An Enhanced Pipeline for the Manzini-Savoia Dialect Corpus

Achille Fusco[1,2], **Greta Mazzaggio**[1,3], **Carlo Zoli**[4,5]
[1]University of Florence, DILEF, Via della Pergola, 60 50121 Florence
[2]IUSS Pavia, NeTS Lab, Piazza della Vittoria, 15 27100 Pavia
[3]University of Enna "Kore", Cittadella Universitaria 94100 Enna
[4]Free University of Bozen-Bolzano, Universitätsplatz 1 - Piazza Università, 1 39100 Bozen-Bolzano
[5]University of Verona, Via dell'Artigliere, 8 37129 Verona
achille.fusco@iusspavia.it, greta.mazzaggio@unikore.it, carlo.zoli@smallcodes.com

Abstract

This paper presents a semi-automatic workflow for enriching the Manzini–Savoia Corpus (MSC) of Italian dialects with extended glosses, normalized transcriptions, and projected morpho-syntactic annotations. While the MSC is a unique resource for Romance microvariation, its partial glossing and phonetic transcription in the International Phonetic Alphabet (IPA) pose major challenges for computational processing. We introduce a pipeline for gloss coverage expansion and reliable morpho-syntactic annotation combining rule-based and data-driven components, which includes: (i) automatic completion of truncated verbal paradigms; (ii) hybrid lexical alignment between dialectal tokens and Italian glosses, integrating per-region lexical priors with a dynamic programming alignment algorithm; and (iii) projection-based morpho-syntactic tagging from aligned glosses. The proposed methods offer a reproducible framework for extending partially glossed dialect corpora and contribute new annotated data for research in computational dialectology and cross-variety language modeling.

Keywords: low-resource languages, annotation projection, text-gloss alignment

1. Introduction

Building computational tools for under-represented languages and dialects remains one of the most pressing challenges in natural language processing (Joshi et al., 2020; Lai et al., 2023). Many varieties, particularly spoken dialects, endangered languages, and historically under-documented forms, lack annotated corpora, standard orthographies, or even basic tokenization conventions (Blasi et al., 2022). At least half of the world's languages are under-documented (Bird, 2011), and over 2500 are currently endangered (Moseley, 2010). Estimates suggest that up to 90 percent of the world's languages are at risk of disappearing within the next century (Krauss, 1992). Language loss is driven by multiple factors, including limited available resources, sociopolitical pressure from dominant languages, and government policies that undermine linguistic diversity (Bromham et al., 2022). Linguistic documentation has thus become an urgent priority, not only for scientific research but also for enabling future revitalization efforts (Crowley, 2007).

A central format in language documentation is Interlinear Glossed Text (IGT), which represents linguistic data using aligned layers of transcription, morpheme segmentation, grammatical annotation, and word-by-word translation. While IGT is an invaluable tool for linguistic analysis and revitalization work, its manual creation is extremely time- and labor-intensive. This has led to growing interest in applying machine learning to automate key steps such as segmentation and glossing (Moeller et al., 2018; Yang et al., 2024; Elsner and Liu, 2025). However, the development of such tools is constrained by the scarcity of annotated training data and the lack of standardized formats for many low-resource and dialectal varieties. This situation further deepens the digital divide between majority and minority languages.

In this work, we address this issue by leveraging the **Manzini–Savoia Corpus (MSC)** (Mazzaggio et al., 2025), a comprehensive digital corpus of Italian, Romansch, and Corsican dialects derived from Manzini and Savoia (2005). The resource comprises more than 60,000 sentences elicited from native speakers through extensive fieldwork. A unique feature of this corpus is that each sentence is presented in International Phonetic Alphabet (IPA) and, in most cases, paired with an Italian gloss, making it one of the most valuable resources for microvariation in Romance syntax and morphology. However, while typical IGTs provide detailed morphological segmentation, Manzini and Savoia (2005) adopted a more compact system consisting mainly of literal Italian translations and, where needed, short morpho-syntactic tags (e.g., `Cls` for subject clitics). A substantial subset of the MSC remains without glosses—particularly inflectional paradigms glossed only in the first person singular (e.g., "io dormo, etc."). Moreover, glosses and transcriptions often diverge in token count or structure, complicating token-level annotation projection. These limitations hinder the automatic en-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3379–3393
11–16 May 2026.

Field	Value
index	158
official_placename	Santa Maria di Sala
provincia	Venezia
regione	Veneto
chap	2.1 Presentazione dei dati relativi al clitico soggetto.
text	vj ŋ me fi i
gloss	viene i miei figli
text_preprocessed	vj ŋ me fi i
gloss_preprocessed	viene i miei figli
text_to_grapheme	vien me fioi
tagging	[{'text': 'vj ŋ', 'gloss': 'viene', 'pos': 'VERB', 'morph': {'Mood': 'Ind', 'Number': 'Sing', 'Person': '3', 'Tense': 'Pres', 'VerbForm': 'Fin'}}, {'text': 'me', 'gloss': 'miei', 'pos': 'DET', 'morph': {'Gender': 'Masc', 'Number': 'Plur', 'Poss': 'Yes', 'PronType': 'Prs'}}, {'text': 'fi i', 'gloss': 'figli', 'pos': 'NOUN', 'morph': {'Gender': 'Masc', 'Number': 'Plur'}}]

Table 1: Example corpus entry with fields displayed vertically. Rows highlighted in grey correspond to fields automatically generated by the processing pipeline.

richment and computational exploitation of the corpus.

To address these issues, this paper presents an extended and reproducible pipeline for **automatic annotation projection and corpus enrichment** in the MSC. The approach combines linguistically informed preprocessing, rule-based expansion, and alignment-driven annotation to improve both coverage and structural consistency. The workflow consists of five main stages:

1. **Data normalization:** systematic cleaning of textual content, including removal of metalinguistic parentheses.

2. **First-pass annotation projection:** initial transfer of part-of-speech and morphological features from Italian glosses to dialectal tokens using direct token alignment.

3. **Paradigm completion:** automatic expansion of truncated verbal paradigms (e.g., “io dormo, etc.”) by detecting first-person singular verbs and projecting the full inflectional pattern.

4. **Hybrid lexical alignment:** construction of frequency-based dialect–Italian lexica and implementation of a monotonic alignment algorithm integrating lexical correspondences and token-level similarity.

5. **Second-pass annotation projection:** reapplication of morpho-syntactic tagging on the aligned data, refining the initial projection and improving coverage for unglossed or newly generated forms.

To illustrate the structure of the corpus and the transformations introduced by our pipeline, Table 1 shows an example entry with the original fields and the additional fields generated during preprocessing and annotation.

The resulting system represents the first automatic projection of morpho-syntactic annotations for the MSC, increasing both the proportion of glossed entries and the overall internal consistency of the dataset. Additionally, this paper contributes both a methodological workflow and an enriched dataset that can serve as a foundation for future research in computational dialectology, morphological modeling, and cross-variety annotation transfer.

The rest of the paper is organized as follows: Section 2 reviews related work on annotation projection and dialectal NLP; Section 3 details preprocessing and normalization steps; Section 4 discusses paradigm completion; Section 5 presents hybrid lexical alignment and second-pass projection; Section 6 reports evaluation results and provides the discussion and conclusions.

2. Related Work

2.1. Resources on Romance Dialects

A growing number of resources document the linguistic diversity of Romance dialects, integrating traditional fieldwork with digital infrastructures. Within the Italian area, the *Atlante Italo-Svizzero* (AIS) and its digital extensions (AISr, Loporcaro et al. 2021) provide lexical and phonetic data across hundreds of localities, while the *Atlante Linguistico del Ladino Dolomitico* (ALD, Rührlinger

2004) focuses on Ladin and neighboring dialects. The *Atlante Sintattico d'Italia* (ASIt, Pescarini and Di Nunzio 2010) complements these by systematically investigating syntactic microvariation across more than 200 Italian varieties through structured elicitation and judgment data. Beyond Italy, the *Atlas Lingüístico de la Península Ibérica* (ALPI, García Mouton 2017) and the *Corpus Oral y Sonoro del Español Rural* (COSER, Fernández-Ordóñez 2011) make available geographically distributed corpora of Iberian Romance varieties. For Portuguese, the CORDIAL-SIN corpus (Carrilho, 2010) offers syntactically annotated transcriptions of dialectal speech. The **MSC** contributes to this landscape by providing systematically glossed IPA transcriptions of Italo-Romance varieties, offering a uniquely fine-grained basis for automatic alignment and annotation projection.

2.2. Annotation projection

Annotation projection has been a foundational approach in cross-lingual NLP to bootstrap models for low-resource languages using aligned corpora. Early work by Yarowsky and Ngai (2001) used robust projection to induce POS taggers for French and other languages — a method shown to be surprisingly effective even with noisy alignments, often achieving 96% POS accuracy in target languages. Huck et al. (2019) demonstrated cross-lingual annotation projection for Universal Dependencies using parallel data between high- and low-resource languages, focusing on alignment quality and typological generalization. A related line of work uses Bible translations as a parallel corpus for multilingual projection and transfer (e.g., (Agić et al., 2015)), taking advantage of their broad language coverage and consistent structure. Projection techniques have been extended to parsing (McDonald et al., 2013), NER (Wang and Manning, 2014), and semantic role labeling (Padó and Lapata, 2005), solidifying projection as a standard tool in multilingual NLP. In our case, rather than projecting across languages, we exploit gloss alignments between dialectal IPA forms and Italian glosses to derive indirect supervision in the absence of standard resources.

In addition to projection-based approaches, recent work has explored learning-based methods for generating glosses and morphological annotations automatically, which are reviewed in the following subsection.

2.3. Automatic glossing and morphological generation

A growing body of work has explored computational approaches to automatic glossing and morphological annotation for low-resource languages. Moeller et al. (2018) investigate automatic glossing using sequence-to-sequence models trained on interlinear glossed text. Their approach treats glossing as a supervised learning task, where gloss tokens are predicted directly from the input sentence using neural architectures. While effective when annotated training data are available, such methods require sufficiently large aligned corpora in order to learn reliable mappings between text and glosses.

More recently, Yang et al. (2024) examine multilingual and cross-lingual strategies for gloss generation using large language models. Their work shows that prompting and retrieval-based methods can improve gloss prediction across typologically diverse languages, particularly in settings where annotated data are scarce. However, these approaches still assume the availability of representative glossed examples that can be used either for training or prompting. In a similar line of research, Elsner and Liu (2025) investigate prompting-based strategies for automatic glossing within the SIGMORPHON shared task framework (Nicolai et al., 2025), demonstrating that large language models can produce high-quality gloss suggestions that assist human annotators during language documentation.

Our work differs from these approaches in several important respects. First, while previous studies focus on learning-based gloss generation, our method does not attempt to predict glosses from raw text. Instead, it exploits the structural properties of the Manzini–Savoia corpus and the presence of existing Italian glosses to derive morpho-syntactic annotations through deterministic processing steps. Second, the data considered here are substantially more fine-grained: the Manzini–Savoia corpus comprises sentences collected from more than 400 Romance microvarieties. While this provides exceptionally detailed coverage of dialectal variation, each variety is represented by only a limited number of examples, which poses a challenge for data-hungry learning-based approaches. Finally, unlike many automatic glossing systems that operate across typologically unrelated languages, our approach leverages Italian as a reference language closely related to all the dialectal varieties in the corpus. This linguistic proximity allows us to combine rule-based normalization, alignment, and annotation projection to obtain morpho-syntactic analyses without requiring supervised training data.

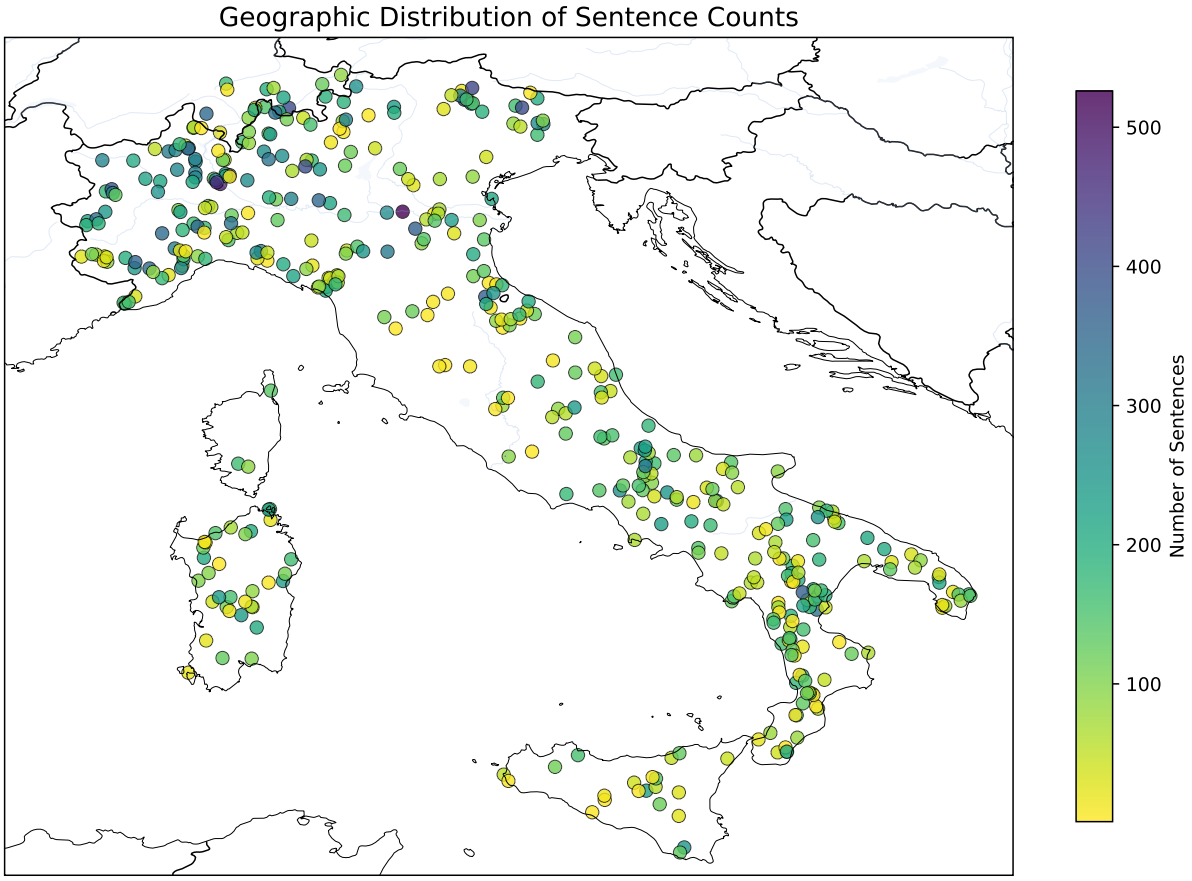

Figure 1: Geographical distribution of sentence counts.

3. Corpus Description and Preprocessing

3.1. Corpus Overview

The Manzini & Savoia corpus (Savoia et al., 2025), based on the monograph by Manzini and Savoia (2005), is one of the most comprehensive resources for the study of morphosyntactic variation in Romance dialects, specifically Italian, Corsican, and Romansh varieties. The original monograph was compiled through extensive fieldwork conducted over several decades by Leonardo Maria Savoia, following a generative-grammar approach that prioritized syntactic judgment and linguistic competence over areal mapping. The corpus collects and digitizes the dialectal examples reported in that work.

The corpus is stored in CSV format and includes data from **487 dialect varieties**, broken down as follows:

- **459** varieties from across the Italian peninsula;
- **9** Corsican varieties;
- **19** varieties from Southern Switzerland.

Each entry in the corpus comprises:

- a dialectal sentence transcribed in **IPA**, typically marking primary stress on content words only;
- a **word-by-word Italian gloss**;

The full dataset currently comprises **64,472 examples**, each associated with its locality of origin and morphosyntactic category, for a total of **253,736** word tokens. The examples also make reference to syntactic phenomena (e.g., clitics and auxiliary selection), reflecting the theoretical orientation of the original work.

At the level of microvarieties, representativeness is quite diverse: the number of documented sentences for each single microvariety ranges from only **1** to **526**, with an average of approximately **129** examples per microvariety. Geographical distribution of the sentences is shown in Figure 1.

This corpus is being digitized and made available within the framework of **Project CHANGES (Cultural Heritage Active Innovation for Sustainable Society)**, a national initiative funded by

Pipeline Step	Match	Mismatch	No-gloss
Step 1: raw preprocessed	31,684 (50.5%)	14,696 (23.4%)	16,395 (26.1%)
Step 2: cleaned/normalized	32,292 (51.4%)	14,088 (22.4%)	16,395 (26.1%)
Step 3: with paradigm fill	38,465 (61.3%)	15,143 (24.1%)	9,167 (14.6%)

Table 2: Match, mismatch, and no-gloss statistics across pipeline steps. Each cell shows raw count (top) and percentage (bottom).

the *Ministry of University and Research* and the *European Union* through the *NextGenerationEU program*, as part of the *Piano Nazionale di Ripresa e Resilienza* (PNRR) (Mazzaggio et al., 2023, 2025; Mazzaggio and Binazzi, 2024).

CHANGES promotes sustainable, open-access strategies for the preservation and valorization of intangible cultural heritage. In this context, the digitization of the Manzini & Savoia corpus responds to the urgent need for systematic documentation of dialectal variation at risk of disappearance (Zoli et al., 2025). The full dataset is openly released on Zenodo (Savoia et al., 2025), and an interactive web platform supports multi-modal exploration of the corpus by syntax, geography, and locality.

3.2. Cleaning

The annotation projection procedure requires a consistent token-level alignment between the dialectal text and its Italian gloss. Because morpho-syntactic information is projected from the gloss to the dialectal tokens, any mismatch in tokenization or segmentation can lead to the loss or duplication of features. Preprocessing is therefore a critical step in ensuring that each dialectal token can be associated with the appropriate gloss element, either through one-to-one alignment or through more flexible matching rules.

To quantify the degree of mismatch, we first computed token counts for both sides of each sentence. Out of roughly 60,000 dialect–gloss pairs, about 50% showed a different number of tokens, reflecting various forms of structural disalignment. Some mismatches arise from asymmetric translation between the dialectal text and the Italian gloss, where a single dialectal token may correspond to multiple Italian words or vice versa. For example, in *la Ma'ria la 'lɛ:ʃ* vs. *Maria ClS-3sf legge*, the definite article *la* preceding the proper name is omitted in Italian, since standard Italian normally does not allow this construction. Additional discrepancies stem from parenthetical notes and metalinguistic comments embedded in the glosses. These may specify grammatical information, such as tense like "*(Pres. Cong.)*" ('present subjunctive'), or offer alternative renderings that conform more closely to standard Italian usage, as in *è* (=*ha*) *mangiato*.

To mitigate such discrepancies that interfere with text–gloss alignment, we systematically removed parenthetical segments containing metalinguistic annotations or explanatory paraphrases through an automatic preprocessing step. For example, "*non ti sieda (2ps Pres. Cong.)*" was simplified to "*non ti sieda*". In cases such as "*lo sono (=ho) chiamato*", we retained the surface form in order to preserve token-level alignment consistency, removing the parenthetical element and keeping "*lo sono chiamato*". Operationally, both preprocessing decisions amounted to deleting all parentheses and their contents from the gloss.

After cleaning, we recomputed token counts and alignment diagnostics, which showed only a minimal net improvement in alignment: the share of matched pairs increased by approximately 1% (see Table 3.1), however, this preprocessing step ensured a more reliable base for subsequent alignment.

A remaining challenge concerns alternative glosses separated by slashes, such as *al fa / fae* ('does'), which represent optional lexical or morphological variants. Such alternation may consist of single or multiple words, with some cases even displaying alternations among options of different number of tokens. Therefore, since these sets of alternations lack explicit delimiters, they could not be automatically disambiguated. Their resolution would require explicit human markup to identify the beginning and end of alternative sets.

4. First-Pass Projection

Before performing any alignment between dialectal text and Italian gloss, a first-pass annotation projection was carried out to provide preliminary morpho-syntactic information for each token. This step serves two main purposes: it establishes a baseline annotation layer for the corpus and provides the linguistic cues necessary for detecting verbal paradigms during gloss completion.

The first-pass projection combines automatic part-of-speech (POS) tagging and morphological analysis of the Italian gloss with a set of hand-crafted templates for dialect-specific clitics and auxiliaries. As already explained in section 3.1, the corpus consists of dialectal utterances transcribed in IPA, with word-level alignments to Italian glosses. Each word-level gloss consists of either a literal Italian translation or a specialized tag-like label indicating its grammatical function (e.g., `ClS-3sm`, `Neg`, etc.).

Accordingly, our projection method distinguishes between two types of glosses:

- **Standard Italian glosses** were annotated automatically using the `spaCy` model for Italian `it_core_news_lg` (Explosion, 2024). This model, trained on the Italian Universal Dependency Treebank (Bosco et al., 2013), provided POS tags and morphological features according to the Universal Dependencies (UD) annotation scheme (De Marneffe and Manning, 2008; De Marneffe et al., 2014).
- **Non-standard or functional glosses**, involving clitics, negative markers or suffixes, were annotated using a hand-written set of annotation rules. These rules map label strings (e.g., `ClS-3sm`) to structured POS tags and morphosyntactic features consistent with the Universal Dependencies framework.

The rule-based component ensured consistent treatment of grammatical elements not easily handled by standard taggers, such as subject clitics and polarity markers. An excerpt of the rule set is shown in Table 4 (see Appendix A for a more comprehensive rules set).

The final annotation for each gloss consisted of:

- the original gloss string (standard or tag-like),
- the assigned POS tag,
- a set of morphological features.

While this initial projection is not yet informed by cross-token alignment or lexical similarity, it provides a useful baseline layer of linguistic information. In particular, it enables the automatic identification of first-person singular (1PS) verb forms, which are required for the paradigm completion step described below.

5. Paradigm Completion

Roughly 26% of corpus entries lack an explicit gloss, the majority of which correspond to verbal paradigms in which only the first-person singular form is translated, followed by the shorthand marker “, etc.” or “, ecc.” For example, a typical gloss such as *ClS dormo, etc.* (‘I sleep, etc.’) introduces a full set of inflectional forms whose glosses are left implicit. Completing these paradigms automatically is essential to increase the coverage and internal consistency of the resource.

Paradigm completion is implemented as a rule-based procedure operating on the tabular structure of the corpus (see Appendix C for a more detailed description of the procedure adopted). Rows whose gloss ends in “etc.” or “ecc.” and that are followed by five empty gloss rows are treated as paradigm starters. Using the morpho-syntactic annotation produced by the first-pass projection (see Section 4), the system identifies tokens corresponding to finite first-person singular verbs within the starter gloss.

The starter line is then used as a template: the trailing “etc.” marker is removed and the token sequence is copied to generate the remaining five rows of the paradigm. Verb tokens are replaced with the corresponding inflected forms retrieved from the Morph-it lexicon (Zanchetta and Baroni, 2005), while subject pronouns, clitic pronouns, and enclitic forms are adjusted to match the target person-number values (2SG–3PL). Additional heuristics handle composite forms with the auxiliary *essere* ‘be’ and minor lexical irregularities. Finally, the modified tokens are recombined to produce the completed gloss lines. The newly generated glosses are then inserted into the corresponding empty rows, producing a complete inflectional paradigm. The resulting column, `gloss_filled_par`, extends gloss coverage by approximately **12%** of the corpus, producing a more homogeneous dataset for subsequent alignment and projection. An overview of the progressive gloss coverage and alignment across the different steps of the pipeline can be viewed in Table 3.1.

By enriching the corpus with consistent paradigmatic information, this step improves the availability of aligned examples for learning dialect–Italian correspondences and enhances the overall reliability of automatic annotation.

6. Hybrid Lexical Alignment

6.1. Motivation

After completing the paradigm expansion step, the next challenge concerns the alignment of dialectal tokens and their Italian glosses. Due to the phonetic nature of the dialect transcriptions and the morphological variability across varieties, one-to-one alignment cannot be assumed even after careful preprocessing. Simple string-based similarity is often insufficient: dialectal forms like /ˈpɔpi/ (‘children’) from their Italian counterparts (*bam-*

Gloss	POS	Morph Features
`ClS-3sf`	PRON	Clitic=Yes, Gender=Fem, Number=Sing, Person=3, PronType=Prs, Case=Nom
`ClS-2p`	PRON	Clitic=Yes, Number=Plur, Person=2, PronType=Prs, Case=Nom
`Neg`	ADV	PronType=Neg
`-1pp`	VERB	Person=1, Number=Plur, VerbForm=Inf

Table 3: Examples of gloss-label to tag projection rules

bini), both orthographically and phonetically. Conversely, purely frequency-based or statistical alignment methods risk producing uninterpretable pairings when working with small or morphologically rich datasets.

To overcome these limitations, we adopt a **hybrid alignment approach** that integrates lexical knowledge derived from the corpus with character-based similarity metrics. This method increases robustness to surface variation while maintaining interpretability, as each alignment can be explained either by a previously attested lexical pairing or by measurable orthographic similarity. The resulting alignments serve as the basis for the refined annotation projection described in Section 7.

6.2. Orthographic Normalization

Dialectal transcriptions in the MSC are provided in the International Phonetic Alphabet (IPA). While phonetically precise, this representation increases the surface distance between dialectal forms and their Italian glosses, as the two strings are encoded using different symbol systems. This reduces the effectiveness of character-level similarity measures used in the alignment step (see Section 6.4). For example, the IPA dialectal form 'dɔpu ('later') is string-wise more distant from its gloss *dopo*, sharing only two characters (*d* and *p*), whereas a grapheme transcription such as *dopu* would share three (*d*, *o*, and *p*).

To reduce this distance, we implemented a rule-based normalization procedure that maps IPA sequences onto an approximate graphemic representation. The conversion rules were manually defined, targeting systematic correspondences such as /ʦ/ → *z*, /ʎ/ → *gli*, and accounting for context-sensitive transcriptions (e.g. the IPA character /k/ is transcribed as *ch* before /e/, /ɛ/, or /i/, and as *c* in all other cases). These mappings were applied deterministically, producing a normalized dialectal column (`text_to_grapheme`) used for subsequent alignment.

Importantly, this procedure does not aim to reconstruct an accurate orthographic representation of each dialect. Rather, it provides a simplified graphemic approximation designed to increase string comparability with Italian glosses. For this reason, the rules follow basic Italian orthographic conventions and favor cross-dialect consistency over orthographic correctness.

Although normalization inevitably introduces a degree of abstraction away from the phonetic detail of the original transcription, it facilitates the transfer of morpho-syntactic information by aligning the dialectal forms more closely with the Italian lexical space, without erasing dialect-specific contrasts.

6.3. Parallel Lexicon Construction

The first step consists in extracting reliable token pairs from those rows in which the number of tokens in the dialectal text and in the Italian gloss is identical. In such cases, a direct one-to-one correspondence can be safely assumed. Each token pair (w_d, w_i), where w_d is a dialectal token and w_i its gloss counterpart, is stored in a bilingual lexicon together with its frequency of occurrence.

To capture systematic correspondences that are characteristic of specific linguistic areas, separate bilingual lexica are built for each administrative region represented in the corpus (e.g. Veneto, Puglia, Sardegna). This regional conditioning allows the alignment algorithm to favor lexically attested matches typical of the variety spoken in the corresponding region, while still retaining a global bilingual lexicon that provides fallback mappings for tokens not attested regionally.

6.4. Alignment Algorithm

Lexical priors The aligner consults two probabilistic lexica built from length-matched pairs: a *per-Regione* lexicon and a *global* one. Each entry stores frequency counts and the conditional probabilities $P(\text{gloss} \mid \text{text})$, $P(\text{text} \mid \text{gloss})$, and PMI. An entry is considered reliable when these values exceed small thresholds (τ_{gt}=0.2, τ_{tg}=0.1, $\tau_{\text{pmi}} \geq 0$). The higher threshold for $P(\text{gloss} \mid \text{text})$ reflects the fact that a dialect token typically maps to a small number of Italian glosses, whereas the reverse direction is more ambiguous due to the large number of dialectal variants corresponding to the same

	Alignment		POS		Morph	
Condition	Sent.	Tok.	Sent.	Tok.	Sent.	Tok.
Token match	1.00	1.00	0.78	0.91	0.74	0.87
Token mismatch	0.87	0.94	0.61	0.80	0.59	0.78
Total	0.94	0.98	0.70	0.86	0.67	0.83

Table 4: Evaluation results for alignment, POS tagging, and morphological tagging at sentence and token level. Sentence-level accuracy requires all tokens in the sentence to be correct.

gloss. Regional evidence is preferred over global evidence.

Pairwise scoring For every candidate pair (w_d, w_i), the system computes a weighted alignment score. If a reliable entry is found in either the regional or the global lexicon, the score combines a base weight (B_{reg}=10 or B_{glob}=6) with log-scaled bonuses for $P(\text{gloss} \mid \text{text})$, $P(\text{text} \mid \text{gloss})$, and the pointwise mutual information (PMI). Regional matches receive a higher base weight, reflecting the expectation that dialect–Italian correspondences are often locality-specific.

$$\text{PMI}(t_d, t_i) = \log_2 \frac{P(t_d, t_i)}{P(t_d)\,P(t_i)}$$

If no reliable lexicon entry is available, the system falls back on a similarity-based score. This score combines scaled character similarity (RapidFuzz ratio), a small bonus for comparable token length, and a fixed reward for exact matches.

This scoring scheme ensures that high-confidence lexical correspondences dominate the alignment whenever available, while still allowing plausible matches for previously unseen token pairs.

Monotonic dynamic programming Since we are working with glosses rather than free translations, we assume that the relative order of corresponding tokens is preserved between the dialectal text and the gloss. We therefore adopt a monotonic dynamic programming approach that aligns each token in the shorter sequence to a token in the longer one, allowing unmatched tokens on the longer side to be skipped.

Let $D = [d_1, \ldots, d_n]$ be dialect tokens and $I = [i_1, \ldots, i_m]$ Italian tokens. The algorithm aligns the shorter sequence to the longer one using a monotonic dynamic program:

$$\text{DP}(i,j) = \max\Big(\text{DP}(i, j{-}1), \text{DP}(i{-}1, j{-}1) + \text{score}(s_i, \ell_j)\Big) \quad (1)$$

with $\text{DP}(0, j)$=0.

Backtracking over the dynamic programming table yields the optimal alignment path together with its total score. For each sentence pair, the algorithm finally returns the set of aligned token pairs and the corresponding alignment score.

7. Second-Pass Tagging

7.1. Motivation and Method

The alignment procedure described in the previous section enables a more precise mapping between dialectal and Italian tokens. While the first-pass tagging relied on a direct projection from the gloss line to the dialectal text, this projection was necessarily limited by token mismatches and by the presence of unaligned or partially glossed forms. The hybrid alignment stage provides a refined correspondence that allows for a *second-pass projection* of linguistic information, improving both coverage and consistency.

In this second pass, the part-of-speech (POS) and morphological features initially obtained from the Italian gloss (via `spaCy`) are re-evaluated and reassigned to the corresponding dialectal tokens according to the alignment pairs.

To ensure interpretability, the projected morphological information preserves the `spaCy` format, storing both POS and morphological features in the `text_tagging_aligned` column. Each token is represented as a structured dictionary, maintaining compatibility with downstream NLP tools and allowing manual or automatic quality checks.

This step effectively bridges the gap between the first-pass coarse tagging and a more linguistically informed annotation grounded in the alignment results. It also creates a foundation for future semi-supervised refinement: once evaluated, these aligned projections could serve as silver-standard training data for region-specific taggers.

7.2. Evaluation

The output of the second-pass tagging was evaluated on a manually verified sample of 200 entries from the corpus. The sample included both easy and difficult cases: 100 text–gloss pairs with

Text & Gloss	POS tagging (output)	Comments
ɛŋ s erɒ / 'hiɛm la'atʃ *ci eravamo lavati*	PRON: ɛŋ → ci AUX: erɒ → eravamo VERB: la'atʃ → lavati	Although the sentences were token-mismatched, the algorithm correctly mapped one of the two alternatives separated by the slash (erɒ / 'hiɛm) to the correct auxiliary verb *eravamo*.
ratə-'m-illu / -illa / -illəə *date-ce-lo / la / li-le*	NULL: ratə-'m-illu → date-ce-lo NULL: -illa → la NULL: -illəə → li-le	Alignment succeeds, but subword segmentation with hyphens and slash-separated alternatives hinders the tagger's ability to correctly identify and tag the Italian word.

Table 5: Examples of successful and unsuccessful alignment and tagging.

matched token counts and 100 pairs with mismatched counts. Although mismatched pairs represent only about 24% of the total data, they were oversampled to better assess performance under challenging alignment conditions.

Each token in the selected subset was manually checked for correctness of the dialect–gloss pairing, part-of-speech (POS) assignment, and morphological tagging. Since the annotation involved verifying deterministic correspondences between dialect tokens, gloss tokens and projected tags, rather than performing interpretive linguistic judgments, the task was largely unambiguous and was therefore carried out by a single annotator (i.e., the first author). Accuracy was computed both at the sentence level and at the token level. Sentence-level accuracy corresponds to the proportion of sentences in which all projected values were correct, such that a single incorrect tag rendered the entire sentence tagging incorrect. Token-level accuracy instead measures the proportion of individual tokens correctly aligned and tagged.

The results indicate high alignment accuracy (94% at the sentence level and 98% at the token level), while POS and morphological tagging achieved 70% and 67% at the sentence level, and 86% and 83% at the token level, respectively (see Table 4). The main sources of residual error include ambiguous or underspecified glosses and cases involving alternative forms separated by slashes ("/") or hyphens ("-") used to segment clitics or inflectional material. These inconsistencies suggest that further orthographic normalization and finer-grained treatment of compound glosses could improve projection reliability.

Illustrative examples of correctly and incorrectly tagged items are reported in Table 5, highlighting both successful handling of complex verb forms and remaining sources of variability.

8. Discussion and Conclusion

In this work, we presented a pipeline of normalization, paradigm completion, and alignment-based projection that substantially improves the coverage and internal consistency of the Manzini–Savoia corpus (MSC). The achieved alignment accuracy (94% at the sentence level and 98% at the token level) shows that the hybrid approach reliably captures token-level correspondences between dialectal and Italian forms despite substantial orthographic and morphological variation.

POS and morphological tagging achieve lower sentence-level accuracy (70% and 67%), largely reflecting noise and inconsistencies in the gloss layer. Token-level evaluation provides a more fine-grained view of system performance, showing that most individual tokens are correctly analyzed (86% POS and 83% morphological accuracy), while sentence-level errors typically arise from isolated tagging mistakes within otherwise correct analyses.

Remaining errors are mainly associated with ambiguous or underspecified glosses, alternative forms separated by slashes and inconsistent use of hyphens. These issues suggest that further refinement of gloss normalization.

More broadly, this study provides a reproducible methodology for enriching dialectal corpora with structured linguistic annotation, leveraging rule-based linguistic knowledge and similarity-driven alignment.

Future work will pursue three directions. First, the high-quality alignments obtained here can be used to expand the dialect–Italian lexicon and improve coverage across varieties. Second, the enriched corpus provides a basis for developing models that generate literal Italian glosses from dialectal input, facilitating semi-automatic interlinear glossing. Third, the projected morpho-syntactic annotations can serve as silver-standard supervision for training dialect-specific tagging models.

9. Bibliographical References

Željko Agić, Dirk Hovy, and Anders Søgaard. 2015. If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 268–272.

Steven Bird. 2011. Bootstrapping the language archive: New prospects for natural language processing in preserving linguistic heritage. *Linguistic Issues in Language Technology*, 6.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic Inequalities in Language Technology Performance across the World's Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Lindell Bromham, Russell Dinnage, Hedvig Skirgård, Andrew Ritchie, Marcel Cardillo, Felicity Meakins, Simon Greenhill, and Xia Hua. 2022. Global predictors of language endangerment and the future of linguistic diversity. *Nature ecology & evolution*, 6(2):163–173.

Ernestina Carrilho. 2010. Tools for dialect syntax: the case of CORDIAL-SIN (an annotated corpus of Portuguese dialects). *Tools for linguistic variation*, pages 57–70.

Terry Crowley. 2007. *Field linguistics: A beginner's guide*. OUP Oxford.

Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal Stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–4592.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8.

Micha Elsner and David Liu. 2025. Prompt and circumstance: A word-by-word LLM prompting approach to interlinear glossing for low-resource languages. In *Proceedings of the 22nd SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics*, pages 1–14, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Inés Fernández-Ordóñez. 2011. Nuevos horizontes en el estudio de la variación gramatical del español: el Corpus Oral y Sonoro del Español Rural. *Noves tendències en la dialectología contempoània*, pages 173–203.

Pilar García Mouton. 2017. El Atlas Lingüístico de la Península Ibérica (ALPI) en línea. geolingüística a la carta.

Matthias Huck, Diana Dutka, and Alexander Fraser. 2019. Cross-lingual Annotation Projection Is Effective for Neural Part-of-Speech Tagging. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 223–233, Ann Arbor, Michigan. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Michael Krauss. 1992. The world's languages in crisis. *Language*, 68(1):4–10.

Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Michele Loporcaro, Stephan Schmid, Chiara Zanini, Diego Pescarini, Giulia Donzelli, Stefano Negrinelli, Graziano Tisato, et al. 2021. AIS, reloaded: A digital dialect atlas of Italy and southern Switzerland. *TRAVAUX DE LINGUISTIQUE ROMANE*, pages 111–136.

Maria Rita Manzini and Leonardo Maria Savoia. 2005. *I dialetti italiani e romanci: Morfosintassi generativa*. Edizioni dell'Orso.

G Mazzaggio, LA Ludovico, MV Vena, Maria Rita Manzini, Leonardo Maria Savoia, et al. 2023. Morphosyntax of Italian and Romance Varieties: Presentation of the Manzini and Savoia (2005) Corpus and Its Digitalization. *Bollettino dell'Atlante Linguistico Italiano*, 2023(47):185–210.

Greta Mazzaggio and Neri Binazzi. 2024. Valorizzare il patrimonio immateriale: un'esperienza di digitalizzazione del dialetto. *DILEF Rivista digitale del Dipartimento di Lettere e Filosofia*, 3:224–242.

Greta Mazzaggio, Carlo Zoli, Neri Binazzi, Luca Andrea Ludovico, Mael Vittorio Vena, M. Rita Manzini, and Leonardo Maria Savoia. 2025. Morphosyntactic Variation in Italian and Romansh Dialects: The Manzini & Savoia (2005) Corpus within Project CHANGES. In *DIGITAL HERITAGE (2025). The Eurographics Association*.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Sarah Moeller, Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 2018. A neural morphological analyzer for Arapaho verbs learned from a finite state transducer. In *Proceedings of the workshop on computational modeling of polysynthetic languages*, pages 12–20.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Garrett Nicolai, Eleanor Chodroff, Frederic Mailhot, and Çağrı Çöltekin, editors. 2025. *Proceedings of the 22nd SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics*. Association for Computational Linguistics, Albuquerque, New Mexico, USA.

Sebastian Padó and Mirella Lapata. 2005. Cross-linguistic Projection of Role-Semantic Information. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 859–866, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Diego Pescarini and Giorgio Maria Di Nunzio. 2010. Il database dell'Atlante Sintattico d'italia (ASIt). *Quaderni di lavoro ASIT*, 10:63–81.

Brigitte Rührlinger. 2004. Atlante linguistico del Ladino Dolomitico e dei dialetti limitrofi (ald). *Bollettino dell'Atlante Linguistico Italiano*, 28:229–243.

Mengqiu Wang and Christopher D Manning. 2014. Cross-lingual projected expectation regularization for weakly supervised learning. *Transactions of the Association for Computational Linguistics*, 2:55–66.

Changbing Yang, Garrett Nicolai, and Miikka Silfverberg. 2024. Multiple Sources are Better Than One: Incorporating External Knowledge in Low-Resource Glossing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4537–4552, Miami, Florida, USA. Association for Computational Linguistics.

David Yarowsky and Grace Ngai. 2001. Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection Across Aligned Corpora. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Eros Zanchetta and Marco Baroni. 2005. Morph-it! a free corpus-based morphological resource for the Italian language. In *Proceedings of corpus linguistics conference series 2005 (ISSN 1747-9398)*, volume 1, pages 1–12. University of Birmingham.

Carlo Zoli, Greta Mazzaggio, and Neri Binazzi. 2025. Small Codes: a platform for digital resources and tools for minority languages and dialects. In *DIGITAL HERITAGE (2025). The Eurographics Association*.

10. Language Resource References

Cristina Bosco, Simonetta Montemagni, Maria Simi, et al. 2013. Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 61–69. The Association for Computational Linguistics.

Explosion. 2024. spaCy: Industrial-strength Natural Language Processing in Python. `https://spacy.io`. Version 3.8.0, model: it_core_news_lg.

Leonardo Maria Savoia, Maria Rita Manzini, Greta Mazzaggio, Ludovico Luca Andrea, Mael Vittorio Vena, Carlo Zoli, Benedetta Baldi, Ludovico Franco, and Neri Binazzi. 2025. The Manzini & Savoia (2005) Corpus: Morphosyntactic Variation in Italian and Romansh Dialects.

Gloss	POS	Morph Features
ClS	PRON	Case=Nom, Clitic=Yes, PronType=Prs
ClS-3sm	PRON	Case=Nom, Clitic=Yes, Gender=Masc, Number=Sing, Person=3, PronType=Prs
ClS-3pm	PRON	Case=Nom, Clitic=Yes, Gender=Masc, Number=Plur, Person=3, PronType=Prs
ClS-2s	PRON	Case=Nom, Clitic=Yes, Number=Sing, Person=2, PronType=Prs
ClS-2p	PRON	Case=Nom, Clitic=Yes, Number=Plur, Person=2, PronType=Prs
ClS-3s	PRON	Case=Nom, Clitic=Yes, Number=Sing, Person=3, PronType=Prs
ClS-3p	PRON	Case=Nom, Clitic=Yes, Number=Plur, Person=3, PronType=Prs
ClS-sm	PRON	Case=Nom, Clitic=Yes, Gender=Masc, Number=Sing, PronType=Prs
ClS-sf	PRON	Case=Nom, Clitic=Yes, Gender=Fem, Number=Sing, PronType=Prs
ClS-pm	PRON	Case=Nom, Clitic=Yes, Gender=Masc, Number=Plur, PronType=Prs
ClS-pf	PRON	Case=Nom, Clitic=Yes, Gender=Fem, Number=Plur, PronType=Prs
Neg	ADV	PronType=Neg
-1ps	VERB	Person=1, Number=Sing, VerbForm=Inf
-2ps	VERB	Person=2, Number=Sing, VerbForm=Inf
-3ps	VERB	Person=3, Number=Sing, VerbForm=Inf
-1pp	VERB	Person=1, Number=Plur, VerbForm=Inf
-2pp	VERB	Person=2, Number=Plur, VerbForm=Inf
-3pp	VERB	Person=3, Number=Plur, VerbForm=Inf

Table 6: Representative functional glosses with manually assigned POS and morphological features.

A. Annotation projection procedure

Morpho-syntactic annotation is assigned to dialect tokens by projecting the analysis of the corresponding gloss tokens. The same procedure is used in both the first and second pass; the only difference between the two runs is the alignment supplied as input. Given a pair of preprocessed columns, `gloss_preprocessed` and `text_preprocessed`, the algorithm produces, for each sentence, a list of token-level dictionaries containing the dialect token, the aligned gloss token, its projected POS label, dependency relation, and morphological features.

Functional glosses Before running the POS tagger, the system identifies a small inventory of pre-coded gloss strings that encode grammatical information directly, such as `ClS-3sm`, `Neg`, or `-1ps`. These items do not behave like ordinary lexical tokens and therefore receive manually specified annotations. Each functional gloss is associated with a fixed bundle of values for `pos`, `dep`, and `morph`. Table 6 illustrates the main classes.

General procedure For each row of the dataframe, the algorithm proceeds as follows.

1. **Input validation.** If either the gloss string or the dialect text is missing or not represented as a string, the row is skipped and an empty annotation is returned.

2. **Extraction of functional glosses.** The gloss is scanned for occurrences of pre-coded strings listed in a manually defined dictionary. For each match, the system records:

   - the matched string,
   - its character offsets in the original gloss,
   - its manually assigned `pos`, `dep`, and `morph` values.

   These matched substrings are then removed from the gloss, yielding a cleaned gloss string.

3. **Automatic analysis of the remaining gloss tokens.** The cleaned gloss is passed to the Italian `spaCy` pipeline. For each resulting token, the tagger returns:

   - token text,
   - part of speech,
   - morphological features.

   These analyses provide the default annotation for lexical gloss material.

4. **Reinsertion of functional glosses.** The manually annotated functional glosses are reinserted into the sequence of automatically analyzed gloss tokens. To preserve the original order, both sets of items are assigned character-based positions and merged according to their start offsets. The result is a single ordered sequence of annotated gloss tokens.

5. **Projection onto dialect tokens.** The dialect text is tokenized by whitespace. The system then traverses the ordered gloss-token sequence and, position by position, replaces the token form in the gloss sequence with the corresponding dialect token. At the same time, the original gloss token is preserved in a

separate `gloss` field. In this way, each output token contains:

- `text`: the dialect token,
- `gloss`: the aligned gloss token,
- `pos`: the projected part of speech,
- `dep`: the projected dependency label,
- `morph`: the projected morphological features.

6. **Output construction.** The projected tokens are stored as a list of dictionaries, one list per sentence. This list constitutes the annotation output for the row.

Assumptions and limitations The procedure assumes that the dialect text and the processed gloss are token-aligned after reinsertion of functional glosses. Projection is therefore strictly position-based once the aligned gloss sequence has been reconstructed. This makes the method simple and deterministic, but it also means that tagging errors may arise whenever residual alignment mismatches persist.

Functional gloss inventory Table 6 gives representative examples of the manually specified functional glosses used during projection.

B. IPA-to-grapheme normalization

Here we describe in greater detail the deterministic rule-based normalization procedure adopted to map each IPA token to a simplified graphemic string (see Section 6.2). The goal of this procedure is not to reconstruct dialect-specific orthographies, but to reduce symbol-set mismatch and make dialect and gloss strings more comparable for subsequent character-based alignment (as described in Section 6.4).

Input and output The procedure takes as input a sentence in IPA (stored in `text_preprocessed_new`) and produces a normalized sentence (stored in `text_to_grapheme`). Normalization is applied token-by-token by whitespace splitting. Each token is transformed by a fixed cascade of rewrite rules described below.

Overview of the rule cascade Let t be an IPA token. The function NORMALIZE(t) applies the following steps in order:

1. **Removal of suprasegmental marks**

   Primary and secondary stress symbols (e.g. ˈ, ˌ) are removed.

2. **Simplification of gemination and length**

   Length marks (ː) are interpreted as consonant gemination. When the preceding segment corresponds to an ASCII consonant, the consonant is duplicated (e.g. pː → pp). Residual length marks are then removed.

   For selected IPA segments that may appear geminated, repeated or length-marked sequences are collapsed to a single symbol:

   $$ʃː \to ʃ \quad ʃʃ \to ʃ \quad ɲː \to ɲ \quad ʎː \to ʎ$$

3. **Direct mappings for palatal consonants**

   Some IPA consonants are mapped to standard Italian graphemic representations:

   $$ɲ \to \texttt{gn} \quad ʎ \to \texttt{gli} \quad ŋ \to \texttt{n}$$

4. **Vowel normalization**

   Open and central vowels are mapped to plain Latin vowels in order to reduce token variability:

   $$ɛ \to \texttt{e} \quad ɔ \to \texttt{o} \quad ə \to \texttt{e} \quad ɪ \to \texttt{i} \quad ʊ \to \texttt{u}$$

   This step removes phonetic distinctions that are not relevant for the alignment task.

5. **Normalization of additional IPA symbols**

   Various IPA consonants and diacritics not typically used in Italian are mapped to approximate Latin equivalents or removed. Examples include:

   $$β \to \texttt{v} \quad θ \to \texttt{t} \quad ð \to \texttt{d} \quad ɾ \to \texttt{r} \quad ʔ \to$$

   Certain fricatives such as ɕ and ʂ are temporarily treated as members of the ʃ class to allow uniform contextual mapping in the following step.

6. **Context-sensitive mappings**

   Several consonants are rewritten depending on the following vowel, approximating common Italian orthographic conventions.

   (a) Velars:

   $$\begin{aligned} \texttt{k} &\to \texttt{ch} \quad / \_ \{\texttt{e}, \texttt{i}\} \\ \texttt{k} &\to \texttt{c} \quad \text{elsewhere} \end{aligned}$$

   Similarly,

   $$\begin{aligned} \texttt{g} &\to \texttt{gh} \quad / \_ \{\texttt{e}, \texttt{i}\} \\ \texttt{g} &\to \texttt{g} \quad \text{elsewhere} \end{aligned}$$

   (b) Affricates:

   $$\begin{aligned} \widehat{\texttt{tʃ}} &\to \texttt{c} \quad / \_ \{\texttt{e}, \texttt{i}\} \\ \widehat{\texttt{tʃ}} &\to \texttt{ci} \quad / \_ \{\texttt{a}, \texttt{o}, \texttt{u}\} \end{aligned}$$

Similarly,

$$\widehat{\text{dʒ}} \rightarrow \text{g} \quad / \_ \{\text{e}, \text{i}\}$$
$$\widehat{\text{dʒ}} \rightarrow \text{gi} \quad / \_ \{\text{a}, \text{o}, \text{u}\}$$

(c) Postalveolar fricatives:

$$\text{ʃ} \rightarrow \text{sc} \quad / \_ \{\text{e}, \text{i}\}$$
$$\text{ʃ} \rightarrow \text{sci} \quad / \_ \{\text{a}, \text{o}, \text{u}\}$$

7. **Affricate simplification**

   Clusters corresponding to ts and dz are optionally collapsed to z, reflecting a simplified Italian-style approximation.

8. **Intervocalic z weakening**

   A single z between vowels is rewritten as s (while leaving zz unchanged), e.g.:

   $$\text{aza} \rightarrow \text{asa}$$

9. **Final cleanup**

   Residual IPA tie-bars (e.g., ‿) and other leftover diacritics are removed. Apostrophes are normalized by removing surrounding whitespace and repeated spaces are collapsed.

C. Automatic paradigm completion

Paradigm completion is implemented as a rule-based procedure operating on the tabular structure of the corpus. The algorithm scans the dataset sequentially and detects candidate paradigm blocks based on editorial conventions of the corpus.

1. **Detection of paradigm starters.** A row is identified as a paradigm starter if three conditions hold: (i) the gloss ends with “etc.” or “ecc.”, (ii) the gloss contains at least one token tagged as a finite first-person singular verb (POS $\in$ {VERB, AUX}, Person=1, Number=Sing, VerbForm $\neq$ Part) according to the morpho-syntactic annotation produced in the first-pass projection (see Section 4), and (iii) the following five rows contain empty gloss fields.

   When these conditions are satisfied, the six rows are treated as an inflectional paradigm block.

2. **Template construction.** The starter gloss is stripped of the trailing “etc./ecc.” marker and tokenized using an apostrophe-aware tokenizer that preserves contractions (e.g. *l’ho*). The resulting token sequence is used as a template representing the first-person singular form of the paradigm.

3. **Identification of verb positions.** Within the template, the algorithm identifies all token positions corresponding to finite 1SG verbs using the morpho-syntactic information obtained during the first-pass projection. Multiple verbs within the same gloss are supported.

4. **Paradigm retrieval.** For each detected verb token, the system retrieves its lemma and inflectional paradigm from Morph-it (Zanchetta and Baroni, 2005), a lexicon of Italian inflected forms with their lemma and morphological features. Candidate analyses are filtered to retain only finite 1SG verb forms. Once a correct analysis is identified, the lemma and morphological class are used to retrieve all lexicon entries belonging to the same paradigm, which provides the set of inflected forms used to generate the remaining person-number variants.

5. **Generation of the remaining paradigm rows.** The remaining five rows of the block are generated by systematically modifying the template to produce the forms corresponding to:

   2SG, 2SG, 1PL, 2PL, 3PL.

   For each person-number combination, the following operations are applied:

   (a) Replacement of the verb token(s) with the corresponding inflected form obtained from the retrieved paradigm.

   (b) Adjustment of subject pronouns and clitic pronouns (e.g. *io* $\rightarrow$ *tu*, *mi* $\rightarrow$ *ti*) according to the target person-number specification.

   (c) Rewriting of enclitic forms such as *-mi* to the corresponding person-specific clitic.

   (d) Agreement adjustments for participles when the auxiliary is *essere*.[1] In plural persons, participles are pluralized using a simple Italian pluralization heuristic.

   (e) Minor normalization heuristics to correct known lexical irregularities (e.g., correcting truncated 3PL forms such as *han* (‘have.3PL’) in favor of the more common *hanno*).

6. **Reconstruction of the gloss line.** After token-level rewriting, the tokens are recombined into a string while preserving apostrophe attachment rules (e.g. *l’* + *ho* $\rightarrow$ *l’ho*).

[1] In composite verbal forms of Italian, formed by an auxiliary verb and the past participle of the main verb, the participle is inflected for gender and number.

The resulting five lines are written into the previously empty rows of the paradigm block, while the original starter line is preserved as the 1SG template.