Robust Language Identification for Romansh Varieties

Charlotte Model **Sina Ahmadi** **Jannis Vamvas**
University of Zurich
charlotte.model@uzh.ch, sina.ahmadi@uzh.ch, vamvas@cl.uzh.ch

Abstract

The Romansh language has several regional varieties, called *idioms*, which sometimes have limited mutual intelligibility. This linguistic diversity motivates the need for a language identification (LID) system that can distinguish between these idioms, yet to date there has been no well-documented effort to build one. Since Romansh LID should also be able to recognize Rumantsch Grischun, a supra-regional variety that combines elements of several idioms, this makes for a novel and interesting classification problem. In this paper, we present a LID system for Romansh idioms based on an SVM approach. We evaluate our model on a newly curated benchmark across two domains and find that it reaches an average in-domain accuracy of 97%, enabling applications such as idiom-aware spell checking or machine translation. Our classifier is publicly available.[1]

1 Introduction

Language identification (LID) is the task of automatically determining the language of text or speech. As a foundational component in natural language processing (NLP) pipelines, LID enables downstream applications such as machine translation, information retrieval, content moderation, and multilingual text processing by routing to the most appropriate language-specific system. While state-of-the-art LID systems achieve high accuracy for widely-used languages (Chen et al., 2024), distinguishing between closely-related language varieties remains a significant challenge (Burchell et al., 2024). Unlike unrelated languages that exhibit clear phonological, morphological, and lexical differences, closely-related varieties have fewer unique linguistic features, making them difficult to differentiate automatically. The challenges of LID are compounded for varieties spoken in limited geographic regions, where the scarcity of diverse digital data hinders the development and generalization of robust LID systems.

In this paper, we present a LID system for Romansh, a collection of closely-related linguistic varieties native to Switzerland and spoken by approximately 40,000 people (Grünert, 2024). Due to Romansh's low-resource status, no well-documented effort in the NLP community has developed a LID system capable of distinguishing between its varieties; existing multilingual LID systems either ignore it entirely or treat it as a monolithic language. Our contributions are threefold:

(1) We curate a novel benchmark dataset spanning multiple domains (news, broadcast transcripts, textbooks, and newsroom notes) to enable systematic evaluation of idiom classification;

(2) We develop an SVM-based classifier that achieves 97% average in-domain accuracy across all six varieties; and

(3) We provide qualitative analysis on the LID task for Romansh idioms.

Our work demonstrates that effective LID for closely-related low-resource varieties is achievable with carefully designed features and appropriate training data, providing a model for similar efforts in other endangered and regional language contexts.

2 Romansh and its Varieties

The term *Romansh* refers to a collection of closely related linguistic varieties of Rhaetian descent native to the canton of Grisons in Switzerland. These varieties, known as *idioms*, comprise five historically distinct forms: *Sursilvan*, *Sutsilvan*, *Surmiran*, *Puter*, and *Vallader*. To facilitate official communication with the government, a sixth, artificially-created variety was introduced in 1982 to act as

[1] https://github.com/ZurichNLP/romansh-lid

the established written standard, known as *Rumantsch Grischun* (RG) (Grünert, 2024). Communities speaking different idioms are distributed within an area of approximately 7,000 km^2 across Grisons.

The geographical proximity of certain idioms is reflected in their linguistic similarity. According to Lia Rumantscha, the organization tasked by the Swiss government to represent and preserve the Romansh identity, the two idioms spoken in the Rhine valley, Sursilvan and Sutsilvan, are lexically closely related, as are Puter and Vallader, spoken in the upper and lower Engadine valley respectively (Lia Rumantscha, 2015). Surmiran, spoken in the central region, often acts as a bridge between these two idiom groups. Finally, since RG was developed on the basis of Sursilvan, Vallader, and Surmiran, it shares features with multiple idioms (Anderson, 2016).

3 Related Work

3.1 Language Identification

LID has been widely studied in NLP, with growing importance for filtering the multilingual corpora used to train language models (Foroutan et al., 2026). When applied to closely-related varieties and dialects, especially within low-resource languages, the task poses significant challenges due to the limited availability of linguistic resources and annotated datasets. Research highlights the difficulty of accurately distinguishing among variants, as demonstrated in studies focused on dialectal Arabic (Dahou et al., 2025) and Indo-Aryan languages such as Bhojpuri and Assamese (Mundotiya et al., 2021). Recent techniques explore machine learning and deep learning architectures, leveraging cross-lingual transfer learning and community-driven approaches that emphasize regional linguistic families (Salesky et al., 2021).

3.2 Romansh NLP

Ongoing projects at Lia Rumantscha for Romansh NLP focus on offering digital services to translate from Romansh into German, the most widespread official language of Switzerland, and vice versa. These services include idiom-specific dictionaries hosted across several portals: Pledari Grond covers Surmiran, Sutsilvan, RG and most recently Sursilvan directly in its interface[2], the Uniun dals Grischs[3] serves Puter and Vallader, and the Dicziunari Rumantsch is an app that aggregates this data on mobile[4]. Pledari Grond also provides a spell-checker UI for RG[5]. One property shared by all these projects is that the user must select the desired idiom before translation or spell-checking can take place, limiting their usability to cases where the user already knows what idiom a text is in.

More recently, resources and models have been released that specifically focus on Romansh, including the Mediomatix Corpus (Hopton et al., 2026), containing parallel sentences extracted from schoolbooks, and WMT24++ (Vamvas et al., 2025), containing Romansh translations of the WMT24++ benchmark in machine translation. Additionally, the Swiss-made LLM Apertus has specifically incorporated Romansh data for post-training (Hernández-Cano et al., 2025, p. 94).

4 Data

For the training and evaluation of Romansh LID systems, we compile five sources of textual data spanning dictionary entries, journalistic articles, broadcast transcripts, newsroom notes, and school textbooks:

- **Pledari Grond** (PG), a comprehensive Romansh-German dictionary covering all Romansh idioms.
- **La Quotidiana** (LQ), a Romansh newspaper with daily idiom-annotated content. We use WordPress dumps from 2021 to 2025.
- **Radiotelevisiun Svizra Rumantscha** (RTR), validated speech transcripts from Romansh broadcasts, annotated by idiom.
- **RTR Telesguard Notes** (TG), pre-broadcast notes written by journalists in their native idioms (excluding RG).
- **Mediomatix Textbooks** (TB), parallel scholastic material per idiom (excluding RG), recently released by Hopton et al. (2026).

4.1 Preprocessing

For each source we extract the idiom label and the main text fields. We then apply minimal cleaning to all samples by i) removing intra-class duplicates (exact string match within idiom); ii) stripping HTML/markup and collapsing repeated whitespace

[2] https://www.pledarigrond.ch

[3] https://www.udg.ch

[4] https://www.dicziunari.ch

[5] https://www.pledarigrond.ch/rumantschgrischun/spellchecker

and newlines; iii) dropping empty/None/non-letter-only items. Additionally, we remove source artifacts that do not carry language cues such as dictionary markers, e.g., sense numerals, "cf." stubs, editorial signatures, worksheet placeholders and long underscore sequences. RTR content showed no recurring artifacts. Noisy or non-Rumantsch snippets were discarded and clear mislabels, if noticed visually, were manually corrected by dictionary checks.

Appendix A summarises the export statistics for each data source after this cleaning, and provides links to the publicly available datasets.

Following Bernier-colborne et al. (2023), we run exact and (where feasible) near-duplicate detection across all the data sources after merging and before splitting them. Exact duplicates judged valid in multiple idioms are kept but routed to the training split.

4.2 Named-Entity Masking

To reduce reliance on lexical memorisation, we produce two training variants: `masked` and `unmasked`. For `masked`, we run a fine-tuned named entity recognition model based on SwissBERT (Vamvas et al., 2023) (zero-shot on RG) with conservative heuristics (min length, standalone tokens, score ≥ 0.98) and replace matched spans with `$NE$`. Pledari Grond is left unmasked due to very few named entities and high inference cost. Because masking replaces spans in place rather than dropping samples, both variants contain the same 487,172 samples and differ only in token count (12.43M unmasked vs. 12.38M masked).

5 Experimental Setup

5.1 Data Splits

Following these steps, we finally split all the datasets into train, validation and test sets. For the latter, we use multiple test sets to disentangle domain, balance, and comparability effects.

- **Train** (`train-set`): PG + LQ + RTR + TB (unbalanced). Both masked and unmasked variants used in training experiments.
- **Dev** (`dev-set`): Balanced, in-domain RTR, **6,000** samples (1k/idiom; avg. 45.9 tokens).
- **Test-A** (`test-a`): In-domain, *unbalanced* LQ, **6,000** samples (avg. 528.5 tokens).
- **Test-B** (`test-b`): In-domain, *balanced* RTR, **6,000** samples (avg. 45.7 tokens).
- **Test-C** (`test-c`): In-domain, approximately *balanced* TB (no RG), **6,000** samples (avg. 85.8 tokens).
- **Test-D** (`test-d`): *Out-of-domain* TG (no RG), **9,607** samples (avg. 151.9 tokens).

To reflect realistic LID inputs, we did not lowercase or strip punctuation on dev/test; we only removed non-letter-only items and ensured non-empty (post-strip) text/labels.

5.2 Classification

We frame our task as a supervised multi-class classification problem where the task is to assign each text sample to one of six varieties.

We extract bag-of-n-grams features combining word unigrams and overlapping character n-grams ($n \in \{1, 2, 3, 4\}$), represented as TF–IDF vectors with sublinear term-frequency scaling and ℓ_2 normalisation. We compare four families of linear classifiers, all from `scikit-learn` (Pedregosa et al., 2011), each with explicit ℓ_2 regularisation by default: (i) Logistic Regression (`LogisticRegression`, multinomial, ℓ_2 penalty, inverse regularisation strength C, `saga` solver, 5,000 max iterations); (ii) linear SVM (`LinearSVC`, squared-hinge loss, ℓ_2 penalty, `dual=False`, inverse regularisation strength C); (iii) two stochastic-gradient variants (`SGDClassifier`)—one with hinge loss matching the SVM objective (SGD-SVM), the other with log loss matching the LR objective (SGD-LR), both with ℓ_2 regularisation strength $\alpha = 10^{-4}$, 5,000 max iterations, and early stopping; (iv) Multinomial and Complement Naïve Bayes baselines (additive smoothing $\alpha = 1.0$).

Baseline values were the `scikit-learn` defaults ($C = 1.0$ for SVM and LR), and were subsequently tuned by randomised search (§ 5.3). The two SGD variants share the SVM and LR loss functions respectively, but differ in optimiser: `LogisticRegression` uses the full-batch `saga` solver, whereas `SGDClassifier` uses mini-batch SGD with a fixed learning-rate schedule. This isolates optimiser effects from loss-function effects.

5.3 Hyperparameter Optimization

We perform 40-iteration randomized searches over pipeline hyperparameters using stratified

Classifier	Accuracy	Macro F1	Weighted F1	Macro Recall
Logistic Regression (LR)	79.5	76.9	79.1	74.6
Linear SVM	78.1	75.4	77.9	73.9
SGD (Linear SVM)	76.3	72.4	75.2	68.0
Naive Bayes (counts)	75.1	72.1	74.6	70.6
Naive Bayes (TF)	75.1	72.1	74.6	70.6
Naive Bayes (TF-IDF)	73.7	72.0	73.9	71.7
SGD (Logistic Regression)	73.5	69.0	72.0	63.6
Majority Baseline	35.1	8.7	18.3	16.7

Table 1: Results from the preliminary experiments with different baseline classifiers, where classifiers are ordered by macro F1 score.

5-fold cross-validation on a stratified 20% subset of the training data, optimising for macro F_1 (`RandomizedSearchCV`, `random_state = 42` throughout). The search space included the regularisation parameters of each classifier (C drawn log-uniformly from $[10^{-2}, 4]$ for SVM and $[10^{-2}, 2]$ for LR), the penalty norm (ℓ_2, ℓ_1, elastic net for LR; ℓ_2 and ℓ_1 for SVM), the elastic-net mixing ratio $l_1\_ratio \in [0.05, 0.9]$, the SVM loss (hinge / squared-hinge), the character n-gram range ($(1,3)$ or $(1,4)$), the word n-gram range ($(1,1)$ or $(1,2)$), and the minimum document frequency (1 or 2). Final models were trained on both unmasked and masked variants of the training set.

The best LR configuration (penalty = ℓ_1, $C \approx 1.94$, $l_1\_ratio \approx 0.47$, character $(1,4)$-grams, word $(1,2)$-grams) reached 96.8 macro F_1 on the dev set, while the best SVM configuration (penalty = ℓ_1, $C \approx 0.62$, squared-hinge loss, character $(1,4)$-grams with $\text{min\_df} = 2$, word unigrams only) reached 97.1 macro F_1 on the dev set.

6 Results

6.1 Overall Results

Table 1 reports the preliminary baseline comparison. All classifiers substantially outperformed the majority-class baseline, with macro F_1 scores clustering within 8 points of each other. LR achieved the highest score (76.9), followed closely by linear SVM (75.4), suggesting that for this task feature representation matters more than classifier choice. We retained LR and SVM for hyperparameter optimisation.

After tuning (§ 5.3), the ordering reversed: SVM (97.1 dev macro F_1) slightly outperformed LR (96.8). We therefore selected SVM for the remaining experiments.

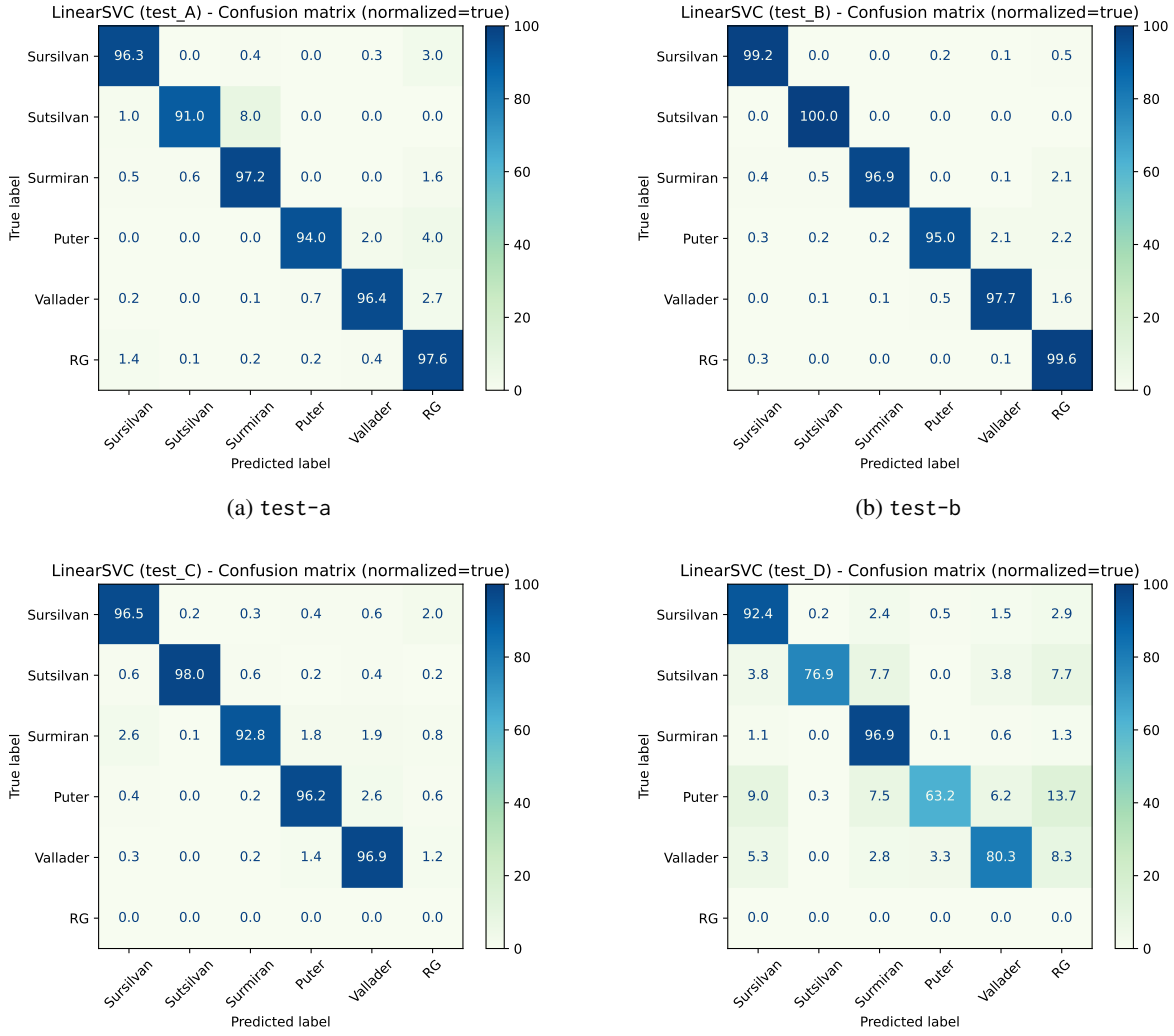
Test Set	Domain	Balance	Acc.	F1
`test-a`	in-domain	unbal.	96.8	94.7
`test-b`	in-domain	bal.	98.1	98.1
`test-c`	in-domain	bal.	96.2	80.5
`test-d`	out-domain	unbal.	90.7	69.1

Table 2: Accuracy and macro F1 scores per test set. The highest scores are achieved on the balanced (bal.) vs. unbalanced (unbal.), in-domain set `test-b`.

Table 2 presents the performance of our optimized linear SVM classifier across the four different test sets. The model achieves macro F1 scores ranging from 98.1 on the balanced in-domain test set (`test-b`) to 69.1 on the unbalanced out-of-domain set (`test-d`), demonstrating this classical machine learning approach is effective for Romansh LID under favorable conditions while struggling with out-of-domain and noisy data. These results compare favorably to similar language discrimination tasks. Recent VarDial shared tasks on Italian dialects and French varieties report best macro F1 scores of 74.6 and 34.4 respectively, while the Dravidian Language Identification task achieved a score of 93, closer to our in-domain results (Chakravarthi et al., 2021; Aepli et al., 2022).

6.2 Per Idiom Performances

Figure 1 shows confusion matrices for each test set. The model achieves consistently high recall across idioms on `test-b` (95.0–100.0), with perfect classification of Sutsilvan samples. Performance varies more on other sets: `test-a` shows recalls of 91.0–97.6, `test-c` shows 92.8–98.0, while `test-d` exhibits the widest range (63.2–96.9). The



(a) `test-a` (b) `test-b`

(c) `test-c` (d) `test-d`

Figure 1: Row-normalized confusion matrices for all test sets. The model achieves near-perfect classification on the balanced in-domain set (`test-b`), while confusion increases on out-of-domain data (`test-d`).

Test Set	Accuracy		Macro F1	
	Masked	Unmasked	Masked	Unmasked
`test-a`	96.8	96.9	94.7	94.5
`test-b`	98.1	98.8	98.1	98.8
`test-c`	96.2	96.2	80.5	80.4
`test-d`	90.7	90.5	69.1	69.0

Table 3: Comparison of masked vs. unmasked training. No major effect can be observed.

notably low recall for Puter in `test-d` (63.2) warrants investigation, as samples were frequently misclassified as RG. This may reflect the informal, potentially noisy nature of the Telesguard notes data. Our analysis of misclassifications reveals several data quality issues. In `test-a`, many errors involved samples containing primarily named entities or German text—artifacts from the journalism data source. In `test-c`, misclassifications often occurred on noisy samples containing mainly numbers, punctuation, or very short texts from the textbook data.

Idiom	1st	Type	2nd	Type	3rd	Type
Sursilvan	ei	char	scha	char	iu	char
Sutsilvan	àn	char	ù	char	eing	char
Surmiran	eir	char	dall	char	ous	char
Puter	ron	char	aunt	char	ọ	char
Vallader	ạ	char	ọn	char	ị	char
RG	␣	char	vegnis	word	ì	char

Table 4: Top 3 most discriminative features per idiom. "char" denotes character n-gram and "word" denotes word unigram. Character n-grams emerge as the most discriminative features across all varieties, with empty space ␣ being the most informative one for RG.

6.3 Discriminative Features

The most informative features are predominantly character n-grams rather than word unigrams, aligning with findings from similar LID tasks. We present these features in Table 4. Several patterns emerge: notably, Puter and Vallader features include underdotted characters (ọ, ụ, ạ, ị, ẹ) that are unique to these idioms in the Pledari Grond data. Although these features are highly discriminative for our classifier, they are not part of the standard orthography of Puter and Vallader: in Pledari Grond they function as phonetic stress markers, and would not appear in running text written in either idiom outside a dictionary setting. The whitespace character appearing as the top feature for RG is an artefact of severe class imbalance in the training data (≈171K RG samples versus 44K–86K for other idioms; see Appendix B), since whitespace sequences are normalised to a single space during preprocessing and RG samples then dominate single-space n-grams.

We tested two ablations to address these artefacts: (i) restricting character n-grams to within-word boundaries only (rather than allowing them to span across whitespace), and (ii) removing character 1-grams from the feature set entirely. Both decreased overall macro F_1, most plausibly because they remove the underdotted characters and other diacritics that supply most of the discriminative signal for Puter and Vallader.

Future work should explore balancing the training data, which we expect to be a more robust remedy for the whitespace artefact than feature-side filtering.

6.4 Impact of Named Entities

Comparing models trained on masked vs. unmasked data, as presented in Table 3, shows negligible performance differences ($\leq$ 0.1 F1 points on most sets; on `test-b` the unmasked variant is 0.7 F1 higher, the largest gap). This confirms that named entities provide no idiom-specific information—unsurprising given the geographic proximity of Romansh communities. News outlets like La Quotidiana and RTR report region-wide events in multiple idioms, textbooks cover similar subjects across idioms, and RG serves as a pan-idiom standard.

7 Conclusion

We presented the first documented language identification system for Romansh idioms, capable of distinguishing between the five historical varieties (Sursilvan, Sutsilvan, Surmiran, Puter, and Vallader) and the standardised RG. Using an SVM classifier with character and word n-gram features, our model achieves up to 98% macro F1 on balanced in-domain data. Our experiments confirm that character n-grams are the most discriminative features for this task, with idiom-specific diacritics and orthographic patterns providing strong classification signals. Furthermore, we found that named entity masking has negligible impact on performance, suggesting that the classifier relies on genuine linguistic features rather than memorizing location-specific proper nouns.

Limitations

Our work has several limitations. First, the training data exhibits substantial class imbalance, with RG overrepresented due to the Pledari Grond dictionary data, which may explain artifacts such as whitespace emerging as a top discriminative feature. Second, performance degrades on out-of-domain data (`test-d`), where macro F1 drops to 69.1, indicating limited generalization to informal text genres like newsroom notes. Third, some discriminative features, particularly the underdotted characters unique to Pledari Grond, may not generalize to texts following standard orthographic conventions. Finally, our evaluation is restricted to written text; spoken language identification for Romansh remains unexplored. Despite these limitations, our system provides a practical tool for downstream applications such as idiom-aware spell checking and machine translation routing, and establishes

a baseline for future work on Romansh and other low-resource regional language varieties.

Acknowledgments

This work is based on a Bachelor's thesis that was presented to the University of Zurich.[6] We thank Lia Rumantscha and RTR for their support and for facilitating access to the data sources used in this study, including Pledari Grond, La Quotidiana, and RTR materials. We also thank Uniun dals Grischs for making dictionary data for Puter and Vallader available to us for research use, and Ignacio Pérez Prat for helpful feedback. Their commitment to preserving and promoting the Romansh language made this research possible.

References

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Stephen R. Anderson. 2016. Romansh (rumantsch). In Adam Ledgeway and Martin Maiden, editors, *The Oxford Guide to the Romance Languages*, pages 169–184. Oxford University Press, Oxford.

Gabriel Bernier-colborne, Cyril Goutte, and Serge Leger. 2023. Dialect and variant identification as a multi-label classification task: A proposal based on near-duplicate analysis. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.

Laurie Burchell, Alexandra Birch, Robert P. Thompson, and Kenneth Heafield. 2024. Code-switched language identification is harder than you think. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 646–658. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Mihaela Gaman, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial evaluation campaign 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–11, Kyiv, Ukraine. Association for Computational Linguistics.

Wei-Rui Chen, Ife Adebara, Khai Duy Doan, Qisheng Liao, and Muhammad Abdul-Mageed. 2024. Fumbling in babel: An investigation into chatgpt's language identification ability. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4387–4413. Association for Computational Linguistics.

Abdelghani Dahou, Abdelhalim Hafedh Dahou, Mohamed Amine Chéragui, Amin Abdedaiem, Mohammed A. A. Al-qaness, Mohamed Abd Elaziz, Ahmed A. Ewees, and Zhonglong Zheng. 2025. A survey on dialect arabic processing and analysis: Recent advances and future trends. *Acm Transactions on Asian and Low-Resource Language Information Processing*.

Negar Foroutan, Jakhongir Saydaliev, Grace Kim, and Antoine Bosselut. 2026. ConLID: Supervised contrastive learning for low-resource language identification. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6693–6708, Rabat, Morocco. Association for Computational Linguistics.

Matthias Grünert. 2024. Rätoromanisch. In Elvira Glaser, Johannes Kabatek, and Barbara Sonnenhauser, editors, *Sprachenräume der Schweiz. Band 1: Sprachen*, pages 156–184. Narr Francke Attempto, Tübingen.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, and 82 others. 2025. Apertus: Democratizing open and compliant llms for global language environments. *Preprint*, arXiv:2509.14233.

Zachary Hopton, Jannis Vamvas, Andrin Büchler, Anna Rutkiewicz, Rico Cathomas, and Rico Sennrich. 2026. The mediomatix corpus: Parallel data for Romansh language varieties via comparable schoolbooks. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 290–306, Rabat, Morocco. Association for Computational Linguistics.

Lia Rumantscha. 2015. Facts. https://www.liarumantscha.ch/sites/default/files/2023-07/PDF%20cumplet_d.pdf. Accessed: 2025-06-17.

Rajesh Kumar Mundotiya, Manish Kumar Singh, Rahul Kapur, Swasti Mishra, and Anil Kumar Singh. 2021. Linguistic resources for bhojpuri, magahi, and maithili: Statistics about them, their similarity estimates, and baselines for three applications. *Acm*

[6] https://seafile.ifi.uzh.ch/f/96df2a17539546e7a192/

Transactions on Asian and Low-Resource Language Information Processing.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Elizabeth Salesky, Badr M. Abdullah, Sabrina Mielke, Elena Klyachko, Oleg Serikov, Edoardo Maria Ponti, Ritesh Kumar, Ryan Cotterell, and Ekaterina Vylomova. 2021. SIGTYP 2021 shared task: Robust spoken language identification. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 122–129, Online. Association for Computational Linguistics.

Jannis Vamvas, Johannes Graën, and Rico Sennrich. 2023. SwissBERT: The multilingual language model for Switzerland. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 54–69, Neuchatel, Switzerland. Association for Computational Linguistics.

Jannis Vamvas, Ignacio Pérez Prat, Not Soliva, Sandra Baltermia-Guetg, Andrina Beeli, Simona Beeli, Madlaina Capeder, Laura Decurtins, Gian Peder Gregori, Flavia Hobi, Gabriela Holderegger, Arina Lazzarini, Viviana Lazzarini, Walter Rosselli, Bettina Vital, Anna Rutkiewicz, and Rico Sennrich. 2025. Expanding the WMT24++ benchmark with Rumantsch Grischun, Sursilvan, Sutsilvan, Surmiran, Puter, and Vallader. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1028–1047, Suzhou, China. Association for Computational Linguistics.

A Data Sources

Source and URL to Dataset (if available)	Idiom	Total # of Samples	Total # of Tokens	Avg. # of Tokens/Sample
Pledari Grond https://www.pledarigrond.ch	Sursilvan	34,724	92,182	2.65
	Sutsilvan	32,175	307,799	9.57
	Surmiran	43,954	301,390	6.86
	Puter	59,852	963,940	16.11
	Vallader	72,470	1,141,688	15.75
	RG	172,616	785,831	4.55
	total	415,791	3,592,830	8.64
La Quotidiana (LQ) https://huggingface.co/datasets/ZurichNLP/quotidiana	Sursilvan	6,088	3,030,828	497.84
	Sutsilvan	363	192,933	531.50
	Surmiran	1,750	1,008,736	576.42
	Puter	878	430,122	489.89
	Vallader	2,395	1,250,042	521.94
	RG	2,567	1,365,988	532.13
	total	14,041	7,278,649	518.39
Radiotelevisiun Svizra Rumantscha (RTR) https://developer.srgssr.ch/en/apis/rtr-linguistic	Sursilvan	6,979	353,205	50.61
	Sutsilvan	3,074	156,674	50.97
	Surmiran	7,196	245,931	34.18
	Puter	6,016	225,389	37.46
	Vallader	5,787	275,516	47.61
	RG	4,359	232,020	53.23
	total	33,411	1,488,735	44.56
Telesguard Notes (TG) n.a.	Sursilvan	4,931	641,844	130.17
	Sutsilvan	27	7,122	263.78
	Surmiran	3,033	261,001	86.05
	Puter	578	162,754	281.58
	Vallader	1,053	389,428	369.83
	RG	0	0	0.00
	total	9,622	1,462,149	151.96
Mediomatix Textbooks (TB) https://huggingface.co/datasets/ZurichNLP/mediomatix	Sursilvan	12,233	971,643	79.43
	Sutsilvan	12,116	1,005,356	82.98
	Surmiran	6,698	534,563	79.81
	Puter	12,238	1,023,779	83.66
	Vallader	12,283	1,019,339	82.99
	RG	0	0	0.00
	total	55,568	4,554,680	81.97

Table 5: Summary of the number of samples extracted from each data source for each idiom, along with the number of whitespace tokens across all collected samples per data source and the average number of tokens per sample.

B Data Split Statistics

Set	Idiom	Total # of Samples	Total # of Tokens	Avg. # of Tokens/Sample
	Sursilvan	55,574	3,663,641	65.92
	Sutsilvan	44,188	1,342,065	30.37
	Surmiran	55,543	1,457,845	26.25
train-set	Puter	74,355	2,353,323	31.65
	Vallader	86,313	2,372,016	27.48
	RG	171,199	1,190,808	6.96
	total	487,172	12,379,698	25.41
	Sursilvan	1,000	50,500	50.5
	Sutsilvan	1,000	50,523	50.52
	Surmiran	1,000	35,133	35.13
dev-set	Puter	1,000	36,891	36.89
	Vallader	1,000	48,482	48.48
	RG	1,000	53,577	53.58
	total	6,000	275,106	45.85
	Sursilvan	1,000	502,709	502.71
	Sutsilvan	100	51,535	515.35
	Surmiran	800	454,843	568.55
test-a	Puter	100	48,195	481.95
	Vallader	2,000	1,046,133	523.07
	RG	2,000	1,067,778	533.89
	total	6,000	3,171,193	528.53
	Sursilvan	1,000	50,324	50.32
	Sutsilvan	1,000	51,542	51.54
	Surmiran	1,000	33,906	33.91
test-b	Puter	1,000	37,401	37.4
	Vallader	1,000	48,440	48.44
	RG	1,000	52,523	52.52
	total	6,000	274,136	45.69
	Sursilvan	1,250	99,690	79.75
	Sutsilvan	1,250	113,106	90.48
	Surmiran	1,000	79,629	79.63
test-c	Puter	1,250	112,670	90.14
	Vallader	1,250	109,764	87.81
	RG	0	0	0
	total	6,000	514,859	85.81
	Sursilvan	4,923	640,746	130.15
	Sutsilvan	26	7,120	273.85
	Surmiran	3,031	260,720	86.02
test-d	Puter	576	162,488	282.1
	Vallader	1,051	388,699	369.84
	RG	0	0	0
	total	9,607	1,459,773	151.95

Table 6: Detailed statistics for all data splits by idiom, including number of samples, tokens, and average tokens per sample. The training set exhibits class imbalance, with RG containing the most samples but the shortest average length due to dictionary entries from Pledari Grond. `test-c` and `test-d` lack RG samples by construction.