RUMLEM: A Dictionary-Based Lemmatizer for Romansh

Dominic P. Fischer **Zachary Hopton** **Jannis Vamvas**
Department of Computational Linguistics, University of Zurich
{dominicphilipp.fischer, zacharywilliam.hopton, jannisnikos.vamvas}@uzh.ch

Abstract

Lemmatization – the task of mapping an inflected word form to its dictionary form – is a crucial component of many NLP applications. In this paper, we present RUMLEM, a lemmatizer that covers the five main varieties of Romansh as well as the supra-regional standard variety Rumantsch Grischun. It is based on comprehensive, community-driven morphological databases for Romansh, enabling RUMLEM to cover 77–84% of the words in a typical Romansh text. Since there is a dedicated database for each Romansh variety, an additional application of RUMLEM is variety-aware language classification. Evaluation on 30'000 Romansh texts of varying lengths shows that RUMLEM correctly identifies the variety in 95% of cases. In addition, a proof of concept demonstrates the feasibility of Romansh vs. non-Romansh language classification based on the lemmatizer.

https://github.com/ZurichNLP/rumlem

1 Introduction

Romansh is a minority Romance language spoken by approximately 40'000–60'000 speakers in several Alpine valleys of Switzerland. It comprises five regional varieties, or *idioms* (Sursilvan, Sutsilvan, Surmiran, Puter, and Vallader), as well as Rumantsch Grischun (RG), the supra-regional standard variety of Romansh. The varieties differ to such an extent that mutual intelligibility is often limited, highlighting the need for variety-specific NLP-tools.

The proposed lemmatizer RUMLEM, given Romansh text, uses morphological databases to **(1)** infer the possible lemmas of word forms, **(2)** identify morphological features of word forms, and **(3)** identify the likely Romansh variety of the input (cf. Figure 1). Together with Model et al. (2026), our system is among the very few systems to reliably perform such a classification; with the additional benefit that it can be used to distinguish between Romansh and non-Romansh text. It thus provides a transparent complement to machine learning approaches for language identification and a core component to variety-aware Romansh NLP-tools.

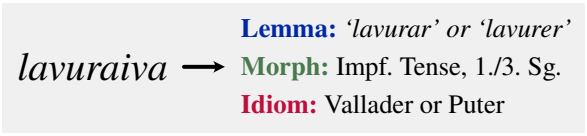

Figure 1: The main functionalities of RUMLEM.

Our approach builds on existing, maintained, and community-driven dictionary data (cf. Table 1), which we process into 725'005 unique word forms mappable to 178'467 lemmas (cf. Table 2).

These data allow our lemmatizer to cover around 80% of a typical Romansh text (cf. Section 4.1) and to identify the variety correctly in 95% of cases (cf. Table 4). Language identification experiments (cf. Figure 2) show that a threshold of ca. 0.6 (i.e., 60% of words recognised as a particular Romansh variety) serves to distinguish Romansh texts from the most closely related Romance languages.

2 Dictionary Resources for Romansh

2.1 Bilingual Dictionaries

Pledari Grond,[1] the dictionary underlying RUMLEM, covers all six Romansh varieties, with translations provided in German (DE), as well as, in part, additional annotations (cf. Table 1). Users may report potentially erroneous German–Romansh pairs and suggest alternative translations.

The dictionaries for Rumantsch Grischun, Surmiran, Sursilvan and Sutsilvan are openly licensed (© Lia Rumantscha 1980–2025). The Vallader and Puter dictionaries were kindly provided by Uniun dals Grischs for use solely as part of this lemmatizer (© Uniun dals Grischs. All rights reserved).

[1] https://pledarigrond.ch/

Proceedings of the 11th Edition of the Swiss Text Analytics Conference (SwissText 2026) – Scientific Track, pages 126–133
June 10, 2026

	Unique Entries	Single Words	DE Translations	POS Tags	Gender	Infl. Verbs
Sursilvan	147,977	93,211	147,971	67,201	72,460	5,031
Sutsilvan	58,584	39,191	58,581	42,817	30,876	3,021
Surmiran	74,986	44,365	74,986	39,224	36,956	2,947
Puter	89,908	32,084	89,807	13,712	36,918	3,383
Vallader	106,690	35,322	106,438	10,854	48,435	3,779
RG	249,169	94,291	249,165	98,046	161,942	3,867
Total	727,314	338,464	726,948	271,854	387,587	22,028

Table 1: Description of the *Pledari Grond* dictionary for each Romansh idiom as well as Rumantsch Grischun. A single 'ǀ' means 'thereof': Unique entries, *thereof* X Single Words, German Translations, etc. *Infl. Verbs* refers to the number of unique verbs for which inflected forms are provided.

2.2 Spellchecking

Pledari Grond also provides a Romansh spell-checking system based on HUNSPELL.[2] With the focus lying primarily on orthographic conventions rather than on inflectional or derivational morphology across the different varieties, we do not use the spell-checker for inflectional processing. We restrict its use to providing a fallback vocabulary, together with the *Mediomatix* corpus (Hopton et al., 2026) and the Rumantsch Grischun newspaper *La Quotidiana*.[3] Our lemmatizer uses said vocabulary to check non-lemmatizable words against a variety's lexicon.

3 Software Design

3.1 Preprocessing of Dictionary Data

Lemma mappings constitute the central building block of RUMLEM, making the transformation of the available dictionary data into this format a key step. We treat the morphologically rich and frequently annotated parts of speech (POS) nouns, verbs, and adjectives separately. Other POS (where present) or entries lacking POS tags were treated jointly. Where possible, we used conservative, rule-based heuristics to assign missing POS tags – for example, treating entries whose German translation begins with a capital letter as nouns.

The entries in the Pledari Grond dictionaries exhibit a wide range of structural patterns, with e.g. approximately 200 distinct patterns each for nouns and adjectives. Pattern recognition distinguishes between single (w) and multiple words (w+), punctuation symbols, as well as special morphological tags (marked below as MT; e.g., m., f., sg., pl.). Two of the most frequent noun patterns and a less frequent one serve as illustration (# occurrences):

w (208,000): armaziun; f; Bewaffnung

w, w (5067): admiratur, admiratura; m/f; Bewunderer(in)

w (w, MT); w (w, MT) (93): arrestà (arrestats, pl); arrestada (arrestadas, pl); m/f; Gefangene

Based on such recurring patterns, informed decisions could be made about how to process the data; for example, that a **w, w** entry such as the one above should yield two separate entries, *admiratur; m; Bewunderer* and *admiratura; f; Bewundererin*. To ensure clean and consistent processing, we automatically generated test skeletons for each distinct pattern occurring more than ten times in each of the four POS categories (N, V, ADJ and other), and manually annotated the corresponding gold-standard outputs. 200 such tests, covering 99.9% of the input data, contribute to high data quality (cf. Appendix A). Table 2 presents the resulting data available to the lemmatizer.

3.2 Variety Identification Process

RUMLEM takes a text and an optional variety. If none is given, it predicts the most likely variety based on the input text. More specifically, the text is tokenized using an adapted version[4] of the Italian Moses tokenizer (Koehn et al., 2007); then, for each variety, the system counts lemmatizable tokens and tokens found in the variety-specific vocabulary and divides this count by the total number of tokens.

[2] https://hunspell.github.io/

[3] https://huggingface.co/datasets/ZurichNLP/quotidiana

[4] The adaptations consist of regex-based preprocessing and protected token patterns designed to correctly tokenize apostrophe-based contractions in each Romansh variety.

	Vocab	Mapped Forms	Lemmas	Noun	Adj	Verb	Other
Sursilvan	223,826	222,860	36,505	23,206	4,977	5,858	2,464
Sutsilvan	129,519	87,902	19,326	12,467	2,671	3,033	1,155
Surmiran	149,078	84,481	22,838	15,145	3,107	3,204	1,382
Puter	180,361	107,758	26,201	15,534	3,122	3,102	4,443
Vallader	165,354	109,090	30,479	19,841	4,821	3,625	2,192
RG	180,690	112,914	43,118	31,200	6,099	4,049	1,770
Total	1,028,828	725,005	178,467	117,393	24,797	22,871	13,406

Table 2: RUMLEM's data coverage, single words only. *Mapped Forms* describes the amount of entries linked to a lemma. A single '|' means 'thereof': Vocab, *thereof* X Mapped Forms, *thereof* X Lemmas, etc.

3.3 Lemmatization Process

Consider the sentence *La vuolp d'eira darcheu üna jada fomantada* ("The fox was once again hungry"). RUMLEM identifies the variety correctly as Vallader, and returns per-token analyses, as shown for *fomantada* in Table 3: assuming Vallader, it may correspond to the feminine form of the adjective *fomantà*, a feminine noun *fomantada*, or the past participle of the verb *fomantar*.

3.3.1 Unknown Word Forms

To try and map unknown word forms to known lemmas, the lematizer's edit tree component may be invoked. Adapting a system for unsupervised morphological paradigm completion (Jin et al., 2020; Kann et al., 2020), this component learns frequent inflectional patterns and their morphological tags for each variety and POS category. For example, masculine adjectives receive an -a when turned feminine, and an -s when pluralized, meaning the edit trees store two adjective-specific lemmatization paths: drop -a (feminine) or drop -s (plural). At inference, upon encountering an unknown word, paths across all POS are searched and transformations applied if suffixes match. The resulting potential lemmas are checked against existing lemmas within the same POS-tag. Matches are collected, and the candidate with the shortest edit distance to the original word form is selected.

4 Evaluation

4.1 Lemmatization Coverage

The data used for this task as well as for variety identification consisted of 3000–7000 texts for each variety, covering a range of input lengths. Shorter texts consist of validated speech transcripts from Romansh broadcasts by Radiotelevisiun Svizra Rumantscha (RTR). Texts longer than 300 tokens are taken from a set of children's stories called *Babulins*, which exist in each Romansh variety. Note that the distributions are not even and differ between varieties; we report them in Appendix C.

We define lemmatization coverage as the percentage of word forms in a Romansh text for which our lemmatizer returns an analysis (i.e., excluding forms in the fallback vocabulary). Removing the high-frequency punctuation symbols "., !?; :", we find that RUMLEM lemmatizes around 80% of all word forms, with variety-specific coverages ranging between 77% and 84%. The edit trees component manages to cover another 5% of words, raising the coverage to around 85%; as it is designed conservatively (cf. Section 3.3.1) and learns

Variety	Form + features	Gloss
RM-SURMIRAN	`fomanto [PoS=ADJ; Gender=FEM; Number=SG]`	hungrig
RM-SURMIRAN	`fomantar [PoS=V; VerbForm=PTCP; Tense=PST; Gender=FEM; Number=SG]`	aushungern
RM-VALLADER	`fomantà [PoS=ADJ; Gender=FEM; Number=SG]`	ausgehungert
RM-VALLADER	`fomantà [PoS=ADJ; Gender=FEM; Number=SG]`	hungrig
RM-VALLADER	`fomantada [PoS=N; Gender=FEM; Number=SG]`	Ausgehungerte
RM-VALLADER	`fomantada [PoS=N; Gender=FEM; Number=SG]`	Hungrige
RM-VALLADER	`fomantar [PoS=V; VerbForm=PTCP; Tense=PST; Gender=FEM; Number=SG]`	jn aushungern

Table 3: RUMLEM's output given the token 'fomantada', with potential lemmas and morphological annotations of the form itself returned, as well as the German translation (variants).

from high-quality dictionary data, the risk of erroneous lemma mappings remains limited. We report detailed coverage scores in Appendix B.

4.2 Dictionary-based Variety Identification

We also evaluated the performance of our lemmatizer in terms of variety classification accuracy. The results, summarized in Table 4, show that RUMLEM accurately recognizes the variety of the vast majority of Romansh texts, especially longer texts. We note that the text genre might, in addition to text length, play a role in classification accuracy.

Length	2–10	10–50	50–300	300–800	800+	All
Sursilvan	0.85	0.85	0.87	1.00	1.00	0.86
Sutsilvan	1.0	0.99	1.0	1.00	1.00	1.00
Surmiran	0.92	0.94	0.99	1.00	1.00	0.95
Puter	0.97	0.98	0.99	1.00	1.00	0.98
Vallader	0.94	0.91	0.93	1.00	1.00	0.92
RG	0.89	1.00	1.00	1.00	1.00	1.00
All	0.94	0.94	0.95	1.00	1.00	

Table 4: Classification accuracy by Rumantsch variety across length buckets (number of tokens).

These scores are comparable to what Model et al. (2026) reports on balanced in-domain data, as well as unbalanced in-domain data with longer samples (avg. ca. 530 tokens). On shorter unbalanced in-domain data (avg. ca. 85 tokens) and out-of-domain data, their SVM classifier struggles, with F1 scores dropping to ca. 0.8 and 0.7, respectively.

4.3 Dictionary-based Language Identification

We selected about 5000 texts from Fineweb[5] in Romansh itself as well as the four Romance languages French, Italian, Catalan and Romanian, as these languages are typologically close to Romansh and therefore most likely to exhibit overlapping dictionary forms (cf. Table D). We record the "winning" scores, i.e., the highest score assigned to a text across the Romansh varieties. Figure 2 shows the Romansh score distributions in turquoise and the non-Romansh ones in rust color for three different setups: as-is, using the sets of words, and removing Romance-language (FR, IT, CA, RO) stopwords.

Figure 2 and App. D.2 show that, using the sets of words, a separating threshold can be found for all three tested buckets. Perfect separation was achieved apart from bucket 50–300; however, each of the misclassified samples was highly noisy, containing a mix of languages (cf. App. D.3). Further manual inspection revealed the presence of many similar samples on the lower end of the Romansh distribution; the ideal threshold may thus lie higher.

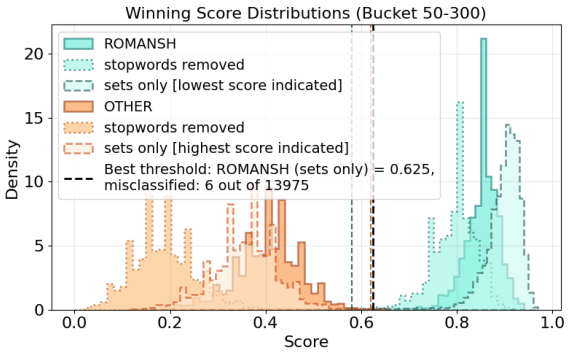

Figure 2: Distributions of Romansh (turquoise) and other Romance languages (rust) according to the highest Romansh variety score assigned to each text sample.

These results suggest that a straightforward Romansh language identification system could be built using RUMLEM. Since the distributions are well-separated, a small validation set would suffice to determine the optimal classification threshold.

5 Related Work

NLP for Romansh Romansh and its varieties are not yet covered by popular NLP tools and resources such as SpaCy, Universal Dependencies (de Marneffe et al., 2021), and UniMorph (Batsuren et al., 2022), motivating the development of dedicated tools. Recent years have nonetheless seen progress in other areas of Romansh NLP, including contextualized token embeddings and named entity recognition (Vamvas et al., 2023), word alignment (Dolev, 2023), and machine-learning-based variety identification (Model et al., 2026). This paper extends this line of work by providing a dictionary-based system for (context-agnostic) lemmatization and morphosyntactic analysis that, as we show, can also serve as a basis for variety and language identification.

Dictionary-based Lemmatization Our work is situated in the tradition of rule- and lexicon-based computational morphology (Koskenniemi, 1984; Schmid et al., 2004). While more recent, neural approaches to lemmatization and morphosyntactic analysis can take into account the context of word forms and even generalize to unseen forms (Straka et al., 2016; McCarthy et al., 2019; Qi et al., 2020),

[5]The data is made up of webpages crawled by CommonCrawl between 2013 and 2024.

they require supervised data, typically in the form of treebanks. In the absence of such treebanks, dictionary-based lemmatization is a viable alternative when a comprehensive dictionary is available for a language. Dictionary-based lemmatizers have been proposed for, among others, German (based on Wiktionary; Liebeck and Conrad, 2015), Middle English (Karimov et al., 2016), Latin (Passarotti et al., 2017), and Somali (Mohamed and Mohamed, 2023). Our system builds on six large-scale, highly consistent dictionaries for the Romansh varieties that include inflection tables and German translations, enabling relatively high word coverage and a rich feature set.

6 Conclusion

We presented RUMLEM, a dictionary-based lemmatizer covering all six Romansh varieties. Beyond lemmatizing around 80% of a given Romansh text, RUMLEM reliably identifies Romansh varieties – averaging 95% accuracy across varieties and text lengths – and can be used to distinguish Romansh even from its most closely related Romance languages. RUMLEM's transparent design makes it a useful complement to machine learning approaches for Romansh NLP.

Limitations

RUMLEM's performance is inherently bounded by its dictionary coverage and quality: words absent from Pledari Grond cannot be lemmatized, and annotation errors will be propagated as-is. Since the lemmatizer cannot account for context, ambiguous forms may receive multiple analyses without disambiguation. Future work could explore the use of statistical or neural approaches to make it more context-aware.

Finally, we note that while the software of RUMLEM itself is released open-source including the postprocessed dictionary data, the Vallader and Puter dictionaries are released without an open-source license, and use of these dictionaries for research beyond RUMLEM will require written permission from the copyright holders.

Acknowledgments

We thank the Swiss Federal Office of Culture, the Lia Rumantscha, and the Uniun dals Grischs for their support, the participants of the digidi 2025 workshop for fruitful discussions, and Sina Ahmadi for helpful feedback.

References

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, and 76 others. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Eyal Liron Dolev. 2023. Does mBERT understand Romansh? evaluating word embeddings using word alignment. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 41–53, Neuchatel, Switzerland. Association for Computational Linguistics.

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.

Huiming Jin, Liwei Cai, Yihui Peng, Chen Xia, Arya McCarthy, and Katharina Kann. 2020. Unsupervised morphological paradigm completion. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6696–6707, Online. Association for Computational Linguistics.

Katharina Kann, Arya D. McCarthy, Garrett Nicolai, and Mans Hulden. 2020. The SIGMORPHON 2020 shared task on unsupervised morphological paradigm completion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 51–62, Online. Association for Computational Linguistics.

Raoul Karimov, Maria Samkova, Svetlana Nikitina, and Andrei Akinin. 2016. Using a hybrid algorithm for lemmatization of a diachronic corpus. In *Proceedings of the Workshop on Computational Linguistics and Language Science*, volume 1886 of *CEUR Workshop Proceedings*, pages 1–8, Moscow, Russia. CEUR-WS.org.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion*

Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Kimmo Koskenniemi. 1984. A general computational model for word-form recognition and production. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, pages 178–181, Stanford, California, USA. Association for Computational Linguistics.

Matthias Liebeck and Stefan Conrad. 2015. IWNLP: Inverse Wiktionary for natural language processing. 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 414–418, Beijing, China. Association for Computational Linguistics.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Charlotte Model, Sina Ahmadi, and Jannis Vamvas. 2026. Robust language identification for Romansh varieties. In *Proceedings of the 11th edition of the Swiss Text Analytics Conference*, Zurich, Switzerland. Association for Computational Linguistics.

Shafie Abdi Mohamed and Muhidin A. Mohamed. 2023. Lexicon and rule-based word lemmatization approach for the Somali language. In *4th Workshop on African Natural Language Processing*.

Marco Passarotti, Marco Budassi, Eleonora Litta, and Paolo Ruffolo. 2017. The lemlat 3.0 package for morphological analysis of Latin. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 24–31, Gothenburg. Linköping University Electronic Press.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

A Preprocessing Tests

The example below illustrates a preprocessing test case. The first line shows the raw dictionary data, while the lines following '>>>' show the validated format – i.e., how the data is processed before being fed into the lemmatizer."

```
'antalg(iant)evel, antalg(iant)evla'; adj
>>> antalgevel:
        antalgevel; ADJ;MASC;SG
        antalgevla; ADJ;FEM;SG
    antalgiantevel:
        antalgiantevel: ADJ;MASC;SG
        antalgiantevla: ADJ;FEM;SG
```

B Lemmatization Coverage

Table 5 shows coverage values across texts of varying lengths and varieties. The text data used are the same as in the variety identification experiments – shorter texts come from Fineweb, longer ones from Babulins (cf. Section 4.1). Overall, coverage does not vary significantly with text length or genre, except for fragmentary texts, where individual missing word forms have a greater impact.

Of the approximately 40,000 missing (i.e., non-lemmatizable) tokens across all varieties, around 11,000 are proper nouns or German nouns. Another 4,000 consist of tokens containing numbers, special tokens, or characters such as dashes and hyphens. Notable cases include contractions that are absent from the Pledari Grond dictionaries. Out of the remaining 25'000 – mostly Romansh – words, around 5000 were not lemmatized due to the respective texts being misclassified, the remaining 20'000 are indeed absent from the lemmatizer. The edit trees component cuts this number down to about 13'500 genuinely missing word forms.

Length	2–10	10–50	50–300	300–800	800+	All
Surs.	0.76	0.81	0.84	0.83	0.78	0.84
Suts.	0.71	0.77	0.77	0.73	0.73	0.77
Surm.	0.73	0.80	0.84	0.82	0.79	0.82
Puter	0.79	0.83	0.84	0.84	0.81	0.84
Vall.	0.66	0.77	0.80	0.83	0.81	0.80
RG	0.93	0.79	0.79	0.80	0.78	0.79
All	0.75	0.80	0.82	0.81	0.79	

Without edit trees

Length	2–10	10–50	50–300	300–800	800+	All
Surs.	0.79	0.84	0.86	0.83	0.79	0.86
Suts.	0.81	0.84	0.84	0.82	0.80	0.84
Surm.	0.80	0.86	0.90	0.86	0.84	0.88
Puter	0.83	0.88	0.89	0.86	0.83	0.89
Vall.	0.72	0.82	0.85	0.86	0.84	0.84
RG	0.96	0.86	0.86	0.82	0.80	0.86
All	0.80	0.86	0.87	0.84	0.82	

With edit trees

Table 5: Coverage ratios by text length and variety. Values for each variety and bucket are averaged across samples. 'All' shows the total number of lemmatizable tokens divided by the total tokens in the bucket/variety.

C Variety ID

C.1 Samples per Variety and Bucket

Table 6 shows the number of samples per Romansh variety and bucket. The same samples were used for both coverage evaluation as well as variety identification.

Length	2–10	10–50	50–300	300–800	800+	Tot
Surs.	68	2647	4173	7	5	6900
Suts.	6	1190	1795	5	7	3003
Surm.	1113	3751	2204	6	6	7080
Puter	660	2783	2468	5	7	5923
Vall.	277	2209	3202	5	7	5700
RG	9	1218	3052	6	6	4291
Tot	2133	13798	16894	34	38	29897

Table 6: #samples by Romansh variety across buckets.

D Language ID

D.1 Samples per Variety and Bucket

Table 7 shows the number of samples per Romance variety and bucket, used for Romansh vs. non-Romansh identification.

D.2 Separating Thresholds per Bucket

Note that the best threshold is defined as, primarily, the one that best separates the data, and, secondarily, the one with the widest margin of separation.

Length	50–300	300–800	800–2000	Tot
French	2517	1551	693	4761
Italian	2493	1620	671	4784
Romanian	2128	1513	922	4563
Catalan	2595	1575	593	4763
Romansh	4242	661	75	4978
Tot	13975	6920	2954	23849

Table 7: #samples by Romance language across buckets.

In both buckets 300–800 and 800–2000, all methods resulted in perfect separation. However, using the sets of words provided the widest margin, indicating that using sets results in the best separation across the different buckets.

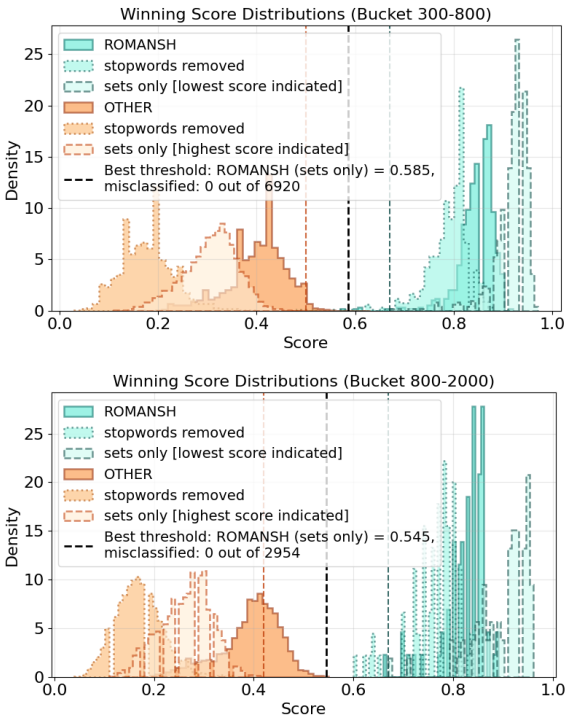

Figure 3: Winning variety score distributions for Romansh (turquoise) and other Romance languages (rust), for token length buckets 300–800 (top) and 800–2000 (bottom).

This was confirmed when we repeated the experiment with different Romansh data, namely the data from the variety classification task. Here too, reducing the texts to their sets of words before being processed by the lemmatizer produced the best threshold in all three token budget settings.

For all experiment settings, we also calculated the average score distribution across varieties instead of the winning score, which resulted in slightly more misclassifications and less inter-

pretable thresholds. This is due to the non-trivial differences between the Romansh varieties, which should be treated separately instead of being conflated.

Finally, note that the ideal threshold decreasing with text length is due to longer texts widening the gap between the Romansh and non-Romansh text, whereby the threshold is placed in the middle.

D.3 Analysis of Misclassified Samples

The six misclassified data samples from Figure 2 were all highly noisy. Four out of the six contained parallel translations (Romansh in bold, Italian in italic):

- "[..] **II cussagl da scoula as cumpuona da duos commembers e dal suprastant dal decasteri.** [...] Der Schulrat setzt sich aus zwei Mitgliedern zusammen sowie dem für diesen Bereich zuständigen Gemeinderatsmitglied. [...]"
- "Herzlich Willkommen Wir begrüssen Sie herzlich auf der Seite unserer Kirchgemeinde und danken Ihnen für Ihr Interesse an unserem Pfarreileben. [...] **Cordial beinvegni silla pagina dalla pleiv catolica Sevgein/Castrisch/Riein. Nus engraziein a Vus che Vus s'interesseis per nossa pleiv.** [...]"
- "**Gorbatschow und Freund Sbalzs classics sün las cordas L'interpret da Balalaika straordinari da nos temp es il virtuos Prof. Andreij Gorbatschow chi viva a Moskau.** [...] Klassische Saitensprünge Der herausragende Balalaika-Interpret unserer Zeit ist der in Moskau lebende Star-Virtuose Prof. Andreij Gorbatschow. [...]"
- "*Un cordiale benvenuto* – Herzlich willkommen - **Cordial bainvegni** *Associazione Spitex dei Grigioni Siamo l'associazione mantello delle 19 organizzazioni Spitex che operano nel Canton dei Grigioni.* [...] Spitex Verband Graubünden Wir sind der Dachverband der 19 im Kanton Graubünden tätigen Spitex-Organisationen. [...] **Federaziun grischuna da spitex Nus essan l'uniun tetgala da las 19 organisaziuns da spitex activas en il chantun Grischun.** [...]"

The remaining two contained what looked like web-scraping artifacts in German:

- "Foto aus dem Akt-Channel Teilnahme am Forum Fotos verkaufen Mehr Foto-Ordner anlegen? Mehr Fotos speichern? [...] **Igl october vargau havein nus era visitau quei marcau ed jeu muossel in maletg ord il casti da Schönbrunn. Amicabels salids giu da Glion Glieci**"
- "**Cla Rauch ha orientà davart l'Archiv Cultural d'Engiadina Bassa (fotografia: Benedict Stecher).** [...] Haben Sie noch kein Konto? Registrieren Sie sich hier [...]"