IruMozhi: Automatically classifying diglossia in Tamil

Kabilan Prasanna
Academies of Loudoun
kabilanprasanna@gmail.com

Aryaman Arora
Stanford University
aryamana@stanford.edu

Abstract

Tamil, a Dravidian language of South Asia, is a highly diglossic language with two very different registers in everyday use: Literary Tamil (preferred in writing and formal communication) and Spoken Tamil (confined to speech and informal media). Spoken Tamil is under-studied in modern NLP systems compared to Literary Tamil written in the Tamil script, as evidenced by a lack of datasets explicitly targetting the Spoken variety. In this paper, we release IruMozhi, a human-translated dataset of parallel text in Literary and Spoken Tamil. Using IruMozhi, we train classifiers on the task of identifying which Tamil variety a text belongs to. We use these models to gauge the availability of pretraining data in Spoken Tamil, to audit the composition of existing labelled datasets for Tamil, and to encourage future work on the variety.

https://github.com/kebathan/diglossia

1 Introduction

Diglossia is a linguistic phenomenon wherein a community maintains two (or more) varieties of their language, with the appropriate variety to use depending on the social context (Ferguson, 1959, 1996). Prototypically, diglossia manifests as two varieties: a **high** variety employed in formal contexts and a **low** variety employed in informal settings. The high variety tends to be standardised and highly preferred in writing and other formal communication (speeches, news broadcasts, etc.), while the low dialect is confined to speech and informal written communication (social media, text messages, etc.) and subject to regional and stylistic variation. Diglossia is thus a challenge for modern NLP systems—accessible training data on the internet usually overrepresents the high variety, while the average user may prefer using the low variety to interact with NLP systems.

English	The tail is also white.
Literary	வாலும் வெள்ளையாக உள்ளது. vaalum vellaiyaaga ulladhu
Spoken (1)	vaalu vellaiye irukku
Spoken (2)	vaalum white-ah irruku
English	Duryodhana's close friend.
Literary	துரியோதனனின் உற்ற நண்பன். thuriyodhananin utra nanban
Spoken (1)	dhuriyodhananoda nalla nanban
Spoken (2)	dhuriyodhanan-oda uyir nanban

Table 1: Two examples from our parallel corpus of Literary and Spoken Tamil showing morphological (1), phonological (2), and lexical differences (3). Spoken (1) and (2) are produced by different annotators.

Tamil is one such highly diglossic language primarily spoken in the state of Tamil Nadu in India, and in Sri Lanka and Singapore (Annamalai and Steever, 2015). Tamil belongs to the Dravidian language family, and is the oldest attested language in this group. Literary Tamil is the standardised (high) variety, preserving a more archaic stage[1] of the language than the low variety termed Spoken Tamil. Spoken Tamil (or Colloquial Tamil) is subject to dialectal variation by geography and caste, but in India there does exist a widely used and understood (but not officially regulated) Standard Spoken Tamil, based primarily on the dialect of educated non-Brahmin urban residents of central Tamil Nadu (Annamalai, 1980; Schiffman, 1998, 1999; Saravanan et al., 2009). Both forms of the language coexist in complementary social contexts, and thus practical NLP systems should endeavour to support both.

Tamil is a rising star in data availability for NLP research (Joshi et al., 2020; Arora et al.,

[1] Literary Tamil traditionally follows the rules described in the *Naṉṉūl*, a 13th-century grammar by Pavaṇanti. However, it has been subject to linguistic change since then by e.g. the coining of new words.

Findings of the Association for Computational Linguistics: NAACL 2024, pages 3096–3103

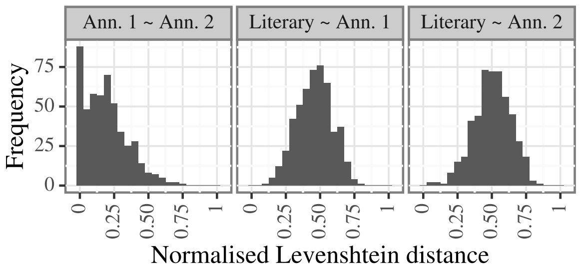

Figure 1: Histogram of normalised Levenshtein distances between parallel sentences from our Literary Tamil corpus and the two Spoken Tamil translators. The two Spoken Tamil sets are much more similar to each other than to Literary Tamil.

Set 1	Set 2	Lev. (↓)	(norm.)	BLEU (↑)	chrF (↑)
Ann. 1	Ann. 2	7.99	0.19	35.34	73.49
Literary	Ann. 1	21.84	0.46	0.83	37.19
Literary	Ann. 2	23.78	0.50	0.73	33.28

Table 2: Text similarity metrics between the transliterated Literary Tamil text and the two Spoken Tamil translators.

2022). However, most recent research, particularly on general-purpose systems like language models, has focused on Literary Tamil to the detriment of the Spoken variety. Combined with a lack of standardisation, we expect existing systems to be much worse at all tasks in Spoken Tamil. To combat this problem, we introduce a corpus of high-quality Literary Tamil sentences paired with human-elicited equivalents in Spoken Tamil. Using this data, we train classifiers to identify Spoken Tamil and audit existing Tamil datasets to measure the representation of the two varieties.

2 Related work

Spoken Tamil. While low varieties of diglossic languages are generally understudied in NLP, there is some previous work on NLP for Spoken Tamil. K and Lalitha Devi (2014) attempted conversion of Spoken Tamil to Literary Tamil using a rule-based system. Nanmalar et al. (2022, 2019) train models to classify diglossic register for Tamil audio. Furthermore, recent work on code-switching in Tamil implicitly uses at least some data in Spoken Tamil, since that is the variety most permissive of code-switching (Chakravarthi et al., 2020, 2021; Banerjee et al., 2018; Mandl et al., 2020).

Diglossia. Diglossia in NLP has largely been studied in the context of Arabic. For example, Zaidan and Callison-Burch (2014); Sadat et al. (2014); Salameh et al. (2018); Bouamor et al. (2019) all train models on the task of Arabic dialect and register classification. However, we were inspired to study diglossia in Tamil by Krishna et al. (2022), the only work on style transfer for Indian languages to our knowledge.

3 Dataset

To study diglossia in Tamil, we created IruMozhi,[2] a dataset of parallel sentences in Literary and Spoken Tamil. We first collected a high-quality set of 499 sentences randomly sampled from a large corpus of scraped Tamil Wikipedia articles, written in Literary Tamil.[3] This initial dataset was then converted to Spoken Tamil by two native-speaker translators. A few examples of the parallel data are presented in Table 1.

3.1 Creation

The dataset from Wikipedia was originally in the Tamil script; however, Spoken Tamil is largely found in the Latin script online. To enable easier comparison to Spoken Tamil and to have parallel romanised training data for both varieties, the dataset was automatically transliterated into the Latin alphabet using a Python program, resulting in the Literary Tamil split of IruMozhi.

Afterwards, two native speaker volunteers, both fluent in Literary and Spoken Tamil, were chosen to translate the sentences into their register of Spoken Tamil. Translator 1 and 2 both grew up in Salem, Tamil Nadu, India, albeit at different times; translator 1 tends to use fewer English loanwords.

The translators were instructed to convert the literary sentences into their register of Tamil while adhering to the original meaning of the sentence as closely as possible. Translator 1 only had access to the Literary sentences (both Tamil and transliterated), whereas Translator 2 had access to Translator 1's conversions as well.

3.2 Augmentation

We also design rules to augment all our data with orthographic variants, resulting in 6,224 Spoken Tamil and 2,410 Literary Tamil sentences. These rules simulate normal orthographic variation in

[2] *IruMozhi* means 'two languages' in Tamil.

[3] The articles were scraped in April 2019 and originally hosted as a Kaggle dataset. We sampled sentences from the first file of the `train` split of the corpus.

Dataset	Ref.	Register	Source	# Lines
IruMozhi	—	Both	Wikipedia	1,497
IruMozhi-AUGMENTED	—	Both	Wikipedia	8,634
Tamilmixsentiment	Chakravarthi et al. (2020)	Spoken?	YouTube	15,744
Offenseval	Chakravarthi et al. (2021)	Spoken?	Social media	39,527
Dakshina	Roark et al. (2020)	Literary	Wikipedia	10,000
HopeEDI	Chakravarthi (2020)	Spoken?	YouTube	18,178
CC-100	Conneau et al. (2020)	Both?	Web	6,243,679

Table 3: Datasets for romanised Tamil that we consider. The register of each corpus is not known in some cases, in which case we indicate our best guess with '?'.

romanised Tamil, which correspond with common characteristics of speech, such as the alternation between ⟨zh⟩ and ⟨l⟩ to represent the voiced retroflex approximant /ɻ/. Other changes include alternating intervocalic ⟨h⟩ and ⟨g⟩, word-initial ⟨ch⟩ and ⟨s⟩, and word-final ⟨le⟩ and ⟨la⟩. Geminated consonants are also shortened to be singled, and long vowels are replaced with other variants. For example, ⟨oo⟩ is replaced with ⟨uu⟩, and ⟨ae⟩ is replaced with ⟨e⟩. Each augmented sentence is added back to the dataset with its respective literary or colloquial tag, as adjusting the orthography should not impact the register of Tamil used in the text. We apply all possible combinations of augmentations to each entry in the dataset.

3.3 Analysis

We measured Levenshtein distance (raw and normalised), BLEU, and chrF between all three pairings of the transliterated Literary Tamil sentences and the two Spoken Tamil translated conversions. The latter two metrics were computed using SACREBLEU (Post, 2018). All metrics are reported in Table 2 and Figure 1. Overall, the two Spoken Tamil translators agree with each other more than they do with Literary Tamil across all of our metrics. However, there is clearly linguistic variation in Spoken Tamil given disagreements between the two translators. For further discussion on linguistic differences between Spoken and Literary Tamil, see appendix A.

4 Experiments

Using IruMozhi, we train models on the task of classifying romanised Tamil text as Literary or Colloquial Tamil. After evaluating our models on a held-out test set, we audit existing datasets of romanised Tamil text to gauge the amount of data available for the two registers. We train two main types of model: **Naïve Bayes** classifiers on n-gram features and **XLM-R** finetuned for sequence classification. For both training and evaluation, we strip punctuation and convert all text to lowercase.

Model		Trained on IruMozhi			
		Acc.	$F1^{ST}$	$F1^{LT}$	$Acc.^{D}$
Gauss. NB	$c = 4$	99.7%	0.998	0.995	52.9%
	$c = 3$	99.8%	0.998	0.996	36.9%
	$c = 2$	99.8%	0.998	0.996	58.5%
Multi. NB	$c = 4$	99.1%	0.994	0.984	70.8%
	$c = 3$	98.7%	0.991	0.978	52.1%
	$c = 2$	98.8%	0.992	0.978	20.3%
XLM-R	base	99.4%	0.996	0.990	81.5%

Table 4: Results averaged over 5 runs, reporting accuracy and per-class F1 on IruMozhi and accuracy on Dakshina (which the models were not trained on). For all Naïve Bayes models we report with $w = 1$. ST and LT refer to Spoken and Literary Tamil splits, respectivity. *For all metrics, larger is better.*

For Naïve Bayes, we featurise our data into char and word n-grams using a sliding window, resulting in a fixed-length vector of counts over features for each text input. We test both Gaussian and Multinomial distributions for the feature likelihood, and tune the maximum n-gram length for characters (c) and words (w) as hyperparameters. We use model implementations from `scikit-learn`.

XLM-R is a 279M-parameter masked transformer language model trained on the CC-100 web text corpus of one hundred languages, including romanised Tamil (Conneau et al., 2020). Using the HuggingFace implementation of `XLMRobertaForSequenceClassification`, we train a classification head on the first token <s>. We finetune the entire model for 4 epochs with a learning rate of $2 \cdot 10^{-5}$ for the Adam optimiser, on a single NVIDIA RTX A6000.

5 Results

We present results in Table 4 (see appendix B for results on more hyperparameters). All model ar-

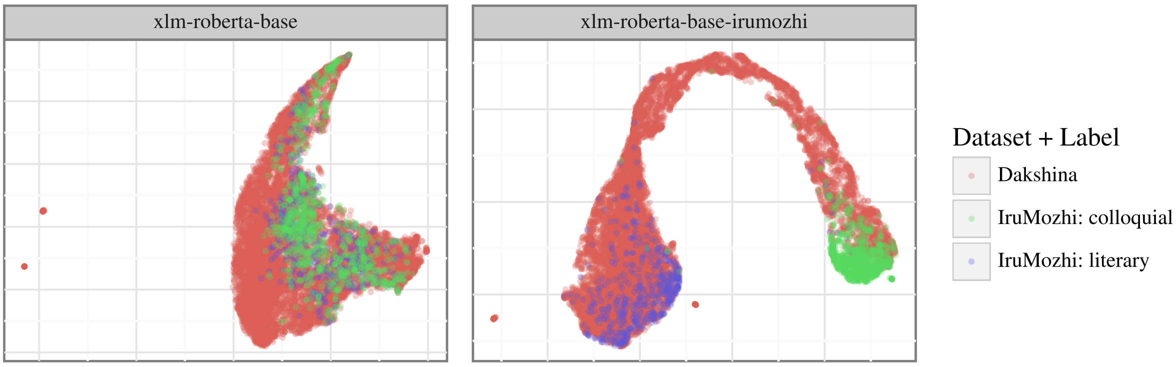

Figure 2: Embeddings of each sentence in IruMozhi and Dakshina taken from XLM-R base (left) and our finetuned version (right), with dimensionality reduced with UMAP (McInnes et al., 2018). Note the separation of Spoken and Literary clusters from IruMozhi after finetuning, with most of the Dakshina data closer to Literary Tamil from IruMozhi. For a different view, see Figure 3.

chitectures reliably converge to near-perfect performance on the held-out portion of IruMozhi.

5.1 Generalisation

It is difficult to decide which classifier is the best due to their similarly high performance on the IruMozhi dev set. Thus, we must further test our models' ability to generalise to another dataset with known labels. Fortunately, the Dakshina dataset (Roark et al., 2020) contains human-translated romanised Literary Tamil from the same data distribution as our dataset (Wikipedia). To measure generalisation ability, we check whether models correctly identify Dakshina to be Literary Tamil when only trained on our dataset.

Finetuning XLM-R leads to the best and most consistent performance on Dakshina. Naïve Bayes models, as one would expect, are less reliable for out-of-domain test data. We plot the resulting embeddings in Figure 2. Some hyperparameter settings for Naïve Bayes reported in appendix B did achieve better accuracy on Dakshina, but due to the highly inconsistent behaviour and complete reliance on orthography (unlike a pretrained language model), we do not suggest using Naïve Bayes approaches for the Tamil variety classification task.

5.2 Out-of-domain audits

Having trained these models, we audited the datasets listed in Table 3 to estimate the proportion of Literary and Spoken Tamil in them. We report these estimates in Table 5. Finetuned XLM-R and Multinomial Bayes ($c = 4, w = 1$) confirm that

Dataset	**XLM-R**	**Multi. NB**
Tamilmixsentiment	6.2%	6.7%
Offenseval	14.1%	19.7%
Dakshina	81.5%	70.8%
HopeEDI	13.1%	20.6%
CC-100	44.0%	13.2%

Table 5: Estimated percentage of Literary Tamil sentences in each available romanised Tamil corpus, according to finetuned XLM-R and Multinomial Naïve Bayes models trained on IruMozhi.

Dakshina is almost entirely Literary Tamil, while Tamilmixsentiment, Offenseval, and HopeEDI are largely Spoken Tamil. Given the genres that these datasets were collected from (formal Wikipedia vs. informal social media), these are reasonable predictions. Finally, testing the first 50k lines, we find a surprisingly high portion of Spoken Tamil in the CC-100 `ta_rom` split. This suggests that XLM-R was indeed trained on a large amount of Spoken Tamil, explaining why our finetuning was successful.

6 Conclusion

We presented IruMozhi, a parallel corpus of Literary and Spoken Tamil translated on Wikipedia text. We trained models on an augmented version of IruMozhi for classifying Tamil diglossia, and audited the composition of existing labelled datasets and the CC-100 pretraining text in romanised Tamil. We found that there are indeed labelled and unlabelled data sources for Spoken Tamil text, indicating hopeful avenues for future NLP research on the variety. Particularly, XLM-R seems to have already been trained on some romanised Spoken Tamil data.

We hope to train style transfer models for the two varieties and study diglossia in other Indian languages. Our aim is to encourage work on lesser-studied languages and dialects in South Asia.

Limitations

This work is one of a handful submitted to *CL venues on Spoken Tamil. However, our definition of Spoken Tamil does not take into consideration dialectal variation in the variety. Particularly, since both of our annotators were from Salem, Tamil Nadu, India, our dataset excludes other regional dialects of Spoken Tamil. This may harm the ability of our trained models to generalise to other dialects of Spoken Tamil. Future work could improve on this paper by collecting translations from a more geographically diverse set of annotators, similar to what has been done in dialectal NLP work on Arabic.

Ethics Statement

We release models for classifying the register of romanised Tamil texts. This could be used to e.g. profile users on social media, but since our classification is not very fine-grained we do not foresee such uses being practical and thus do not have ethical concerns about our models.

References

E Annamalai. 1980. Some syntactic differences between spoken and written Tamil. *South Asian Languages: Structure, Convergence and Diglossia*, pages 289–93.

Elay Annamalai and Sanford B. Steever. 2015. Modern Tamil. In *The Dravidian Languages*, pages 118–175. Routledge.

Aryaman Arora, Adam Farris, Samopriya Basu, and Suresh Kolichala. 2022. Computational historical linguistics and language diversity in South Asia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1396–1409, Dublin, Ireland. Association for Computational Linguistics.

Suman Banerjee, Nikita Moghe, Siddhartha Arora, and Mitesh M. Khapra. 2018. A dataset for building code-mixed goal oriented conversation systems. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3766–3780, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy. Association for Computational Linguistics.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145, Kyiv. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Charles A. Ferguson. 1959. Diglossia. *Word*, 15(2):325–340.

Charles A. Ferguson. 1996. Epilogue: diglossia revisited. *Understanding Arabic: Essays in contemporary Arabic linguistics in honor of El-Said Badawi*, pages 49–67.

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.

Marimuthu K and Sobha Lalitha Devi. 2014. Automatic conversion of dialectal Tamil text to standard written Tamil text using FSTs. In *Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM*, pages 37–45, Baltimore, Maryland. Association for Computational Linguistics.

Kalpesh Krishna, Deepak Nathani, Xavier Garcia, Bidisha Samanta, and Partha Talukdar. 2022. Few-shot controllable style transfer for low-resource multilingual settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7439–7468, Dublin, Ireland. Association for Computational Linguistics.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the HASOC track at FIRE 2020: Hate speech and offensive language identification in Tamil, Malayalam, Hindi, English and German. In *Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 29–32.

Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*.

M Nanmalar, P Vijayalakshmi, and T Nagarajan. 2019. Literary and colloquial dialect identification for Tamil using acoustic features. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, pages 1303–1306. IEEE.

M Nanmalar, P Vijayalakshmi, and T Nagarajan. 2022. Literary and Colloquial Tamil dialect identification. *Circuits, Systems, and Signal Processing*, 41(7):4004–4027.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall. 2020. Processing South Asian languages written in the Latin script: the Dakshina dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2413–2423, Marseille, France. European Language Resources Association.

Fatiha Sadat, Farnazeh Kazemi, and Atefeh Farzindar. 2014. Automatic identification of Arabic dialects in social media. In *Proceedings of the First International Workshop on Social Media Retrieval and Analysis*, pages 35–40.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Vanithamani Saravanan, Seetha Lakshmi, and Imelda S Caleon. 2009. The debate over literary Tamil versus standard spoken Tamil: What do teachers say? *Journal of Language, Identity, and Education*, 8(4):221–235.

Harold F. Schiffman. 1998. Standardization or restandardization: The case for standard spoken Tamil. *Language in Society*, 27(3):359–385.

Harold F. Schiffman. 1999. *A reference grammar of spoken Tamil*. Cambridge University Press.

Omar F. Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

A Linguistic differences between Literary and Spoken Tamil

We briefly discuss the linguistic differences between Literary and Spoken Tamil (Schiffman, 1999). The vowels of Literary Tamil undergo various phonological changes when converted into speech. Vowels, both short monophthongs and diphthongs, are regularly raised in the word-final position. For example, both /-a/ and /-aɪ/ are raised to [-ɛ]. Word-final /u/ (with the exception of names) is shortened to [ɯ]; additionally, an epenthetic-[ɯ] is usually added to the end of words that end with consonants. When not in the word-final position, /e/ and /i/ are relaxed into /ɛ/ and /ɪ/. Additionally, /i/ along with /u/ are lowered to [ɛ] and [o], respectively, when preceding a short consonant followed by /a/ and /aɪ/. Unlike the short vowels, long monophthongs will mostly remain the same quality regardless of position.

Word-final nasal consonants (excluding /ɳ/) also affect preceding vowels. In all cases, the vowel becomes nasalized and the consonant is dropped. For short vowels, however, the nasal may also change the quality of the vowel. For example, /an/ is nasalized to [ã], and then raised to [ɛ̃]. Similarly, /am/ is also nasalized to [ã], but then rounded to [ɔ̃].

Outside of regular vowel changes, various other aspects of Spoken Tamil differ from the literary variety. For example, the locative suffix /-il/ is expressed as [-lɛ]; A suffix like /(-)illaɪ/, indicating negation, is said as [-lɛ] at the end of words and [illɛ] elsewhere. /(-)uɭɭeː/ 'inside' is spoken as [(-)uɭɭɛ]. In some dialects of Spoken Tamil, the 3rd-person irrational ending, /-atu/, can become palatalised to [-Vt͡ʃɯ] in the past tense of strong verbs, with the vowel depending on the verb being conjugated. In general, strong verbs substitute /-tt-/ and /-nt-/ with [-t͡ʃ-] and [-nd͡ʒ-], respectively.

Finally, there are major lexical differences between Spoken and Literary Tamil. For example, there is a large presence of loanwords in the colloquial form of the language, most often taken from English and Sanskrit. These words, alongside some of native Tamil origin, often replace literary words that may seem too formal in speech. An example of this is *ulladhu*, which is almost always replaced with *irukku* in colloquial contexts as the existence copula. Similarly, the Sanskrit loan *sandosham* is preferred over the native Tamil word *magizhcci* for 'happy', although the latter is gaining popularity among the younger generations.

B More results

Model	Params	Trained on IruMozhi				IruMozhi + Dakshina		
		Acc.	$F1^{ST}$	$F1^{LT}$	$Acc.^{Dakshina}$	Acc.	$F1^{ST}$	$F1^{LT}$
Naïve Bayes (Gaussian)	$c=4, w=1$	99.7%	0.998	0.995	52.9%	99.7%	0.998	0.995
	$c=3, w=1$	99.8%	0.998	0.996	36.9%	99.7%	0.998	0.994
	$c=2, w=1$	99.8%	0.998	0.996	58.5%	99.8%	0.999	0.997
	$c=1, w=1$	99.4%	0.996	0.989	91.3%	99.2%	0.995	0.987
	$c=0, w=1$	99.4%	0.996	0.990	1.7%	99.4%	0.996	0.988
	$c=4, w=0$	99.1%	0.994	0.984	48.7%	99.5%	0.996	0.991
	$c=3, w=0$	94.3%	0.959	0.906	29.5%	93.9%	0.956	0.901
	$c=2, w=0$	67.3%	0.708	0.628	43.2%	68.2%	0.718	0.636
	$c=1, w=0$	72.2%	0.761	0.561	29.9%	40.8%	0.330	0.470
Naïve Bayes (Multinomial)	$c=4, w=1$	99.1%	0.994	0.984	70.8%	99.1%	0.994	0.984
	$c=3, w=1$	98.7%	0.991	0.978	52.1%	98.4%	0.989	0.971
	$c=2, w=1$	98.8%	0.992	0.978	20.3%	99.0%	0.993	0.981
	$c=1, w=1$	99.1%	0.993	0.983	2.2%	99.0%	0.993	0.981
	$c=0, w=1$	99.0%	0.993	0.982	74.0%	98.4%	0.989	0.972
	$c=4, w=0$	98.7%	0.991	0.977	76.0%	98.1%	0.987	0.966
	$c=3, w=0$	98.0%	0.986	0.965	65.4%	98.6%	0.990	0.974
	$c=2, w=0$	94.3%	0.960	0.902	50.9%	94.2%	0.959	0.901
	$c=1, w=0$	82.0%	0.880	0.643	34.8%	82.6%	0.884	0.655
XLM-R		99.4%	0.996	0.990	81.5%	99.1%	0.990	0.991

Table 6: Results on more hyperparameter settings.

We trained many variants of Naïve Bayes models, but their erratic generalisation behaviour on Dakshina led us to focus on XLM-R in the main text. We also tried training on IruMozhi and Dakshina together, but this heavily skewed the data distribution towards Literary Tamil since Dakshina is much larger than

IruMozhi, and seemed to harm out-of-domain generalisation; estimated Literary Tamil percentages on other datasets were around 50% and thus basically random.

C UMAP with predicted labels

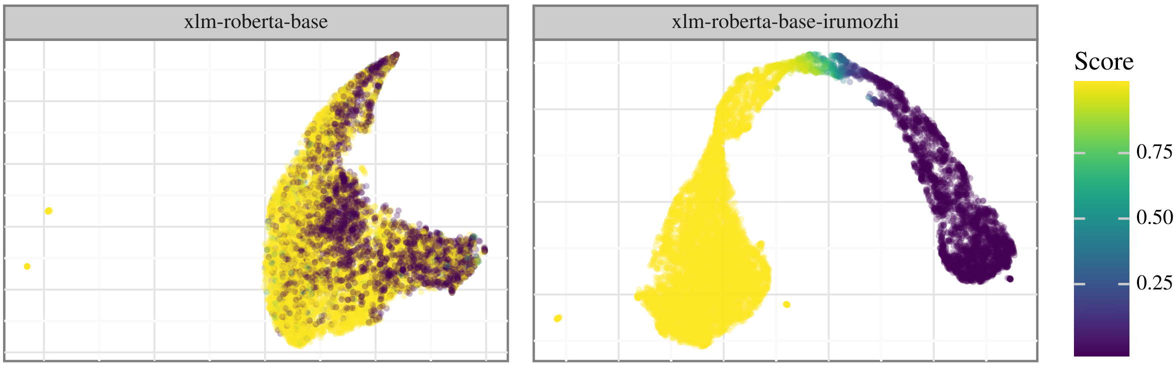

Figure 3: Same embedding map as Figure 2 but with predicted probability for Literary Tamil by XLM-R finetuned on IruMozhi instead of dataset and label. There is no apparent structure in the base XLM-R model.