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

Small Codes: a platform for digital resources and tools for minority languages and dialects

Carlo Zoli[1,2], Greta Mazzaggio[1,3], Neri Binazzi[1,3]

[1]Smallcodes, Italy
[2]Free University of Bozen-Bolzano, Department of Education, Italy
[3]University of Florence, Department of Literature and Philosophy, Italy

Abstract

Small Codes is an open digital infrastructure designed to support the preservation and revitalization of minority languages through scalable, interoperable and user-friendly tools. The platform combines linguistic data management with web-based technologies, offering an integrated suite of software modules—including online dictionaries, spell-checkers, corpus alignment systems, linguistic maps, and multimedia archives—tailored for under-resourced and dialectally fragmented languages.
Unlike standard language technology pipelines designed for dominant languages, Small Codes supports linguistically diverse input and community-led data models. It operates through a federated, semi-industrial development model, balancing long-term sustainability with flexibility for academic and institutional partners.
This paper outlines the system architecture and core functionalities of Small Codes, presents selected implementation scenarios, and discusses its contribution to digital heritage and computational dialectology.

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

In a hyper-connected and increasingly digitized society, access to written language has become essential for full participation in public, educational, and technological life. While dominant national languages benefit from decades of computational development and large annotated corpora, many minority and regional languages are still structurally excluded from the digital ecosystem. These languages—often unwritten, under-resourced, and highly variable across dialects—face a dual challenge: the risk of further marginalization in the digital realm, and the opportunity to reinvent their role in society through inclusive digital infrastructures.

The *Small Codes* reality was conceived to meet this challenge. Its name plays on a dual semantic register: on one hand, it evokes "software code"—the computational architecture that underpins modern language technologies—and on the other, it recalls the linguistic notion of "codes" as in "language codes". In this sense, *Small Codes* refers simultaneously to the technological instruments required for digital infrastructure and to the small, minoritized languages that risk digital invisibility and long-term disappearance. The term encapsulates both the medium and the message: small languages need code, and code needs to serve small languages.

Developed over more than a decade through partnerships with universities, cultural institutions, local administrations, and community groups, *Small Codes* is a unified digital infrastructure for the creation, management, and dissemination of language resources for minority and lesser-used languages. Its modular architecture supports lexicon building, spell-checking, morphological analysis, collaborative neology, corpus alignment, and multimedia publication. These modules are interoperable and scalable, allowing each linguistic community to configure its own instance while benefiting from shared standards and tools.

The platform's guiding vision is articulated through the metaphor of the "dwarfs sitting on the shoulders of giants" [ZSM13].

Over the past ten years, *Small Codes* has been successfully applied to dozens of language projects, especially in Italy and alpine regions, where dialectal fragmentation is extremely high. In the case of Ladin, for example, the platform manages more than 150 phonetic variants for a single lexical item, offering a "standardizing dictionary" that includes both local forms and a shared orthographic norm. For Sardinian, *Small Codes* provides a dialect-aware spell-checker and morphological analyzer that can detect and correct common errors arising from the interference of spoken varieties. For Francoprovençal and Occitan, the platform has supported the digitization of handwritten archival materials and their transformation into searchable, aligned digital corpora.

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.2312/dh.20253331

delivered by EUROGRAPHICS DIGITAL LIBRARY www.eg.org diglib.eg.org

These implementations demonstrate the platform's adaptability to very different linguistic and institutional contexts. While the core codebase remains unified, the data models, orthographies, and interface languages vary across each deployment. The system supports Unicode and TEI, allows export in LMF-compliant XML/JSON, and integrates with tools such as Python SpaCy and Hunspell. Its architecture is built around a relational database and a REST API layer, enabling real-time updates and interoperability with both desktop and web-based applications.

With regard to printed publications, the lexicographical platform can be integrated with Adobe Indesign for importing XML files, which are easily laid out using customised scripts, thus obtaining a fully edited professional dictionary within hours.

A key aspect of the *Small Codes* model is its balance between technological rigor and community engagement. Rather than treating speakers as passive users or informants, the platform encourages participation in all phases of development—from the identification of lexical gaps to the validation of neologisms and the design of user interfaces. This co-design approach is critical not only for the accuracy of the resources, but also for their social legitimacy and uptake.

Standardization, often viewed with suspicion by minority language speakers, is approached pragmatically in *Small Codes*. The platform does not impose a single written form but allows for multiple levels of variation—phonetic, orthographic, dialectal—to be encoded and accessed. Standard forms are suggested not as replacements, but as bridges between local usage and broader communication needs (e.g., education, public administration, media). In doing so, *Small Codes* follows the logic of Ausbau rather than Abstand, enabling languages to expand their functions without erasing their internal diversity [Klo67].

Today, this long-term infrastructure is entering a new era of consolidation and expansion. Thanks to the structural funding and institutional support provided by the Italian *Piano Nazionale di Ripresa e Resilienza (PNRR)* and the broader framework of *CHANGES – Cultural Heritage Active Innovation for Sustainable Society*, *Small Codes* is now part of a national effort to integrate digital humanities, language documentation, and cultural sustainability. Within *CHANGES Spoke 2*—devoted to Creativity and Intangible Cultural Heritage and the collaboration with the Department of Literature and Philosophy (DILEF) of the University of Florence—the platform is being used to digitize and annotate decades of dialectal fieldwork [MS05, MB24, MLV*23, SMM*25].

This new phase is not just a scaling-up in technical terms, but a redefinition of purpose and scope. Moreover, the integration of *Small Codes* within the PNRR framework aligns its development with long-term national priorities: digital transition, cultural sustainability, and open innovation. The project reinforces the idea that intangible heritage—proverbs, dialects, oral narratives—requires the same level of infrastructural investment as physical monuments and museums. In this sense, *Small Codes* is not merely a linguistic tool, but a component of a wider cultural strategy.

Through this collaboration, the platform gains access to new computational resources, institutional networks, and interdisciplinary exchanges. At the same time, it provides CHANGES with a tested, flexible, and community-oriented system capable of supporting a wide range of documentation practices—from lexicography to syntax, from dialect mapping to literary corpus development.

This article presents the architecture, principles, and use cases of the *Small Codes* platform. We argue that this represents a model for digital linguistic infrastructure that is not only technically sound but also culturally sensitive and sustainable. In doing so, *Small Codes* offers a replicable path for the revitalization of linguistic heritage within the broader framework of digital innovation and cultural resilience.

2. Background and Related Work

The intersection of computational linguistics and digital heritage has generated a growing body of tools and platforms aimed at supporting language documentation and revitalization. However, most of these efforts are oriented toward majority or standardized languages, and often lack the architectural flexibility required to handle the fragmentation, orthographic variation, and oral dominance that characterize minority language contexts.

Several projects and frameworks have emerged to address linguistic diversity from a technological perspective. Initiatives like ELAR, DoBeS, and Living Dictionaries have demonstrated the potential of digital repositories for endangered languages. Similarly, the Summer Institute of Linguistics (SIL) has produced widely used tools such as FieldWorks and Toolbox, which are particularly suited for field linguists working with low-resource languages. While these platforms offer robust functionality for lexicon creation and phonological analysis, they tend to operate as standalone applications with limited support for interoperability, distributed collaboration, or web-based delivery.

In the European context, infrastructures such as CLARIN and META-SHARE have promoted standards for resource sharing and metadata encoding, advocating for FAIR principles and long-term preservation. Nonetheless, their focus is primarily on well-resourced languages and lacks the granularity and community-centered features necessary for minority language documentation in highly multilingual regions, such as Italy or Switzerland.

From a methodological standpoint, traditional NLP pipelines (e.g., those relying on off-the-shelf POS taggers or dependency parsers) perform poorly on dialectal data, due to both the scarcity of annotated corpora and the mismatch in linguistic features. Transfer learning and multilingual embeddings have partially mitigated this issue, but require costly fine-tuning and substantial language-specific preprocessing.

The *Small Codes* platform was designed to fill this methodological and infrastructural gap. Its architecture supports dialectally stratified data, hybrid text–audio corpora, and collaborative workflows that allow for both academic and grassroots participation. Moreover, its integration of geospatial visualization and linguistic ontologies facilitates novel forms of linguistic analysis and cultural valorization.

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

3. Architecture and System Design

The *Small Codes* platform is structured as a modular web-based infrastructure designed to support the full lifecycle of digital language resource development: from data ingestion and processing to visualization and dissemination. Its architecture reflects the need to accommodate heterogeneous linguistic data—such as dialectal variation, audio recordings, scanned manuscripts, and morphosyntactic annotations—within a unified and extensible framework.

3.1. System Overview

The platform consists of four main layers:

- **Data Layer**: A relational database schema (abstracting from single SLQ dialects through a middle layer such as Hibernate) optimized for linguistic data, including tables for lexical entries, dialect variants, interlinear glosses, metadata, and geographic coordinates. The schema supports multilingual alignments, user-contributed variants, and linked terminologies. Where applicable, the structure is compatible with standards such as LMF (Lexical Markup Framework) and TEI.
- **API and Service Layer**: RESTful web services expose structured access to all data entities, allowing both internal modules and third-party applications to interact with the corpus. All endpoints are documented via OpenAPI specifications. Authentication and access control are implemented using a role-based permission model to differentiate between contributors, reviewers, and public users.
- **Application Layer**: A set of modular front-end components implemented using modern JavaScript frameworks (e.g., Vue.js or React), enabling rich interactivity for dictionary browsing, term creation, dialect comparison, and geospatial navigation. The interface dynamically adapts to the data available for each language, offering context-specific features (e.g., paradigmatic tables, audio playback, term approval workflows).
- **Integration Layer**: Connectors to external NLP pipelines (e.g., SpaCy for Italian gloss analysis, UDPipe for syntactic parsing), metadata harvesters (e.g., OAI-PMH for CLARIN compatibility), and GIS engines (Leaflet, QGIS) to support map-based visualization. This layer also includes the infrastructure for ingesting legacy datasets, such as digitized notebooks and TEI-encoded corpora.

3.2. Modularity and Scalability

Each component of the platform is designed for independent deployment and versioning. Institutions working on specific languages or dialects can host dedicated instances while benefiting from shared updates and community support. New languages can be onboarded through a structured configuration process that includes lexicon import, dialect mapping, and orthographic normalization profiles. The system supports vertical scaling (through PostgreSQL optimization and caching) and horizontal scaling (via container orchestration tools such as Docker and Kubernetes).

3.3. Standards and Interoperability

To ensure long-term compatibility and openness, *Small Codes* adheres to established standards and best practices in linguistic data management:

- **FAIR principles**: All datasets are published with persistent identifiers, machine-readable metadata, and open licenses whenever possible.
- **ISO/TEI compatibility**: The data model supports morphosyntactic features aligned with ISOcat and data export in TEI-XML for integration with digital archives and corpora.
- **Geographic standards**: All locations are geo-referenced using ISO 3166 and GeoJSON, enabling use in external mapping systems and compatibility with digital atlases.

3.4. Sustainability Model

Unlike typical research-driven software that is discontinued after project funding ends, *Small Codes* is built and maintained under a hybrid model: community-driven governance (via partner institutions) combined with commercial-grade technical development. A federated deployment architecture allows local stakeholders—such as universities, linguistic academies, or municipalities—to retain control over their instances while benefiting from a shared technical core. Continuous updates and support are coordinated via Git-based versioning and issue tracking.

4. Use Cases and Implementation Scenarios

The *Small Codes* platform has been adopted and validated across a range of linguistic and territorial contexts, each characterized by different documentation traditions, institutional actors, and levels of digital readiness. These use cases demonstrate the system's adaptability to complex sociolinguistic settings, as well as its capacity to handle heterogeneous data formats and user needs.

4.1. Italian dialects of Switzerland: Multidialectal Lexicography

In collaboration with the Centro di Dialettologia e Etnografia della Svizzera italiana, *Small Codes* has been deployed to support the creation of a unified yet variant-aware lexicographic system. The platform manages over 150 dialectal variants per lexical entry, offering a normalized reference form (fábrica 'factory') (Figure 1) alongside geolocated dialectal realizations.

The system integrates:

- A collaborative neology module for educational and administrative terminology;
- Automatic export to print layouts via InDesign XML;
- Spell-checking and auto-completion tools customized for dialectal interference patterns.

4.2. Ladin: Language Policy and Terminological Planning

Working with the autonomous administration of Trentino/Alto Adige, *Small Codes* has been employed to consolidate terminological databases used in public administration, education, and health

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

Figure 1: *Example of multi-dialectal semantic and phonetic variation in Italian Switzerland dialects.*

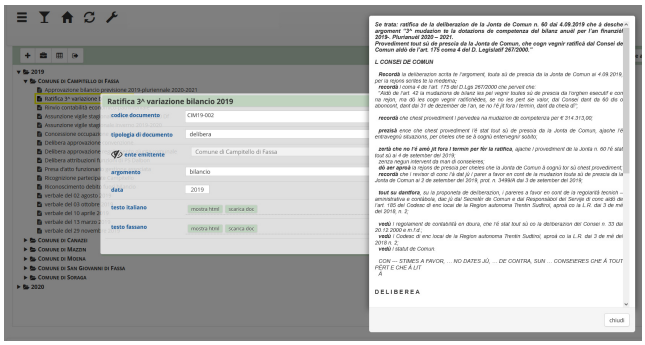

Figure 2: *Text aligner for administrative terminology.*

communication. The system's dialect-aware spell-checker accounts for major orthographic and phonetic differences across southern and central-northern varieties. (Figure 2)

The Ladin instance features:

- Integration with legal and educational translation workflows;
- Web services for terminology lookup in government portals;
- Morphological analyzers for both standard and dialectal forms.

4.3. Francoprovençal and Occitan: Archival Digitization and Lexicon extraction

In Piedmont and the Aosta Valley, *Small Codes* has supported the digitization of handwritten lexical archives, converting them into structured digital corpora. Optical Character Recognition (OCR) pipelines feed into the lexicon builder, which enables fine-grained dialectal mapping. The resulting resources are aligned with regional atlases and used in local schools and cultural programming.

Features include:

- TEI import/export for annotated manuscripts;
- Automatic variant clustering using phonetic similarity and Levenshtein distance;
- Custom dashboards for educators and community animators. (Figure 3)

Figure 3: *Occitan and Francoprovençal digital corpus.*

4.4. Italian dialects: the Manzini & Savoia corpus

The Manzini & Savoia (2005) corpus is an ambitious and influential documentation efforts undertaken for Italian and Romansh dialects, representing the outcome of decades of systematic fieldwork [MS05]. This resource includes data from 459 Italian dialect varieties, 9 from Corsica, and 19 from Switzerland, all collected through direct interaction with native speakers and transcribed in the International Phonetic Alphabet (IPA). What distinguishes the corpus is not only its scope, but its commitment to capturing microvariation: each example is accompanied by an Italian interlinear gloss, facilitating both linguistic analysis and computational processing.

The recent digitization project, conducted within the framework of the CHANGES initiative and under the guidance of the University of Florence and the *Small Codes* platform, has transformed this foundational collection into a dynamic, open-access digital corpus [MLV*23, MB24, SMM*25]. The digital version is structured around a robust PostgreSQL database, interfaced through an interactive web platform (`https://manzinisavoia.changes.unifi.it/`) that enables advanced exploration, visualization, and analysis of the data.

4.4.0.1. Structure and enrichment of the digital corpus. The corpus is organized thematically according to morphosyntactic phenomena—agreement, cliticization, negation, auxiliary selection, and more—mirroring the typological focus of the original publication. For each example, metadata includes the locality (official name and short form), bibliographic reference, and detailed phonetic transcription. The platform supports multiple access paths: by locality, by region/province, by morphosyntactic category, or via an interactive georeferenced map. This multimodal approach enhances both the scholarly depth and the public accessibility of the resource.

A core innovation of the digitization effort is the integration of the original field notebooks of Leonardo Maria Savoia. These handwritten materials, now available as high-resolution PDFs, provide additional data and rich contextual annotation not present in the published volumes, expanding both the empirical base and the transparency of the corpus.

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

4.4.0.2. Computational methodologies and NLP integration. One of the most significant advances introduced by the digital platform is the application of computational techniques to the annotated corpus. Leveraging the Italian glosses, the project employs a semi-automated process for morphosyntactic annotation: pretrained NLP models for Italian (e.g., SpaCy) generate POS tags, morphological features, and dependency structures for the glosses, which are then transferred to the dialectal data through a set of adaptation and alignment rules. Special procedures address elements unique to dialects, such as clitic pronouns or non-standard syntactic orders, ensuring faithful and informative annotation. This process not only enriches the corpus with structured linguistic information, but also provides a testbed for further computational experiments, such as fine-tuning LLMs on dialectal data.

4.4.0.3. Open access and interoperability. All core datasets are made available as Open Access resources on Zenodo [SMM*25], in CSV format with transparent metadata, enabling maximum scientific reuse and interoperability. The digital platform is explicitly designed to be scalable and interoperable, laying the groundwork for integration with other major dialectal corpora, as envisaged by the broader goals of the CHANGES project and in accordance with FAIR principles.

4.4.0.4. Outlook. By digitizing, annotating, and publishing the Manzini & Savoia corpus within a robust digital infrastructure, this initiative not only preserves an invaluable record of Italy's linguistic diversity but also provides a methodological template for future work in corpus linguistics, dialectology, and language technology. The project exemplifies how legacy fieldwork can be transformed—through careful curation and technological innovation—into a living resource at the service of both the scientific community and the wider public.

4.5. Cross-border Experimentation and Academic Networks

Thanks to its modularity and web-based delivery, *Small Codes* has also been used in research projects involving cross-border variation (e.g., Romansh dialects in Switzerland, Corsican-Italian dialects), allowing partners to co-develop resources while maintaining data sovereignty.

The platform facilitates:

- Parallel corpus alignment (e.g., dialect - gloss - standard language);
- Integration with audio and phonetic transcription tools (IPA-based);
- Collaboration between universities, linguistic societies, and citizen archivists.

These examples underline the versatility of the platform and its potential to serve as a replicable infrastructure for diverse linguistic ecologies—particularly in multilingual, decentralized, and geographically fragmented areas.

4.6. Vocabolario del Fiorentino Contemporaneo (VFC)

The Vocabolario del Fiorentino Contemporaneo [Bin18] is a project born in 1994 at the behest of Giovanni Nencioni, then President of the Accademia della Crusca. The idea, to which the editors have remained faithful over time, was to document all that area of the repertory in which nowadays there is a lack of, or partial, overlap between "Florentine lexicon" and "Italian lexicon."

With this in mind, detailed field research verifies the knowledge and use of an extensive corpus that has been compiled from a critical review of various types of historical and contemporary sources (local dictionaries, linguistic atlases, vernacular theatre). Among these, the Nòvo vocabolario della lingua italiana secondo l'uso di Firenze (generally known as 'Giorgini-Broglio', published between 1870 and 1897) stands out for the consistency of the glossary made available for field verification. As will be recalled, this was the result of Manzoni's language policy, which proposed spoken Florentine as the reference for spoken Italian.

In order to make the most of the vast material emerging from the surveys carried out in the area, the lexicographical entries have been organised into a database: in this way, the entries can serve as a starting point for wide-ranging research into contemporary Florentine spoken language. The testimonies of the speakers themselves are at the heart of the project: the core of each entry in the VFC is structured on the basis of the contexts in which the words were actually used.

Methodologically, the careful analysis of a wide and varied documentation, its organisation with a view to field surveys and, finally, the consideration of informants' testimonies as a source of lexicographical attestation tend to place the VFC within the framework of the most up-to-date experiences in dialect lexicography.

With *Small Codes*, the VFC project moves to the next stage: a real demo-anthropological atlas is created, geo-referencing on the territory of the historic city the impressive corpus of ethno-texts that has been the basis the VFC, with anecdotes, economic or cultural peculiarities related to places, through the living voice of the speakers. In this way, in a groundbreaking manner with respect to the traditional linguistic atlas, the new VFC mixes ethnographical / anthropological with purely linguistic data, hooking syntactic/morphological/lexical phenomena of contemporary Florentine to physical historic places. These short ethno-texts with audio synchronization have in addition to ethnographic and obviously linguistic, value, a so called 'third mission', very effective, purpose.

They have met with great favor among the general public, well far from the frequent folklorization attitude towards dialects, but rather with a strong grounding in the territory and an image of solid scientific basis. A pilot project of sharing this linguistic-ethnographic material through 'virtual hot-spots' on the territory is being experimented, in order to involve the very large audience of Italian tourists, advantaged by the fact that Florentine dialect ("vernacular"), unlike many other so-called "Italian dialects," is in fact intelligible to all Italians.

Vocabolario del Fiorentino Contemporane in brief:

- **starting corpus**: 20,000 entries (approx.);
- **cards published**: 3,670;
- **cards estimated**: 6,500;
- **speech recorded in verification surveys**: 250 hours (approx.), almost entirely transferred to MP3;
- **survey locations**: Santa Croce, Rifredi, San Frediano.

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

5. Interoperability and Open Standards

Ensuring interoperability across tools, data sources, and research communities is a core design principle of the *Small Codes* platform. In the domain of digital heritage and computational linguistics, long-term sustainability and reuse depend heavily on adherence to open standards and shared protocols for data representation, metadata encoding, and resource exchange.

5.1. Standards-Based Data Modeling

All linguistic resources within *Small Codes* are modeled following established international standards, enabling export, transformation, and federation with external repositories:

- **Lexical Markup Framework (LMF)** is adopted for lexicon representation, supporting structured encoding of headwords, variants, grammatical features, and multilingual mappings.
- **Text Encoding Initiative (TEI)** guidelines are used for the digitization of historical and dialectal texts, including critical editions, interlinear glosses, and aligned transcriptions.
- **ISOcat and GOLD** provide the basis for feature tagging and morphosyntactic annotation, ensuring semantic compatibility across corpora and languages.

These standards facilitate export in XML, JSON-LD, and RDF formats, making resources consumable by NLP tools, digital libraries, and semantic web applications.

5.2. Metadata and FAIR Principles

All datasets produced or managed by *Small Codes* follow the FAIR data principles (Findable, Accessible, Interoperable, Reusable). Metadata is exposed using Dublin Core, CMDI (CLARIN Component Metadata Infrastructure), and DCAT vocabularies to support harvesting, indexing, and integration with registries such as META-SHARE and Zenodo.

Each resource includes:

- Persistent identifiers (e.g., DOIs or ARKs);
- Machine-readable licensing (e.g., Creative Commons, Open Data Commons);
- Provenance and versioning metadata.

5.3. Integration with External Platforms

Small Codes is designed to serve not only as a stand-alone infrastructure but also as an interoperable node within larger digital ecosystems. Examples include:

- **Harvesting compatibility** with repositories such as CLARIN and DARIAH, via OAI-PMH endpoints;
- **API-level integration** with language technology platforms (e.g., LinguaLibre, SpaCy, UDPipe) for annotation and analysis;
- **GIS interoperability** via GeoJSON exports, enabling map-based overlays with other cultural datasets (e.g., UNESCO, GeoNames).

5.4. Licensing and Governance

All core components of the *Small Codes* infrastructure are distributed under permissive open-source licenses (e.g., MIT, GPLv3), while datasets are released under Creative Commons licenses where applicable. A governance model is in place to support community review of proposed data schemas, API extensions, and localization modules, encouraging transparent collaboration and standard evolution.

Through this standards-driven approach, *Small Codes* ensures that minority language resources are not isolated artifacts but part of a larger, durable, and machine-actionable web of linguistic and cultural knowledge.

6. Future Developments and LLM Integration

As the digital humanities and computational linguistics increasingly intersect with artificial intelligence, the *Small Codes* platform is actively evolving to embrace automation, large-scale data enrichment, and intelligent annotation techniques. The strategic integration of Large Language Models (LLMs) marks a significant shift in the system's development trajectory, offering new possibilities to reduce manual workload, accelerate corpus expansion, and deepen the analytical granularity of dialectally rich and under-resourced linguistic data.

This new phase builds on the recognition that while traditional NLP pipelines are valuable, they often fall short when applied to fragmented, morphosyntactically complex, and low-resource languages. LLMs—particularly transformer-based architectures—offer a scalable solution, provided that they are adapted and fine-tuned to respect the unique properties of regional and minority languages. Rather than replacing human expertise, these models are designed to work in tandem with community knowledge, fieldwork-driven corpora, and language-specific rules.

6.1. Automated Morphosyntactic Annotation

One of the most promising directions of development is the automation of morphosyntactic annotation for dialectal corpora. This task is particularly challenging due to the lack of pre-existing annotated resources and the syntactic non-conformity of many dialects when compared to standard language models.

To address these challenges, *Small Codes* is developing hybrid annotation pipelines that combine rule-based linguistic logic with machine learning techniques. The general strategy involves leveraging aligned glosses in majority languages (typically Italian) to infer morphosyntactic structures for dialectal texts. These glosses act as a bridge between under-resourced varieties and existing NLP resources.

A notable application of this approach is the use of legacy corpora such as the Manzini–Savoia collection, which includes thousands of dialectal examples transcribed in IPA and accompanied by Italian glosses. In this context, tools like SpaCy and UDPipe are used to annotate the gloss layer—generating POS tags, dependency relations, and morphological attributes. These annotations are then projected onto the dialectal tier via alignment heuristics, mapping procedures, and dialect-specific transformation rules.

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

This methodology is not limited to a single corpus, but is designed to be replicable and adaptable to other dialect-gloss corpora, whether legacy or newly developed. The current experimental pipelines being tested within *Small Codes* include:

- **Gloss-to-dialect alignment models**, which rely on word-level correspondences and contextual disambiguation;
- **Transfer learning** techniques that adapt Italian-trained models to dialectal inputs using minimal supervision;
- **Rule-based augmentation modules** to handle linguistic phenomena such as clitic duplication, pro-drop structures, and non-canonical syntactic orders;
- **Dialect-aware error correction**, capable of detecting and normalizing orthographic variants and morphophonological shifts specific to each variety.

These tools will significantly accelerate the annotation of dialect corpora, opening the way to more robust querying, visualization, and statistical analysis, and serving as a foundation for further tasks such as lemmatization, parsing, and search-based exploration of morphosyntactic microvariation.

6.2. Fine-Tuning and Custom LLMs

In parallel, *Small Codes* is preparing to implement fine-tuning procedures on open-source LLM architectures—such as BERT, RoBERTa, and BLOOM—using regionally specific data from fieldwork corpora and manually curated resources. The aim is to train dialect-aware language representations capable of handling high lexical variability, orthographic non-standardness, and syntactic heterogeneity.

Future applications of these adapted models will include:

- **Dialect-sensitive sentence embeddings** for similarity and clustering tasks;
- **Zero-shot and few-shot learning prompts** for fast and context-aware dictionary building;
- **Semantic variant prediction**, where the model proposes possible regional equivalents of a concept based on contextual input;
- **Domain adaptation**, enabling models to support tasks in schools, municipal communication, and cultural heritage documentation.

This shift toward customized LLMs is not intended to replace symbolic linguistic work, but rather to enhance it—serving as a scaffolding layer for structured linguistic databases and terminology banks that remain human-curated.

6.3. Crowdsourcing and Human-in-the-Loop Annotation

Despite the advances offered by LLMs, full automation is neither realistic nor desirable when dealing with minority languages. These languages often exhibit subtle and localized phenomena that require human insight and socio-linguistic sensitivity. For this reason, *Small Codes* will include a robust human-in-the-loop system that blends automated pre-annotation with user validation.

A dedicated web interface will allow native speakers, linguists, educators, and trained volunteers to review, correct, and enrich model-generated annotations. The system will feature intuitive mechanisms for:

- **Tag review and correction**, where users validate POS tags and morphological features;
- **Dialectal variant confirmation**, supporting crowd-curated lexical diversity;
- **Collaborative terminology creation**, integrating LLM suggestions into structured approval workflows.

This semi-supervised approach ensures that the knowledge encoded in the system reflects real usage and community expectations, while also improving training data for future iterations of the model.

6.4. Semantic Web and Knowledge Graph Integration

In the spirit of open science and linked data, *Small Codes* is also planning to integrate selected resources into the Linked Open Data (LOD) ecosystem. Lexical entries, neologisms, and terminological records will be transformed into RDF triples, aligned with external ontologies and repositories such as Wikidata, Glottolog, Lexvo, and the CLARIN Virtual Language Observatory.

By enabling cross-referencing between linguistic and cultural datasets, this integration will:

- **Facilitate semantic search and multilingual discovery** of minority language resources;
- **Enable integration with digital libraries, corpora, and cultural archives**;
- **Enhance visibility and reusability** of linguistic data in a broader heritage and research framework.

This transformation will allow *Small Codes* not only to function as a local documentation tool, but also as a node in a global knowledge infrastructure for intangible cultural heritage.

In summary, the future trajectory of *Small Codes* involves a careful synthesis of AI-enhanced language processing, participatory validation, and open data modeling. These developments will allow the platform to scale up, deepen its analytical capabilities, and offer new pathways for the preservation, accessibility, and enrichment of minority languages in the digital age.

7. Toward a Research and Community Portal

The roadmap also includes the development of a unified portal for scholars, institutions, and communities, featuring:

- Advanced search across dialects, corpora, and annotations;
- Visualization of historical linguistic change via temporal filtering;
- Aggregated dashboards for language vitality metrics, user feedback, and revision histories.

These developments mark *Small Codes*' transition from a project-based infrastructure to a stable, AI-augmented platform ecosystem that supports both scientific research and grassroots linguistic empowerment.

In order to have a compelling impact on the scholarly community and, more broadly, on the speech communities, this project will

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

make a decisive contribution to the development of a data management platform, which will be entirely dedicated to poorly documented varieties.

Such a platform will be configured as an open, long-lived digital and interoperable system for: - the linguistic web-gis - the analysis of 'atlantic' data for use by researchers, - the dissemination of usable materials of the general public

A good starting point (although outdated and to date without a stable development team) is the WALS platform mentioned above, whose code is open and available and which was developed in its time by a very good technical team.

This tool cannot be tied to a specific project because it must be able to survive beyond the project's time horizon and must meet the needs of multiple research groups. Therefore, the project intends to create a public-private partnership foundation with the goal of contributing to the development of open source software to serve the academic community like Praat or Elan.

Only by relying on such an entity can the impact of the project be maximized, and three critical issues that systematically plague academic language documentation and research digital projects be overcome: - 1 the absence of a culture of software reuse and broadening the user base, which instead are the only guarantees of long digital life - 2 the lack of access to the commercial software-house market, due to budget constraints, the inability to enter into multi-year maintenance contracts, and the absence of specialized expertise - 3 the absence, in the research world, of professional figures dedicated to digital communication, new media, and professional cultural marketing

This foundation will have a technical-engineering focus, although it will maintain a preferential relationship with the research world. Such a foundation, which can be configured as a university spin-off, will be nonprofit in nature, but will be immersed in the logic of market sustainability. The characteristics of such an enterprise are:

- totally private nature and therefore total freedom of management, allocation of financial and human resources;

- has technical staff from industry: software development and (digital) marketing are technical trades and must be led by experienced, full-time professionals, preferably well-paid and therefore retainable.

8. Conclusions, Outlook and dissemination 'business' model

This paper has presented the design rationale, architectural foundations, and application scenarios of *Small Codes*—a modular, standards-compliant digital infrastructure developed to support the documentation, analysis, and revitalization of minority languages and dialects.

UUlike traditional language technology pipelines, *Small Codes* has been conceived with the specific needs of lesser-used and geographically fragmented language varieties in mind. Its hybrid model, combining academic rigor with production-grade engineering, enables it to serve both as a research tool and as a long-term service platform for communities, institutions, and educators.

Through real-world deployments across diverse language contexts—ranging from Ladin and Sardinian to Francoprovençal and Occitan—the platform has demonstrated its ability to handle complex lexicographic variation, user-driven terminology creation, geospatial visualization, and multimedia integration. Its commitment to open standards and FAIR data principles ensures that these resources remain accessible, reusable, and interoperable within broader digital heritage ecosystems.

Looking ahead, *Small Codes* is evolving into a next-generation linguistic infrastructure by integrating LLM-based annotation, semantic enrichment, and community-sourced validation mechanisms. These innovations will further extend its capabilities while preserving the integrity and specificity of each language context.

By bridging the gap between digital linguistics, AI, and cultural heritage, *Small Codes* represents a scalable, sustainable response to the growing urgency of language loss in the digital age. It provides not only a technological toolkit, but also a methodological and ethical framework for inclusive and responsible digital language work.

Small Codes will allow a partial decoupling of technical choices from the needs of the individual research project that will use them. 'Partial' because each client of the various research groups will certainly have specific needs: but it will be up to the technical coordination of the spinoff to figure out how to integrate the various needs into the final product, which will be configurable, open, but also sufficiently multi-purpose and not tailor made each time, on pain of certain failure of the initiative.

Small Codes will combine software/NLP/web-GIS development skills and communications agency skills, or at least coordinate them. The promotion of research results, their publication even for the general public, dissemination and popularization initiatives must be coordinated from the beginning with the software tools used for data manipulation: a dialect geolinguistic map, e.g., intended for specialists cannot disregard, from the very beginning of its conception, the design of the 'public' and popularization version of that same type of maps, nor the design of the dissemination campaign on social media, etc. This does not mean taking away rigor from research (which remains the center of the Academy's interest), but getting out of a logic of subalternity of the third mission; a subalternity that, if maintained, would reduce even more the visibility, and the available funds, for these already underfunded fields of research.

Small Codes will have mixed business model. Reaffirming its private nature, the spinoff will set itself up as a nonprofit entity, in the market, and with total freedom to price its services: the patrons, consisting not only of research groups in the digital humanities (particularly in the field of (geo-)linguistics and that of documentation and preservation of under-resourced languages) will have very different financial availability from year to year and from project to project: a guarantee of the long life of the initiative is the release of the single technical realization from the funds theoretically needed to develop it. The logic must be that of a unified project that is sustainable over time, not of many projects each in itself in financial balance, what would simply be impossible. A small-budget research group must be able to use a complex tool; a well-funded research group can contribute to the realization of features even beyond those that would be strictly necessary in that case.

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.

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.

Credit Author Statement

Greta Mazzaggio and Carlo Zoli drafted the manuscript. Neri Binazzi revised the manuscript.

References

[Bin18] BINAZZI N.: Parole che diventano storia. l'esperienza del "vocabolario del fiorentino contemporaneo" (vfc). *Etimologia e storia delle parole. XII Convegno ASLI - Associazione per la Storia della Lingua Italiana, Firenze, 3-5 novembre 2016, Franco Cesati* (2018), 195–206. 5

[Klo67] KLOSS H.: Abstand languages and ausbau languages. *Anthropological Linguistics 9, H. 7* (1967), 29–41. 2

[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. 2, 4

[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. 2, 4

[MS05] MANZINI M. R., SAVOIA L. M.: *I dialetti italiani e romanci: Morfosintassi generativa*. Edizioni dell'Orso, 2005. 2, 4

[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, 4, 5

[ZSM13] ZOLI C., SORIA C., MARIANI J.: Dwarfs sitting on giants' shoulders: How lts for regional and minority languages can benefit from piggybacking major languages. *FEL XVII: Endangered Languages Beyond Boundaries* (2013). 1

© 2025 The Author(s).
Proceedings published by Eurographics - The European Association for Computer Graphics.