing for spatial contextualization, but the primary logic of organization is linguistic.

The digital corpus retains this morphosyntactic structure and complements it with a variety of interactive tools. Users can still browse the data by grammatical category, but also explore them by locality, region, or via an interactive map. This hybrid approach preserves the original theoretical framework while opening up new pathways for corpus navigation and geolinguistic analysis.

3. The Web Platform

3.1. Design and Structure

The web platform developed in the context of Project CHANGES represents a central digital infrastructure for the digitization, cataloging, and online visualization of linguistic data from the Manzini & Savoia corpus. The platform is publicly accessible at https://manzinisavoia.changes.unifi.it/.

Architecturally, the platform is built upon three main technological pillars. First, a relational database implemented in *PostgreSQL* serves as the digital archive, ensuring robust organization, querying, and preservation of the linguistic data. The dataset publicly available on Zenodo represents a filtered dump of such a database in comma-separated values (CSV) format.

Second, the backend management is handled through *Strapi*, an open-source headless Content Management System (CMS), which provides an efficient and secure environment for the structured insertion, editing, and validation of linguistic examples and associated metadata. The use of *Strapi*'s customizable content models and granular permissions allows for a flexible and collaborative workflow, supporting both technical and non-technical contributors (such as linguists, annotators, and research assistants) while maintaining full traceability and control over data modifications.

Third, the frontend is developed as a responsive web application, implemented using modern web-compliant languages (HTML, CSS, JavaScript) to ensure intuitive access from any device and to facilitate multilingual use. The default interface is in English, supporting international accessibility and future expansion to additional languages or specialized interaction modes. The landing page is shown in Figure 3.

The platform is continuously evolving, with new features and datasets being added as the project progresses.

3.2. Graphical User Interface

The graphical interface has been carefully designed to support multiple types of users and research goals. The structure of the corpus is mirrored in the platform layout, with linguistic examples organized and accessible via a modular navigation system. Metadata and source references are included for each example, maintaining transparency and facilitating scholarly citation. Additionally, the

Figure 3: *Landing page of the web platform.*

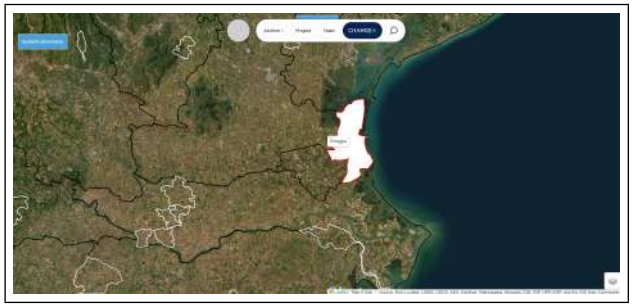

Figure 4: *Explore by map.*

Figure 5: *Explore by area.*

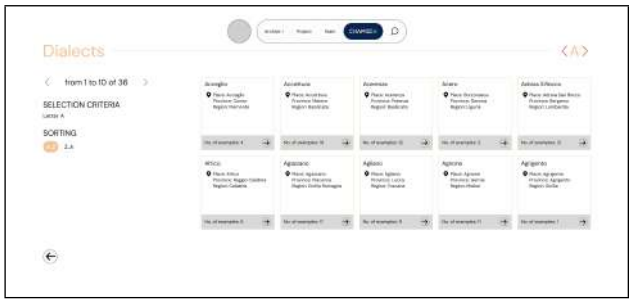

Figure 6: *Explore by name.*

Figure 7: *Explore by syntax.*

interface is designed to accommodate further expansion, including the integration of glosses, translations, and interlinear annotations.

3.3. Navigation Paths

One of the central innovations of the current digital platform lies in the flexibility it offers for accessing and exploring the corpus data, particularly with the fully interactive georeferenced map, which presents a visually intuitive and spatially grounded entry point into the corpus data. The system is designed to accommodate different user profiles–from researchers conducting in-depth linguistic analysis to educators, students, and members of the general public interested in exploring dialectal variation across Italy and beyond.

To this end, the platform provides multiple and complementary navigation paths:

- **Explore by map** – An interactive georeferenced map enables users to visualize all dialectal varieties covered in the corpus. Each data point is linked to a specific locality and can be selected to access the corresponding linguistic examples. This map-based interface promotes a spatial understanding of morphosyntactic variation and offers intuitive access for non-specialist users (Figure 4);
- **Explore by area** – This function allows users to browse the data by administrative divisions (region and province). It is particularly useful for those interested in comparing dialects within specific geographical units or conducting regionally constrained studies (Figure 5);
- **Explore by name** – Dialectal varieties can also be retrieved alphabetically by locality name. This index-based navigation is suited for targeted queries and for users familiar with specific places of interest (Figure 6);
- **Explore by syntax** – Reflecting the structure of the original Manzini & Savoia (2005) volumes, this path organizes data according to morphosyntactic phenomena. Users can navigate through major grammatical categories such as agreement, negation, auxiliary selection, and cliticization. This linguistically motivated path is especially useful for comparative research and typological studies (Figure 7);
- **Search function** – A dedicated search engine will allow for more complex queries, combining multiple filters (e.g., locality, syntactic feature, phonological form). A simple search bar supports quick lookups, while an advanced mode enables precise control over search parameters for linguistic research.

These various pathways are not siloed but interconnected: users can seamlessly switch from one mode to another. For example, a syntactic query may lead to a specific locality, which can then be explored spatially via the map interface. This cross-referencing mechanism ensures a fluid and multimodal user experience. By diversifying the modes of access, the platform supports a pluralistic approach to linguistic inquiry–geographic, formal, lexical, or

Figure 8: *Detail page of a dialect.*

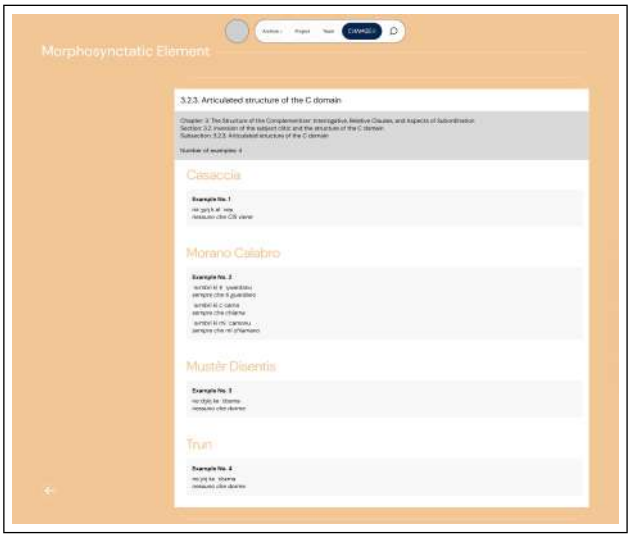

Figure 9: *Detail page of a morphosyntactic element.*

sociolinguistic–enhancing both scholarly depth and public engagement with the data.

Figures 8 and 9 show, respectively, the detail page of a dialect and that of a specific morphosyntactic element. It can be noted that, in both cases, the examples collected by Savoia are displayed, but they are grouped according to different criteria.

3.4. Zenodo Dataset

Beyond the interactive functionalities of the web platform, the entire raw dataset is freely available as an Open Access resource on Zenodo [SMM*25]. The corpus consists of a total of 64,472 linguistic examples, each comprising a dialectal sentence transcribed in IPA together with its Italian gloss. For each example, additional metadata is provided in the header line of the CSV file, with self-explanatory field names such as:

- *Locality_official_name*: Official name of the municipality or village;
- *Locality_short*: Shortened version of the locality name;
- *Chapter_title*: Title of the chapter in Manzini & Savoia (2005) where the example appears;
- *Chapter_subtitle*: Subchapter title (if applicable);
- *Example*: Dialectal sentence (IPA transcription);
- *Glossa*: Italian gloss of the example.

All data are provided in standard CSV format, enabling researchers to download, analyze, visualize, or combine the corpus with other resources as needed. This ensures maximum transparency, scientific reuse, and the possibility of independent or collaborative research extending well beyond the functionalities of the web platform.

3.5. Toward Integration with Other Corpora

The modular and scalable architecture of the current platform has been explicitly designed with interoperability and extensibility in mind, anticipating future developments in the digital humanities and language documentation. In the contemporary landscape, one of the primary challenges facing corpus-based dialectology is the fragmentation and isolation of valuable resources, each often relying on different data formats, metadata standards, and interface solutions.

Our goal is to move beyond the limitations of stand-alone databases and toward a unified digital environment where heterogeneous dialectal corpora–including not only Romance but also Germanic, Slavic, and minority language varieties–can be seamlessly integrated, explored, and compared. Such integration requires not only robust technological infrastructure, but also the adoption of common data models, crosswalks for metadata, and standardized linguistic ontologies.

By laying the groundwork for a shared and open ecosystem, the platform fosters cross-corpus analysis, comparative studies, and the cumulative enrichment of linguistic resources. This approach also aligns with Open Science and FAIR (Findable, Accessible, Interoperable, Reusable) data principles, ensuring long-term sustainability, traceability, and reproducibility of linguistic research.

Ultimately, the platform aims to serve as both a reference implementation and a flexible node within a broader network of digital linguistic archives. In this way, it offers not only a digital edition of the Manzini & Savoia corpus, but also a scalable, future-proof foundation for the collaborative study and preservation of linguistic diversity in the digital age.

Please note that the integration with non-linguistic datasets produced by other research teams within Project CHANGES will be addressed in Section 6.

4. Georeferencing

In line with the principles of accessibility, interactivity, and user-oriented design, the platform provides an alternative exploration mode through a georeferenced map. The map displays all the localities where dialectal examples have been collected, offering a visual and spatial representation of linguistic data.

Each locality is marked on an open-source, interactive map. Instead of using traditional pinpointing, thus marking very specific locations, the map allows for the visualization and selection of the areas where the dialect is spoken. By selecting an area, a pop-up

window with dialect details is opened, and a link lets users be redirected to a dedicated page containing examples associated with that location. This approach allows users to engage with the archive more intuitively, facilitating comparative studies across regions and supporting both lay and scholarly audiences.

The integration of geographic metadata not only enriches the browsing experience but also enhances the analytical possibilities, enabling cross-referencing with sociolinguistic, historical, and demographic data. From a research perspective, the geospatial interface contributes to the study of linguistic variation and distribution, while from an educational standpoint, it supports didactic activities by visualizing language diversity in context.

This map-based access is fully interoperable with other sections of the archive, such as thematic queries by morphosyntactic elements or dialect names, reinforcing the multimodal nature of the platform.

Beyond its utility in the linguistic domain, the use of georeferencing is a shared feature across several non-linguistic research areas involved in Project CHANGES. As an example, in studies focusing on anthropology, ethnomusicology, or food culture, geographic positioning provides essential context and enhances accessibility. By adopting open maps as a common interface, the platform establishes a coherent framework through which heterogeneous contributions from the participating research groups can be explored. This layered approach–where data from multiple domains can be superimposed or filtered–enables interdisciplinary insights and fosters a more integrated understanding of the studied phenomena. Users are not only able to browse specific datasets but can also identify intersections among different disciplines through their spatial dimension. Please note that this unifying exploration interface is being developed for the global project web platform, and a more detailed description would go beyond the scope of the present paper.

5. LLMs and Experimental Computational Applications

The application of Natural Language Processing (NLP) methods to dialectal data remains a significant challenge [AKG*23], primarily due to the scarcity and fragmentation of available resources–a situation that mirrors the difficulties encountered with other low-resource or small-sized training datasets [FBPB*24]. While Large Language Models (LLMs) have demonstrated impressive performance across mainstream NLP tasks [CWW*24, Xue24, Zan24], their success typically depends on vast quantities of annotated or easily tokenizable data [WZW*23, VAL*24], a condition not met by most dialectal corpora. In this context, the Manzini & Savoia corpus offers a unique advantage: its systematic pairing of dialectal data with Italian interlinear glosses.

We explore a methodology that leverages these glosses to enable automatic morphosyntactic annotation of dialectal sentences. First, glosses are processed using a pre-trained Italian model (`it_core_news_lg`, see https://spacy.io/models/it), which generates part-of-speech tags, morphological features, and dependency relations. These annotations are then transferred to the corresponding dialectal tokens via alignment rules, preserving syntactic structure while allowing for cross-linguistic variation.

Consider the following example from the corpus:

(1) ik'ke ttu f'fai? (Firenze)
(2) *cosa ClS fai?*
'what ClS are you doing'

Here, the gloss contains both standard Italian words and abstract tags like `ClS`, representing a subject clitic not present in Italian. Our system annotates the gloss, reintegrates such dialect-specific elements with predefined morphosyntactic labels, and projects the enriched information onto the dialectal layer. The result is a set of fully annotated sentences that can be queried, visualized, and used for linguistic or typological analysis.

This method opens two key perspectives: first, the development of an advanced search interface allowing users to filter data not only by geographical or syntactic criteria, but also by lexical or morphological patterns; second, the possibility of fine-tuning NLP models directly on dialectal data, progressively reducing reliance on the Italian layer. Such a system would represent a substantial step toward inclusive computational linguistics, offering tools capable of addressing under-resourced languages with precision and linguistic sensitivity.

6. Integration with Other Research Areas

Project CHANGES provides the Italian national framework within which the research activities described in this paper were carried out. The project seeks to establish an international centre of excellence for training, research, and technology transfer in the field of culture and cultural heritage. Its general goal is to enhance Italian cultural assets by promoting innovative and sustainable approaches to their preservation and valorization, fostering long-term collaborations between academia and industry, and generating new employment opportunities within the cultural sector.

Specifically, Work Package 1 (WP1) – "Mapping Intangible Cultural Heritage in Performing Arts, Music, Audiovisual Media, Design and Fashion, Craftsmanship, and Linguistic Diversity" focuses on the development of a tool for mapping, describing, and enhancing the intangible cultural heritage. The goal of WP1 is to provide structured data relevant to intangible cultural heritage, organized into thematic maps, archives, and audiovisual documents.

To populate the platform, seven research groups have been established, each pursuing both the general goal and specific research objectives:

1. *Cinema Group*
2. *Design, Fashion, and Craftsmanship Group*
3. *Ethnomusicology and Cultural Anthropology Group*
4. *Food Culture Group*
5. *Linguistics Group*
6. *Musicology Group*
7. *Theatre Group*

In line with the overall objectives of the project, points of contact and hybridizations between the activities of the various research groups can be highlighted. In this sense, the *Linguistics Group*

presents numerous opportunities for collaboration with the other research groups within WP1, fostering a multidisciplinary approach to the study of cultural heritage.

One clear connection is with the *Cinema Group*. The use of dialects in Italian cinema provides a rich source for the *Linguistics Group*'s study, as films often represent regional identities and social dynamics through language. This offers a fertile ground for examining how dialects are portrayed and the extent to which they reflect or shape cultural narratives. Collaborative work could involve linking dialectal variations in films with sociolinguistic trends, as well as analyzing how films have been subject to censorship based on dialect use, which in turn could provide insight into the historical and cultural significance of regional dialects in Italy. As a result of the dialogue between these research groups, a dedicated web portal has already been released. The platform is publicly available at https://dialettialcinema.changes.unimi.it/

Although seemingly distant, the *Design, Fashion, and Craftsmanship Group* also has potential connections with the *Linguistics Group*. Many artisanal practices, particularly those with deep regional roots, use terminology that is specific to local dialects. The two groups could document and analyze dialectal lexicons associated with local craftsmanship, design, and fashion. This could include linking dialect terms to particular techniques or regions, as well as exploring how language and material heritage intersect in Italian traditions of design and craftsmanship.

The *Food Culture Group* provides another interesting avenue for collaboration. Food-related terms and dishes are often named and described in local dialects, making them an essential aspect of regional linguistic heritage. A collaborative project between groups could focus on compiling a dialectal glossary of food-related terminology, exploring the linguistic diversity within Italian culinary traditions. Additionally, the two groups could investigate how dialects are used to promote regional food identities in both local and global contexts, highlighting how food and language together contribute to the broader cultural heritage.

Another significant intersection is with the *Ethnomusicology and Cultural Anthropology Group*. The focus on urban soundscapes and performance practices can be linked to the study of dialectal variation, especially within specific communities and urban spaces. A joint ethnographic study could investigate the use of dialects in community performances, festivals, or religious rituals. By mapping how dialects are used in these settings, both groups could explore the role of dialect in constructing identity, especially in the context of migration and cultural exchange within urban environments.

The *Music Group* also intersects with the *Linguistics Group*'s work in a compelling way. Many Italian songs, particularly those from traditional or folk genres, are performed in dialects. These musical expressions offer an opportunity for the *Linguistics Group* to analyze phonological and syntactic features of dialects, while exploring the sociocultural contexts in which these songs are produced and consumed. A collaborative effort could focus on creating a dialect-tagged corpus of song lyrics, allowing for a deeper exploration of the relationship between language and musical tradition, and examining how dialects serve as cultural markers in the Italian music scene.

Finally, the *Theatre Group* presents another natural connection. Theatrical performances often employ dialects, especially in regional and folk theatre traditions. This offers an excellent opportunity to collaborate on cataloging and analyzing theatre scripts or recorded performances that feature dialectal language. This partnership could focus on how dialects function as dramatic tools in theatre, as well as providing a historical overview of dialect usage in Italian dramaturgy, from classical playwrights like Goldoni to contemporary theatrical works.

7. Conclusion

This paper has presented the current state of the digitization project of the Manzini & Savoia (2005) corpus, a foundational resource for the study of morphosyntactic variation in Italian and Romansh dialects. Building upon previous work [MLV*23], the updated platform introduces a range of technological enhancements aimed at ensuring the long-term preservation, accessibility, and analytical usability of the corpus data.

The new digital infrastructure offers multiple modes of exploration–by locality, region, morphosyntactic category, and interactive map–thereby broadening the research potential and extending access to a wider audience. Alongside the digitized data from the 2005 volumes, the inclusion of previously unpublished fieldwork notebooks in PDF format will significantly expand the empirical scope of the resource. These documents offer an unfiltered view of in-situ elicitation practices and linguistic variation, and will be made available for consultation as part of the corpus infrastructure.

In line with the objectives of Project CHANGES, this work contributes to the development of sustainable, open-access tools for linguistic and cultural preservation.

Future directions include the integration of computational linguistic tools such as POS-tagging and morphological analyzers powered by large language models (LLMs), which will further enhance the corpus' analytical depth. Moreover, the modular structure of the platform lays the foundation for connections with other domains–such as audiovisual media, music, and geolinguistics–thereby fostering interdisciplinary collaboration and enriching the representation of dialects within broader cultural frameworks.

Ultimately, this project reaffirms the relevance of dialectological data in contemporary linguistic inquiry and cultural heritage management. By translating decades of fieldwork into a dynamic, extensible digital format, it ensures that the voices and structures of Italy's linguistic diversity continue to inform theory, pedagogy, and public discourse in the digital age.

Acknowledgements

This research is a part of Project CHANGES, funded by the National Recovery and Resilience Plan (NRRP), Mission 4 (Education and Research), Component 2 (From Research to Enterprise), Investment 1.3 (Extended Partnerships), Theme 5 (Humanities and Cultural Heritage as Laboratories of Innovation and Creativity). We gratefully acknowledge its financial support, which made this work

possible. The authors would like to especially thank Alice Biancardi, Politecnico di Milano, who designed the graphical interface of the web platform, and Achille Fusco, University of Florence, for his collaboration in the development and implementation of the LLM-based annotation system.

Credit Author Statement

Greta Mazzaggio (Sections 1, 2, 3, 5, 7), Carlo Zoli (Sections 3 and 4) and Luca Andrea Ludovico (Sections 1, 3, 6, 7) drafted the manuscript. Neri Binazzi, Mael Vittorio Vena, Maria Rita Manzini and Leonardo Maria Savoia revised the manuscript.

References

[AKG*23] AHIA O., KUMAR S., GONEN H., KASAI J., MORTENSEN D., SMITH N., TSVETKOV Y.: Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (Singapore, Dec. 2023), Bouamor H., Pino J., Bali K., (Eds.), Association for Computational Linguistics, pp. 9904–9923. URL: https://aclanthology.org/2023.emnlp-main.614/, doi:10.18653/v1/2023.emnlp-main.614. 6

[Cho95] CHOMSKY N.: *The minimalist program*. MIT press, 1995. 3

[CWW*24] CHANG Y., WANG X., WANG J., WU Y., YANG L., ZHU K., CHEN H., YI X., WANG C., WANG Y., ET AL.: A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology 15*, 3 (2024), 1–45. 6

[FBPB*24] FUSCO A., BARBINI M., PICCINI BIANCHESSI M. L., BRESSAN V., NERI S., ROSSI S., SGRIZZI T., CHESI C.: Recurrent networks are (linguistically) better? an (ongoing) experiment on small-LM training on child-directed speech in Italian. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)* (Pisa, Italy, Dec. 2024), Dell'Orletta F., Lenci A., Montemagni S., Sprugnoli R., (Eds.), CEUR Workshop Proceedings, pp. 382–389. URL: https://aclanthology.org/2024.clicit-1.46/. 6

[MB24] MAZZAGGIO G., BINAZZI N.: Valorizzare il patrimonio immateriale: un'esperienza di digitalizzazione del dialetto. *DILEF Rivista digitale del Dipartimento di Lettere e Filosofia 3* (2024), 224–242. 1, 2

[MLV*23] MAZZAGGIO G., LUDOVICO L., VENA M., MANZINI M. R., SAVOIA L. M., ET AL.: Morphosyntax of italian and romance varieties: Presentation of the manzini and savoia (2005) corpus and its digitalization. *Bollettino dell'Atlante Linguistico Italiano 2023*, 47 (2023), 185–210. 1, 2, 7

[MS05] MANZINI M. R., SAVOIA L. M.: *I dialetti italiani e romanci: Morfosintassi generativa*. Edizioni dell'Orso, 2005. 1, 2, 3

[SMM*25] SAVOIA L. M., MANZINI M. R., MAZZAGGIO G., LUCA ANDREA L., VENA M. V., ZOLI C., BALDI B., FRANCO L., BINAZZI N.: The manzini & savoia (2005) corpus: Morphosyntactic variation in italian and romansh dialects, Feb. 2025. URL: https://doi.org/10.5281/zenodo.14803999, doi:10.5281/zenodo.14803999. 2, 3, 5

[VAL*24] VIEIRA I., ALLRED W., LANKFORD S., CASTILHO S., WAY A.: How much data is enough data? fine-tuning large language models for in-house translation: Performance evaluation across multiple dataset sizes. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)* (Chicago, USA, Sept. 2024), Knowles R., Eriguchi A., Goel S., (Eds.), Association for Machine Translation in the Americas, pp. 236–249. URL: https://aclanthology.org/2024.amta-research.20/. 6

[WZW*23] WANG Z., ZHONG W., WANG Y., ZHU Q., MI F., WANG B., SHANG L., JIANG X., LIU Q.: Data management for large language models: A survey. *CoRR* (2023). 6

[Xue24] XUE Q.: Unlocking the potential: A comprehensive exploration of large language models in natural language processing. *Applied and Computational Engineering 57* (2024), 247–252. 6

[Zan24] ZANGANA H. M.: Harnessing the power of large language models. *Application of Large Language Models (LLMs) for Software Vulnerability Detection* (2024), 1. 6