Dialectometry and Evaluation of the ePark Corpus for Low-Resource Formosan Language Dialects

Henry Gagnier
Pittsford Sutherland High School
Pittsford, NY, USA
henrygagnier9@gmail.com

Abstract

Formosan languages are a critically endangered branch of the Austronesian family spoken in Taiwan, and many of their dialects remain poorly understood and computationally understudied. Subgrouping relationships in these languages are often contested and unresolved. We provide the first evaluation of the ePark corpus as a dialectal NLP resource, identifying its strengths and gaps for future NLP work, and present the first large-scale corpus-based computational analysis of dialect similarity across all officially recognized Formosan languages. We use the ePark corpus to analyze 42 dialects in 16 Formosan languages, and through word-level TF-IDF cosine similarity, Jaccard similarity over shared vocabulary, and Levenshtein distance, we quantify pairwise dialectal relationships within the Amis, Atayal, Seediq, Bunun, Paiwan, Rukai, and Puyuma languages. We find that simple lexical similarity methods can recover and confirm linguistically established dialectal subgroupings. We find that in multiple cases the two metrics diverge, offering insights on contested subgroupings such as Mantauran Rukai. This work establishes a scalable methodological framework for dialectometry in low-resource languages, demonstrates the value of the ePark corpus for Formosan NLP research, and encourages future work in NLP on Formosan dialects.

Keywords: dialectometry, Formosan languages, dialects, ePark corpus, corpus linguistics

1. Introduction

Formosan languages are a primary branch of the Austronesian language family spoken by under 200,000 people in Taiwan (Lin et al., 2025; Li, 2008). Despite their importance to historical linguistics (Zeitoun and Goudin, 2024), most Formosan languages are severely under-resourced and endangered, with many exhibiting substantial internal variation. Formosan dialects often differ in lexicon, phonology, and morphosyntax, reflecting complex histories of migration, contact, and isolation (Li, 2014). Understanding relationships between these dialects is crucial to linguistic classification and documentation and supports the development of language technology that can enable preservation, revitalization, and education efforts for Formosan languages (Zeitoun et al., 2003; Meighan, 2021).

Computational methods in dialectology known as dialectometry have emerged as a promising method for dialect analysis in low-resource languages, aggregating a large number of individual linguistic features to produce satisfactory results (Bompolas, 2023). Rama et al. (2017) uses Levenshtein distance and autoencoders to analyze Gondi dialects. (Dunn, 2021) analyzes the dialects of high-resource languages using two Spearman frequency-based similarity measures across multiple domains. (Kwaik et al., 2018) uses the Jaccard similarity, a vector space model (VSM), latent semantic indexing, Hellinger distance, and correlation coefficients between frequent words. Bompolas and Melissaropoulou (2025) uses grammatical and lexical features to quantify dialect differences in Inner Asia Minor Greek varieties. These computational methods provide extremely important insights, especially for dialects without current support in natural language processing (NLP).

Research on Formosan languages in NLP is beginning but is still extremely limited. Zheng et al. (2024) investigates machine translation (MT) methods for low-resource Formosan languages. (Lin et al., 2025) develops FormosanBench, a benchmark for MT, automatic speech recognition (ASR), and summarization in Atayal, Paiwan, and Amis, and finds that existing LLMs perform poorly on Formosan NLP tasks. (Hsieh et al., 2024) uses self-supervised models to obtain and analyze their speech representations and develop a model to classify Formosan languages, and uses the speech representations to construct a linguistic phylogeny of Formosan languages. While work has begun in dialectometry and Formosan languages, Formosan dialects and their dialectal relationships have not been studied in NLP. Considering the importance of dialectometric analysis and the absence of work on Formosan language dialects in NLP, it is necessary to begin computational work on these under-resourced Formosan dialects.

The purpose of this study is to (1) provide the first large-scale corpus-based computational measurement of dialect similarity across all major For-

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 124–134
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

mosan languages, (2) test whether lexical similarity measures recover relationships established in prior research, and (3) provide an evaluation of the ePark corpus as a dialectal resource. This study also aims to improve the inclusion of low-resource Formosan languages and dialects in NLP research and contribute to the evaluation of dialectal corpus resources, particularly the ePark corpus, as a NLP resource for language and dialect analysis.

2. Methodology

2.1. Data

We use the ePark corpus[1] (Mohamed et al., 2024; Indigenous Languages Research and Development Foundation, 2020), a high-quality, comprehensive resource for the preservation, learning, and revitalization of Indigenous languages of Taiwan. It was developed by the Indigenous Languages Research and Development Foundation and covers 42 official dialects of the 16 different Formosan languages being Amis, Atayal, Saisiyat, Thao, Seediq, Bunun, Paiwan, Rukai, Truku, Kavalan, Tsou, Kanakanavu, Saaroa, Puyuma, Yami, and Sakizaya. The corpus is split into sections for vocabulary, situational dialogue, cultural texts, short passages, daily conversation, and reading and writing, and is composed of raw Formosan language text.

The full corpus contains 2680175 tokens and 587 hours of audio and translations into Mandarin, enabling language preservation, education, and research. For each dialect, we take all the text from the six sections and use this text for analysis.

2.2. Preprocessing

For each dialect corpus, we convert all text to lowercase, remove punctuation by replacing non-alphanumeric characters with whitespace with the exception of and ’ as these are used in multiple Formosan languages. We then normalize all whitespace by collapsing runs of whitespace to a single space and removing leading and trailing whitespace. Finally, we perform tokenization by splitting text on boundaries for word-level tokens. Reliable morphological analysers do not exist for most Formosan languages, so we do not perform any segmentation beyond word-level tokenisation.

2.3. Lexical Similarity Measures

We compute three lexical similarity measures for each pair of dialect varieties within a language group, excluding identical pairs.

[1] https://ai4commsci.gitbook.io/formosanbank/

First, we compute word-level TF-IDF cosine similarity to capture the weighted similarity of each dialect. We implemented TF-IDF vectors and cosine similarity using `scikit-learn` (Pedregosa et al., 2011). We set `min_df` to 2 in order to exclude terms appearing in only one document and limit the effect of noise that likely does not affect dialect-level differences. We set `sublinear_tf` to True in order to reduce the influence of high-frequency words, which tend to carry little discriminative information for dialect analysis (Manning et al., 2008).

We then compute Jaccard similarity over the presence of word types to measure unweighted vocabulary overlap. We reduce noise from low-frequency terms, which are likely to be noise or corpus-specific artifacts, by restricting the vocabulary to words with a corpus frequency of at least 5 in each dialect.

Third, we compute character-level Levenshtein distance on the 500 most frequent word types in each dialect, averaged symmetrically across both directions and normalized by word length to capture formal similarity between word types beyond the binary presence/absence distinctions made by Jaccard.

To assess the statistical reliability of our similarity estimates and control for corpus size variation across dialects, we compute bootstrap confidence intervals for each language group. For each dialect pair, we draw 500 bootstrap samples of 10,000 tokens with replacement from each dialect’s token list, fixing the sample size across all pairs to ensure comparability. We recompute TF-IDF cosine similarity, Jaccard similarity, and Levenshtein for each sample and report the mean width of the 95% confidence interval across all pairs within each language group as a measure of estimate stability. To encourage further research on low-resource Formosan dialects and increase reproducibility, we release code at https://github.com/henrygagnier/formosan-dialectometry.

3. Results

3.1. Corpus Statistics

We first look at the overall corpus statistics from ePark (Table 1), which summarize the size and diversity of each dialect. Token counts and types across all dialects and languages are comparable, ranging from 46,177 to 67,968 and 9,530 to 13,390, respectively. Token-type ratios are also fairly consistent, ranging from 0.184 to 0.240. Seven of the sixteen languages have multiple officially recognized dialects. Hapax rates are high across all dialects, ranging from 0.721 to 0.793, indicating that a large proportion of each dialect’s vo-

cabulary is represented by a single occurrence in the corpus. This sparsity is a corpus-wide property and motivates the restriction of Jaccard similarity to types with a frequency greater than 5.

The vocabulary section constitutes between 21.9% and 32.6% of tokens per dialect, suggesting that section composition does not introduce large cross-dialect imbalances due to the vocab size consistency. Among languages with multiple dialects, the shared core vocabulary of the types appearing in all dialects at a frequency greater than 5 varies from 8.8% to 24.5%. The dialect-unique vocabulary, or types that appear in only one dialect ranges from 53.8% to 79.2%. This implies that languages with higher shared core percentages such as Paiwan and Seediq are good candidates for cross-dialect model transfer, while Atayal and Rukai's high dialect-unique percentage suggest the dialect-specific resources are necessary for effective language technology.

3.2. Dialect Similarity

We now look at the overall similarity of each Formosan language with multiple dialects (Table 2). We find that the Seediq dialects are the most similar among the languages based on cosine similarity, with a mean cosine similarity of 0.820. We also find that the Atayal and Rukai dialects differ the most among Formosan languages, with a mean cosine similarity of 0.605 and 0.648, respectively. Paiwan, Amis, Puyuma, and Bunun all have moderate variance, with the mean cosine similarity ranging from 0.676 to 0.781 and the mean Jaccard similarity ranging from 0.333 to 0.447. Bunun, Puyuma, Amis, Paiwan, and Seediq all had relatively similar Levenshtein distances ranging from 0.830 to 0.862. Rukai and Atayal had lower Levenshtein distances of 0.761 and 0.792, respectively.

Looking at the cosine similarity by section (Table 3), we see that the vocabulary section is the most similar in all languages, with cosine similarity scores of 0.811 to 0.930, while the cultural and reading and writing sections are the most different, with cosine similarity scores between 0.412 and 0.672 and 0.425 and 0.708, respectively. Due to the large differences in dialectal similarity in cultural, reading and writing, and daily conversation topics, in dialectal analysis, similarity and distance scores must be compared only with scores computed using the same set of topics, and NLP work must consider these differences.

We now zoom into the per-dialect pairwise TF-IDF cosine similarity, Jaccard similarity, and Levenshtein distance for each Formosan language (Figure 1).

In Amis, it can be seen that the Central, Chengkung-Kwangshan, and Southern dialects are similar with pairwise cosine values ranging from 0.773 to 0.828. The Northern Amis dialect is an outlier, with cosine similarities to other dialects ranging from 0.640 to 0.684, and its Jaccard similarities with other dialects are lower than other pairs, ranging from 0.323 to 0.345. Tavalong-Vataan has moderate relative cosine similarity with other languages, ranging from 0.736 to 0.793. This computational finding aligns with Li and Tsuchida (2022) which finds that Amis dialects in the north are far more diverse than central and southern Amis dialects. Li and Tsuchida (2022) also finds that the Central and Southern Amis dialects are extremely close to each other and that Tavalong-Vataan can be grouped with either the Northern or Southern dialects through lexical differences. Atayal has the lowest mean cosine similarity of the Formosan languages, reflecting high diversity among its dialects. We find that the Skikun-Squliq and the Matu'uwal-Plngawan dialect pairs have the high cosine similarities of 0.728 and 0.719, respectively. Interestingly, while the Skikun-Squliq fair has a high cosine similarity and Jaccard similarity, the Matu'uwal-Plngawan pair has a low Jaccard similarity of 0.238. Other pairs with relatively high cosine similarity include S'uli-Klesan and S'uli-Skikun. The Matu'uwal-Klesan pair has the lowest cosine similarity of 0.526. (Goderich, 2020) groups Atayal into two main subgroups: Northern Atayal and Southern Atayal. Northern Atayal is further grouped into Matu'uwal and Nuclear Northern Atayal (Skikun and Squliq), and Southern Atayal is further grouped into Plngawan and Nuclear Southern Atayal (Klesan, S'uli). The high cosine similarity of Skikun-Squliq is consistent with the classification of the dialects as Nuclear Northern Atayal, which sets them apart from Matu'uwal. Despite being a member of the proposed Northern Atayal subgroup, Matu'uwal has a fairly low cosine similarity with the other members of the Northern Atayal group and a high cosine similarity with Plngawan, a Southern Atayal dialect. Conversely, the low Jaccard similarity and Levenshtein distance between Matu'uwal supports its grouping as a non-Northern Atayal dialect.

Seediq dialects have the highest mean cosine similarity of any Formosan language of 0.820, indicating that dialects are fairly similar. Toda and Truku are the most similar, with a cosine similarity of 0.895 and a high Jaccard similarity of 0.425. The Tgdaya-Truku and Tgdaya-Toda pairs have a lower cosine similarity of 0.765 and 0.799, respectively. While lower, these dialects are still fairly similar compared to dialect similarities in other Formosan languages. (Ogawa and Asai, 1935) identifies Paran/Tgdaya Seediq and Truku Seediq as the two main dialectal groups of Seediq, with Truku Seediq encompassing Truku and Toda. Our computational findings match this finding, with Toda

Language	Dialect	Tokens	Types	TTR	Hapax rate	Vocab%	Core%	Uniq%
Amis	Northern	47,885	9,979	0.208	0.791	31.6	19.6	53.8
	Tavalong-Vataan	51,790	10,767	0.208	0.774	29.7		
	Central	50,338	10,341	0.205	0.770	29.9		
	Chengkung-Kwangshan	46,967	10,315	0.220	0.793	31.9		
	Southern	50,603	10,727	0.212	0.771	30.1		
Atayal	Squliq	56,282	10,853	0.193	0.737	26.8	8.8	71.0
	S'uli	54,378	10,274	0.189	0.742	28.0		
	Matu'uwal	57,312	10,819	0.189	0.756	27.2		
	Plngawan	46,177	9,009	0.195	0.764	32.6		
	Skikun	57,914	10,824	0.187	0.728	26.1		
	Klesan	54,355	10,200	0.188	0.779	28.1		
Saisiyat	Saisiyat	52,442	9,662	0.184	0.759	29.5	—	—
Thao	Thao	49,846	9,530	0.191	0.750	29.4	—	—
Seediq	Toda	51,636	9,957	0.193	0.750	29.3	23.8	61.5
	Tgdaya	51,686	10,028	0.194	0.763	29.1		
	Truku	58,746	10,872	0.185	0.748	25.7		
Bunun	Takbunuaz	49,958	10,715	0.214	0.774	30.4	15.2	61.0
	Takivatan	51,953	11,107	0.214	0.772	29.2		
	Takituduh	47,699	10,391	0.218	0.784	32.1		
	Takibaka	51,127	10,526	0.206	0.771	30.3		
	Isbukun	47,868	10,409	0.217	0.774	31.6		
Paiwan	Southern/Central	49,376	10,536	0.213	0.770	30.4	24.5	56.7
	Northern	59,902	12,633	0.211	0.769	25.3		
	Central	59,479	12,353	0.208	0.767	25.5		
	Eastern	60,056	12,745	0.212	0.781	25.5		
Rukai	Tanan	48,165	10,600	0.220	0.778	31.3	9.3	79.2
	Budai	47,800	10,835	0.227	0.796	31.4		
	Labuan	48,737	10,634	0.218	0.769	30.8		
	Tona	52,733	11,469	0.217	0.769	29.0		
	Maga	50,614	12,168	0.240	0.781	29.8		
	Mantauran	56,976	12,620	0.221	0.777	26.6		
Truku	Truku	51,578	9,747	0.189	0.754	29.3	—	—
Kavalan	Kavalan	57,141	11,026	0.193	0.721	25.7	—	—
Tsou	Tsou	59,577	10,972	0.184	0.740	25.9	—	—
Kanakanavu	Kanakanavu	51,278	10,467	0.204	0.746	29.4	—	—
Saaroa	Saaroa	56,707	12,492	0.220	0.763	26.6	—	—
Puyuma	Nanwang	67,968	13,390	0.197	0.740	21.9	14.8	63.9
	Katipul	56,315	11,186	0.199	0.748	27.0		
	Kasavakan	56,943	11,051	0.194	0.733	26.8		
	Ulivelivek	52,773	10,839	0.205	0.760	28.9		
Yami	Yami	52,209	10,894	0.209	0.775	26.7	—	—
Sakizaya	Sakizaya	50,565	10,477	0.207	0.773	29.9	—	—

Table 1: Corpus statistics and resource evaluation metrics for all Formosan languages and dialects in the ePark corpus.

and Truku being extremely similar and Tgdaya being relatively different from the Truku Seediq dialects.

Bunun dialects have a moderate relative variance with a mean cosine similarity of 0.676. Computationally, we find that the Takivatan-Takbunuaz and Takituduh-Takibaka dialect pairs have the highest cosine similarities of 0.774 and 0.767, respectively. Moreover, these two pairs have the highest Jaccard similarity in Bunun as well. Other pairs have low or moderate relative similarity with cosine similarities ranging from 0.619 to 0.674. Ogawa and Asai (1935) proposes three subgroups of Bunun: Southern Bunun (Ishbukun), Central Bunun (Takbanuaz and Takivatan), and Northern Bunun (Takibaka, Takituduh). Our results confirm this grouping with extreme precision, with the Takivatan-Takbunuaz and Takituduh-Takibaka

		Cosine similarity			Jaccard similarity			Levenshtein distance			
Language	**N**	Mean	SD	CI	Mean	SD	CI	Mean	SD	CI	**Core%**
Seediq	3	0.820	0.055	0.018	0.362	0.047	0.015	0.848	0.024	0.005	23.8
Paiwan	4	0.781	0.015	0.019	0.434	0.015	0.014	0.861	0.011	0.004	24.5
Amis	5	0.728	0.064	0.019	0.447	0.098	0.015	0.862	0.035	0.007	19.6
Puyuma	4	0.726	0.068	0.017	0.333	0.087	0.014	0.830	0.042	0.005	14.8
Bunun	5	0.676	0.051	0.018	0.366	0.041	0.014	0.844	0.016	0.006	15.2
Rukai	6	0.648	0.063	0.014	0.265	0.048	0.012	0.761	0.038	0.006	9.3
Atayal	6	0.605	0.062	0.016	0.270	0.045	0.013	0.792	0.028	0.007	8.8

Table 2: Summary of pairwise dialect similarity by language group

Language	Vocabulary	Situational	Cultural	Short pass.	Daily conv.	Read/Write	All sections
Seediq	**0.930**	0.876	0.672	0.881	0.733	0.708	0.820
Paiwan	**0.888**	0.845	0.603	0.816	0.661	0.634	0.781
Amis	**0.853**	0.771	0.534	0.777	0.559	0.527	0.728
Puyuma	**0.861**	0.788	0.576	0.748	0.612	0.581	0.726
Bunun	**0.847**	0.724	0.520	0.727	0.484	0.447	0.676
Rukai	**0.847**	0.727	0.439	0.696	0.465	0.439	0.648
Atayal	**0.811**	0.662	0.412	0.643	0.455	0.425	0.605
Mean	*0.863*	*0.770*	*0.537*	*0.755*	*0.567*	*0.537*	*0.712*

Table 3: Mean pairwise TF-IDF cosine similarity between dialects, computed separately for each ePark corpus section and across all sections combined

pairs lining up exactly with the Central Bunun and Northern Bunun pairs. This precise recovery of an established three-subgroup structure from lexical similarity demonstrates the potential of corpus-based dialectometry as a scalable complement to traditional methods in dialectology.

Paiwan's dialects are very similar, with a mean cosine similarity of 0.781. The differences in similarity in dialect pairs are extremely small, with a standard deviation of the cosine similarity being 0.015 and cosine similarity ranging from 0.765 to 0.802. (Ho, 1978) proposed the grouping of Paiwan in Southeast and Northwest Li (1999) concluded that the internal relationship of Paiwan still remains unclear. Computationally, we are unable to make conclusions regarding the subgrouping of Paiwan, as all dialect pairs have similar cosine similarity.

Rukai dialects vary greatly, with a mean cosine similarity of 0.648 and the lowest mean Jaccard similarity of all Formosan languages of 0.265. The Budai and Labuan dialects have a high similarity of 0.802. Tona, Mantauran, and Maga have relatively moderate to high cosine similarities with each other, ranging from 0.674 to 0.714. Looking at Jaccard similarity, different trends emerge. Budai and Labuan are consistently the most similar, while Tona, Mantauran, and Maga have low Jaccard similarities. Labuan, Tanan, and Budai have moderate-high Jaccard similarities, and Mantauran is an outlier with low Jaccard similarity with all other Rukai dialects. Dialectal relationships among Rukai dialects are under debate, with Tanan, Budai, and Labuan being designated as the Eastern Rukai Group and Tona and Maga in the Western Rukai Group (Hsin, 2000; Chen, 2011). Computationally, through Jaccard similarity, we are able to see the Western Rukai Group clearly. We are also able to see the Eastern Rukai Group using cosine similarity. The position of Mantauran is controversial, with hypotheses that Mantauran groups with Tona and Maga (Li, 1997), Mantauran is the first offshoot of the Rukai family (Starosta, 1994, 1995), and that Mantauran forms a subgroup with Tanan, Budai, and Labuan (Zeitoun, 2007). We find that Mantauran aligns with the Western Rukai Group through cosine similarity. In contrast, we find that Mantauran does not align with either group using Jaccard similarity and Levenshtein distance, supporting Starosta (1994).

Puyuma has moderate dialectal variation with a mean cosine similarity of 0.726. We find that the Katipul and Kasavakan dialects are the most similar, with a cosine similarity of 0.821. The Katipul-Ulivelivek and Kasavakan-Ulivelivek pairs also have high cosine similarity and Jaccard similarity. The Puyuma Nanwang dialect stands out as an outlier with low cosine similarity scores and low Jaccard similarity scores. This matches with established work, which groups the Nanwang dialect as an early offshoot of the Puyuma and as the most conservative Puyuma dialect (kuei Li, 1991; hsin

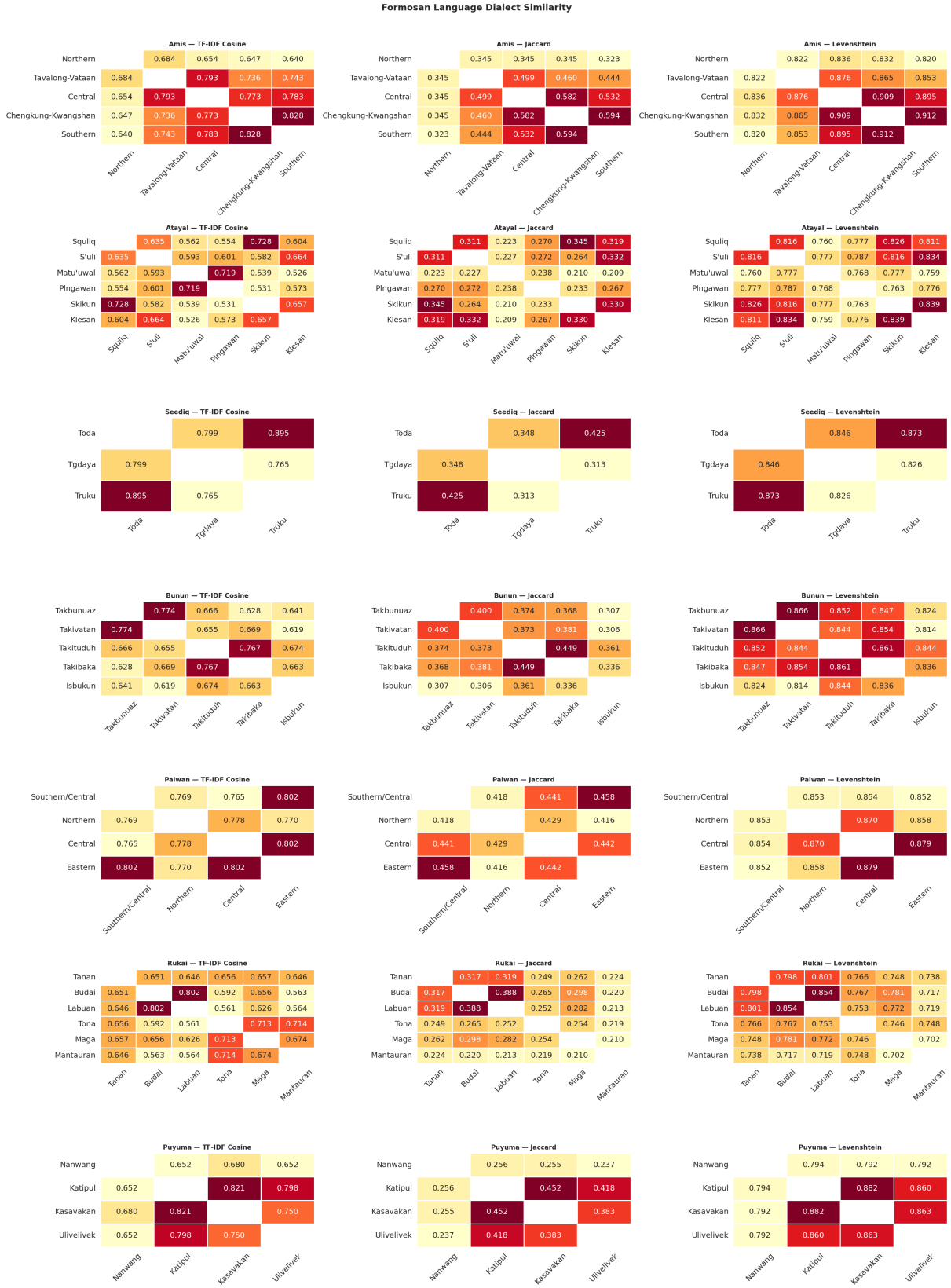

Figure 1: Per-dialect TF-IDF cosine similarity, Jaccard similarity, and Levenshtein distance for all seven Formosan languages with multiple dialects

Ting, 1978).

4. Discussion

4.1. ePark as a Dialectal NLP Resource

The ePark corpus was well-suited to this study in many ways. ePark is publicly accessible, which is

significant for low-resource dialect research when resources are often difficult to obtain. It comprehensively offers data on all Formosan languages and dialects in the same format, with text separated into six thematic sections, allowing for accurate comparisons across dialects without bias from domain differences. Each dialect had a similar token count, ensuring that imbalances in size did not greatly affect results. The corpus also focuses on language education and revitalization, meaning that it prioritizes everyday vocabulary, which is useful for dialect comparison.

However, our analysis identifies limitations that researchers should consider. The consistently high hapax rates indicate that the corpus captures a relatively shallow slice of each dialect's vocabulary. Many word types are attested only once, limiting the reliability of frequency-based similarity measures and making morphological analysis, which would require multiple attestations of inflected forms, especially difficult, unless measures are taken to mitigate these issues, such as only analyzing words of moderate and high frequency. The vocabulary section contributes such a high percentage of the tokens in each corpus, and similarity scores are highest among vocabulary, meaning that the vocabulary may overrepresent shared core vocabulary. The goals of education and language revitalization in ePark may present some challenges. Text in the corpora is curated and may underrepresent regional vocabulary that would be useful for measuring dialect similarity. The corpus also only contains official dialects and languages, meaning that all Formosan language communities and dialects may not be represented. While not utilized in this study, the ePark corpus provides audio for all text corpora, enabling further dialect research. Despite these limitations, the section structure of ePark is an advantage, allowing researchers to control for potential biases and obtain similarity estimates on text closer to naturalistic speech.

Based on our evaluation, we offer four recommendations for researchers using the ePark corpus for dialectal NLP work. First, for dialect similarity and dialect identification tasks, we recommend using the full corpus or situational dialogue and short passage sections, as analysis of the vocabulary section may underrepresent differences across Formosan dialects. Second, researchers should apply minimum frequency thresholds when computing vocabulary-based metrics due to the high hapax rates, meaning that many low-frequency types may be corpus noise rather than dialect signals. Third, for languages with a high percentage of unique vocabulary, dialect-specific models or fine-tuning are recommended instead of cross-dialect transfer, while in languages with high shared core percentages, transfer learning should be tested. Fourth, the audio data provided in the ePark corpus for all text corpora is a significantly underutilized resource that can enable phonological and ASR research in future work. We recommend ePark as a strong resource for Formosan dialectal NLP, particularly for dialect similarity modeling, dialect identification, and transfer learning experiments, while we note that tasks that require morphological annotation and speaker metadata will need supplementary resources.

4.2. Lexical Similarity as a Tool for Formosan Dialectometry

We perform the first corpus-based lexical computational analysis of dialect similarity for Formosan languages. We find that this approach can effectively capture dialect relationships in Formosan languages at scale. Using TF-IDF cosine similarity, Jaccard similarity, and Levenshtein distance computed on the ePark corpus, we find and quantify meaningful variation within and across language groups. Seediq dialects are the most internally similar, while Atayal and Rukai have the greatest internal diversity. These findings have implications for NLP development for low-resource Formosan languages and their dialects. Dialect-transfer approaches may work better for languages with high internal similarity, such as Seediq and Paiwan. Conversely, outlier dialects such as Northern Amis and highly differing languages may require dialect-specific resources for effective NLP. This work also displays the value of the ePark corpus that enables quantitative linguistic analysis on low-resource dialects and Formosan languages.

Through this study, we sought to assess whether computational lexical similarity measures recover subgrouping established through traditional linguistic methods. Among most groups, our results are extremely close to prior scholarship. For Amis, our finding that Northern Amis is an outlier from Central and Southern dialects is consistent with Li and Tsuchida (2022). For Seediq, the high similarity between Toda and Truku dialects and the divergence of Tgdaya matches the two-branch grouping proposed by (Ogawa and Asai, 1935). In Bunun, our results recover the three-subgroup structure of Southern, Central, and Northern Bunun proposed by Ogawa and Asai (1935) with precision as the highly similar Takivatan-Takbunuaz and Takituduh-Takibaka pairs correspond to Central and Northern Bunun. For Puyuma, the outlier status of Nanwang aligns with its characterization as an early offshoot and the most conservative Puyuma dialect (kuei Li,

1991; hsin Ting, 1978). These similarities give empirical support to the validity of corpus-based computational approaches as a complement to traditional methods in Formosan linguistics.

In Atayal, Matu'uwal has a high cosine similarity with Plngawan, a Southern Atayal dialect, despite its classification as a Northern Atayal dialect (Goderich, 2020). Matu'uwal's low Jaccard similarity with Plngawan suggests that while the two dialects may share high-frequency vocabulary, their overall vocabulary is limited. The divergence between cosine similarity, Jaccard similarity, and Levenshtein similarity demonstrates that each metric captures complementary aspects of lexical similarity, and that neither measure provides a full picture of dialect relationships. Cosine similarity, weighted by TF-IDF, is more sensitive to frequent shared words, while Jaccard similarity captures the broader vocabulary distribution equally. This suggests that dialects can share a common high-frequency core while diverging substantially in their wider lexicons. In Rukai, the position of Mantauran in the Rukai dialect subgrouping is contested in previous work. Our results do not resolve this debate but offer partial evidence as cosine similarity places Mantauran with the Western Rukai Group, while Jaccard similarity and Levenshtein similarity sets Mantauran apart from other Rukai dialects, which is consistent with proposals that Mantauran is the first offshoot of the Rukai language Starosta (1994). For Paiwan, dialect pairs are extremely similar, so we were unable to resolve Paiwan's structure computationally, attesting to the difficulty of determining Paiwan subgroupings (Li, 1999).

We provide the first large-scale computation analysis of dialect similarity across all recognized Formosan languages and dialects, using the ePark corpus to compute TF-IDF cosine similarity, Jaccard similarity, and Levenshtein distance. Our results confirm subgroupings established in prior linguistic research while providing computational insight into unresolved Formosan dialect subgroupings, suggesting that these methods can validate and complement traditional linguistic approaches. We hope to further encourage work on Formosan languages and dialects as an important but underresearched domain in NLP research, and we encourage researchers to use the ePark corpus for future work on low-resource language technologies.

5. Future Work

In future work, phonological and morphological features should be incorporated into computational approaches, which may allow for more accurate similarity estimates for languages where cosine similarity and Jaccard similarity are in disagreement. Formosan language morphological analysis tools should be developed, which would enable future research on low-resource Formosan dialects. These differences between cosine and Jaccard similarity in many cases should be investigated to see if they stem from linguistic properties or corpus composition. Computational analysis on Formosan dialects should include syntactic or pragmatic features to provide a more complete view of Formosan dialect relationships computationally. This work should also be extended to support downstream NLP tasks such as dialect identification, cross-dialect transfer learning, and automatic speech recognition (ASR), where the similarity estimated could inform models for extremely under-resourced Formosan language dialects. Future work should also compare corpus-based lexical similarity measures against neural and embedding-based approaches, such as multilingual language model representations, to assess whether more expressive similarity measures provide additional insight into Formosan dialect relationships. Corpora of dialect-specific text should continue to be developed and evaluated, especially for Formosan languages, due to the importance of researching and supporting dialects in NLP.

6. Conclusion

This study presents the first computational analysis of dialect similarity in low-resource Formosan languages and an evaluation of the ePark corpus for dialect analysis. We computed word-level TF-IDF cosine similarity, Jaccard similarity, and Levenshtein distance over the presence of word types to analyze dialectal relations across all recognized Formosan language dialects using data from the ePark corpus.

We find that relatively simple, corpus-based lexical similarity measures can create meaningful insights even in low-resource and endangered language settings. Our findings show that dialectometry can recover established subgroupings and further work on cases of unresolved or contested subgroupings. The divergence in the two lexical similarity methods demonstrates the importance of using multiple metrics when computationally analyzing dialect relationships. This work also establishes a scalable methodological framework for future computational dialectal analysis on Formosan and other low-resource languages, providing a step towards endangered language documentation with modern NLP techniques.

This research creates a foundation for Formosan language dialects in NLP, strengthens and complements traditional linguistics work on For-

mosan dialects. We recommend ePark as a starting point for Formosan dialectal NLP, particularly for dialect similarity modeling and dialect identification, while noting that downstream tasks requiring morphological annotation or speaker metadata will need supplementary resources. These findings and evaluations contribute to existing work on dialectometry for Formosan languages and advance the inclusion of low-resource Formosan languages and dialects in NLP research.

7. Limitations

There are several limitations that should be acknowledged in this study. First, our analysis relies on lexical similarity measures, which capture one dimension of linguistic relatedness. which may not completely reflect the underlying linguistic structure. Second, we do not perform morphological segmentation beyond word-level tokenization, which may lead to similar tokens being treated differently and limit estimate accuracy. Finally, our study covered the 16 officially recognized Formosan languages and their 42 officially recognized dialects in ePark. This official classification may not align with community-based and linguistic distinctions and varieties that are not officially recognized, and are not included in this analysis, limiting the generalizability of results in this study to all Formosan dialects.

8. Bibliographical References

Stavros Bompolas. 2023. *Computational Dialectology in the Linguistic Varieties of Cappadocian, Pharasiot, and Silliot*. Ph.D. thesis, University of Patras.

Stavros Bompolas and Dimitra Melissaropoulou. 2025. Understanding Dialectal Variation in Contact Scenarios Through Dialectometry: Insights from Inner Asia Minor Greek. *Languages*, 10(1):13.

C. M. Chen. 2011. Phonetic Evidence for the Contact-Induced Prosody in Budai Rukai. *Concentric: Studies in Linguistics*, 37(2):123–154.

Jonathan Dunn. 2021. Representations of Language Varieties Are Reliable Given Corpus Similarity Measures. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 28–38, Kyiv, Ukraine. Association for Computational Linguistics.

Andre Goderich. 2020. *Atayal Phonology, Reconstruction, and Subgrouping*. PhD dissertation, Institute of Linguistics, National Tsing Hua University, Hsinchu, Taiwan. Advisors: Hsiu-Chuan Liao and Hui-Chuan J. Huang.

Dah-an Ho. 1978. A Preliminary Comparative Study of Five Paiwan Dialects. *Bulletin of the Institute of History and Philology, Academia Sinica*, 49(4):565–681.

Shu-Kai Hsieh, Yu-Hsiang Tseng, Da-Chen Lian, and Chi-Wei Wang. 2024. Self-Supervised Learning for Formosan Speech Representation and Linguistic Phylogeny. *Frontiers in Language Sciences*, 3.

T. H. Hsin. 2000. *Aspects of Maga Rukai Phonology*. Ph.d. dissertation, University of Connecticut.

Pang hsin Ting. 1978. Reconstruction of Proto-Puyuma Phonology. *Bulletin of the Institute of History and Philology, Academia Sinica*, 49:321–392. In Chinese.

Paul Jen kuei Li. 1991. *Orthographic Systems for Formosan Languages*. Ministry of Education, Taipei.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. 2018. A Lexical Distance Study of Arabic Dialects. *Procedia Computer Science*, 142:2–13.

Paul J. Li. 1997. *Formosa Folkways*. Chang-ming, Taipei, Taiwan. In Chinese.

Paul Jen-Kuei Li. 1999. *History of Formosan Natives: Linguistic Perspective*. Taiwan Literature Committee, Nantou, Taiwan. In Chinese.

Paul Jen-kuei Li. 2008. The Great Diversity of Formosan Languages. *Language and Linguistics*, 9(3):523–546.

Paul Jen-kuei Li. 2014. Semantic Shift and Variation in Formosan Languages. *Language and Linguistics*, 15(4):465–477.

Paul Jen-kuei Li and Shigeru Tsuchida. 2022. Subclassification of Amis Dialects. 臺灣語文研究, 17(1).

Kaiying Kevin Lin, Hsi-Yu Chen, and Haopeng Zhang. 2025. FormosanBench: Benchmarking Low-Resource Austronesian Languages in the Era of Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16527–16539, Suzhou, China. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Paul J. Meighan. 2021. Decolonizing the Digital Landscape: The Role of Technology in Indigenous Language Revitalization. *AlterNative: An International Journal of Indigenous Peoples*, 17(3):397–405.

Naoyoshi Ogawa and Erin Asai. 1935. *The Myths and Traditions of the Formosan Native Tribes*. Taihoku Imperial University, Taipei. In Japanese.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Taraka Rama, Çağrı Çöltekin, and Pavel Sofroniev. 2017. Computational Analysis of Gondi Dialects. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 26–35, Valencia, Spain. Association for Computational Linguistics.

Stanley Starosta. 1994. Proto-Rukai-Tsouic: Subgroup or Treetop? In *Proceedings of the Seventh International Conference on Austronesian Linguistics*, Leiden, the Netherlands.

Stanley Starosta. 1995. A Grammatical Subgrouping of Formosan Languages. In Jen kuei Li, Cheng hwa Tsang, Ying kuei Huang, Dah an Ho, and Chiu yu Tseng, editors, *Austronesian Studies Relating to Taiwan*, pages 683–726. Institute of History and Philology, Academia Sinica, Taipei.

Elizabeth Zeitoun. 2007. *A Grammar of Mantauran (Rukai)*. Institute of Linguistics, Academia Sinica, Taipei.

Elizabeth Zeitoun and Yoann Goudin. 2024. Language Contact in Formosan Languages. *Handbook of Formosan Languages: The Indigenous Languages of Taiwan*, 2:246–283.

Elizabeth Zeitoun, Ching-hua Yu, and Cui-xia Weng. 2003. The Formosan Language Archive: Development of a Multimedia Tool to Salvage the Languages and Oral Traditions of the Indigenous Tribes of Taiwan. *Oceanic Linguistics*, 42(1):218–232.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuo. 2024. Improving Low-Resource Machine Translation for Formosan Languages Using Bilingual Lexical Resources. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11248–11259, Bangkok, Thailand. Association for Computational Linguistics.

9. Language Resource References

Indigenous Languages Research and Development Foundation. 2020. 族語 *E* 樂園. Accessed: 2026-01-17.

Mohamed, W. and Le Ferrand, É. and Sung, L.-M. and Prud'hommeaux, E. and Hartshorne, J. K. 2024. *FormosanBank*. Electronic Resource; access date: 2026-01-17.

A. Dendrograms

We present dendrograms based on TF-IDF cosine similarity, Jaccard similarity, and Levenshtein distance in Figure 2.

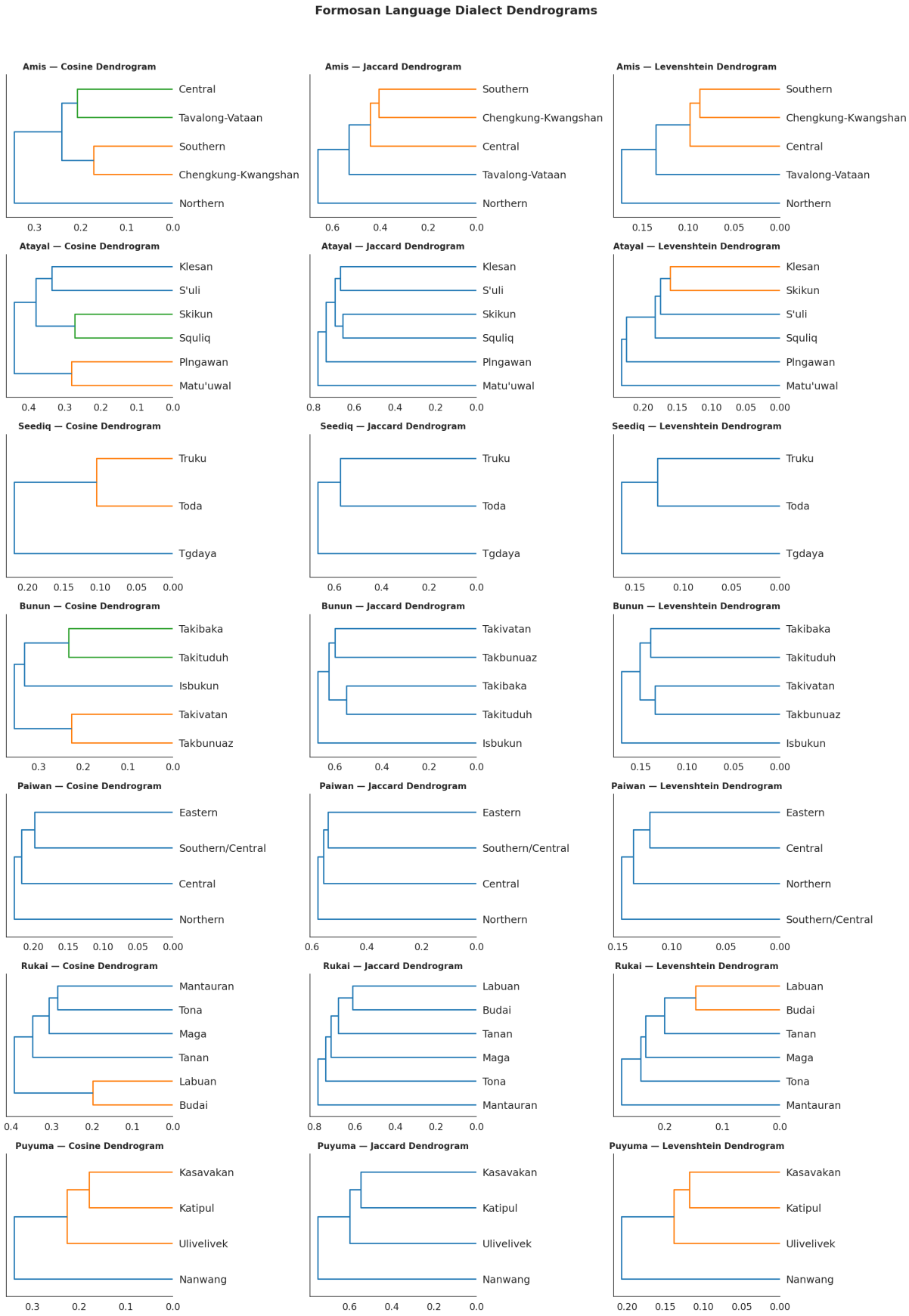

Figure 2: Per-dialect dendrograms based on TF-IDF cosine similarity, Jaccard similarity, and Levenshtein distance for all seven Formosan languages with multiple dialects