Linguistic Changes as Viewed From the 'Knowledge Database of Diaries in Taiwan' (1930s–1960s): A Quantitative Study

Ki Kwong Kenrich Chen
ISA Wuhan Wenhua School, Wuhan 430100, P.R. China
Stanford Pre-Collegiate Studies, Stanford University, Stanford, CA 94305, U.S.

Abstract

Introduction. Focusing on primary materials from the 'Knowledge Database of Diaries in Taiwan', this article aims to outline the linguistic changes in their diaries while exploring the potential roles of migration in linguistic changes.

Method. This study uses web-scraping techniques to collect large-size diaries data and develops a dictionary comparing the frequently used words in Mainland China and Taiwan, China. Then, this study uses computational linguistics to detect linguistic changes.

Result. The frequency of words used in Mainland China increased after the outbreak of Second Sino-Japanese War in 1937 and after China administered the Taiwan region in 1945, respectively. Meanwhile, the frequency of words used in the Taiwan region increased after the Guomindang Government moved to Taiwan in 1949.

Implications. This study sheds light on the role of geopolitical tension and/or migration on language changes and showcases the potential linkages between mobility and linguistic changes from an East Asian perspective.

Keywords: Social-Historical Linguistics, Taiwan, Chinese Language, Migration, Computational Linguistics

1 Introduction

The Second Sino-Japanese War (1937–1945) changed the cultural and social landscapes of East Asia, including the Taiwan region. After the end of the war in 1945, a large number of people from Mainland China migrated to Taiwan, especially after the Guomindang government's defeat in Mainland China in 1949. This migration also marks a turning point in people's language usages. At the time, the Taiwan region had already been a multilingual society. In such a multilingual society, there were Taiwanese Hokkien, Hakka, other Indigenous languages, as well as Japanese as the legacy of Japanese colonial rule. In addition, mainland migrants also brought Mandarin, regional dialects, as well as different speaking habits. Their arrival therefore resulted in a unique linguistic environment that connecting language choice and identity.

Using a quantitative method, this study focuses on how historical materials such as diaries, memoirs, letters, or public documents written between 1945 and the 1960s reflect the linguistic habits of Mainland migrants in Taiwan. This article's first aim is to outline the specific types of linguistic changes in the materials. Second, this article also tries to explore the potential roles of migration in linguistic changes.

A study along this direction holds its theoretical significance by echoing seminal research in linguistics and migration. Previous research as represented by the study of Pérez (2023) and Allibert (2008) has explored the linkage between migration and linguistics. While they have well covered the context of Europe and Malaysia, respectively, East Asia has not been a key focus. This research tries to extend seminal studies by focusing on in-depth case analysis in the East Asian context. A study along this direction will add to our knowledge on the dynamics between linguistics and migration.

Moreover, such an exploration also holds its significance in terms of linguistics practice. Clarifying the linguistic changes of migration population in Taiwan, China well also generate certain practical insights for understanding and governing the migration societies across the world. In and beyond Asia, the coherence of migration societies has been largely maintained by language and culture. Meanwhile, the integration of migration societies into the host country (or region) has also been viewed as a key issue. Exploring the linguistic changes will further our understanding of social and cultural en-

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 353–359
November 6-7, 2026 ©2026 Association for Computational Linguistics

counters in the changing geopolitical environment.

2 Literature Review

Seminal studies on linguistics issues in Taiwan, China have long used a sociolinguistic perspective to observe variational pragmatics between Mainland China and Taiwan, China. For example, Lin et al. (2012) have paid scholarly attention to the compliments in Mainland China and Taiwan, China, especially the influence of region and topic. Meanwhile, Lin (2015) has discussed the role of gender in Mainland Chinese and Taiwan Chinese compliments. Ren (2015) also takes into consideration the socio-pragmatic variation in Mainland Chinese and Taiwan Chinese refusals. Moreover, Liu and Liles (2025) outline the regional variation in conventional expressions, especially of Mandarin (or Guoyu) in Mainland China and Taiwan, China. Most recently, Zhang (2026) also explored the regional variation in tag questions between the two regions. Across the literature, comparing language use (compliments, refusals, tag questions, expressions) between Mainland China and Taiwan Chinese has always been a key focus.

Further research examines language policy, identity, and immigration in Taiwan in transnational contexts. These studies have linked language in Taiwan to political matters, national identity, and social belonging. For instance, Chang and Holt (2011) have analysed how "China" is named in Taiwan to show its connection to the politics, while Yang (2013) uses critical discourse analysis to detail Taiwan's debate on economic ties with Mainland China. Wei and Duann (2019) have also shown that pronouns are also related to collective identity, while Dupré (2016) has discussed attempts to legislate language in Taiwan. Similarly, Lin (2019) uses life story interviews to examine the language experience of Southeast Asian marriage migrants in Taiwan, China, while Le (2025) examines Vietnamese immigrants' language practice in Taiwan, China. In addition, Lin (2018) examines the gender positioning and social expectations of marriage-migrants in Taiwan, China.

Another topic of interest is the endangerment and revitalization of indigenous language(s) in Taiwan, China. Here, the focus is language shift, heritage language proficiency, and digital revitalization among Taiwan's indigenous peoples. The series of studies by Liu et al. (2015) and Liu et al. (2025) have covered the shift of these indigenous languages and the underlying sociocultural factors. In terms of indigenous language education, McNaught (2019) also shows us a landscape of digital revolution in Taiwan, China.

The fourth topic to be mentioned is about language systems and research outputs, especially on phonetics and corpus analysis. According to the study of Fon and Hsu (2007) as well as Hsu and Tse (2009), there is a realization of dipping tones in Taiwan Mandarin. As of now, the English spelling variation has been visible across Mainland China, Hong Kong, and Taiwan, China (Liu and Zhao, 2015). Within Taiwan, China, there is also a strong tension between foreignized and domesticated concerning involving Taiwan subjects (Wang et al., 2009). Overall, the focus here is about tonal phonetics and English spelling variation.

Despite the effort of seminal research, there is a visible research gap in the spectrum of literature. First, while Lin (2019) and Le (2025) do cover the experience of migrant populations from Southeast Asia to the Taiwan region, there is a lack of scholarly attention on the linkage between the linguistic changes and migrant populations from Chinese Mainland to the Taiwan region. Indeed, Mainland China and Taiwan, China, show a different academic and cultural system, which deserves further examination (Huang, 2026). Moreover, while Liu et al. (2015), Liu et al. (2025), as well as McNaught (2019) have discussed the language shift among indigenous language(s) in Taiwan, China, there is a lack of theoretical efforts to summarize the potential roles played by migration in linguistic changes.

In view of the glaring research gap mentioned above, this current study comes up with two research questions. First, what are the linguistic changes among migrant populations from the Chinese Mainland to the Taiwan region? (RQ1). Second, what are the linguistic changes among 'domestic' Taiwanese? (RQ2) Clarifying these two research questions would contribute to our established knowledge about migration and language in an East Asian context.

3 Method

3.1 Materials

This study draws on the diaries collected in the 'Knowledge Database of Diaries in Taiwan' (Taiwan riji zhishiku, https://taco.ith.sinica.edu.tw). Specifically, this study covers the diaries of seven intellectuals, namely Wu Songqing (1901–

	Mainland	Taiwan	Ratio
1947–1950	1525	467	3.266
1947	547	143	3.825
1948	303	78	3.885
1949	341	114	2.991
1950	334	132	2.530

Table 1: Word counts and Mainland/Taiwan ratios in the diaries of Wu Songqing.

1991), Guo Shuzi (1925–2004), Shao Yulin (1909–1984), Wu Xinrong (1907–1967), Ye Shengji (1923–1950), Yang Jizhen (1911–1990), and Lin Xiantang (1881–1956). This study uses web-scraping techniques to collect large-size data. The data were collected from Jun 15 to July 01, 2026.

3.2 Quantitative Method

This study first develops a dictionary comparing the frequently used words in Mainland China and Taiwan, China. This dictionary uses two official sources from Mainland China and Taiwan, China, respectively. The first source is 'Liang'an yongyu duizhao biao' [A Comparable List of Words Used in Mainland China and Taiwan, China] available on the website of the local Council Offices for Taiwan Affairs, People's Republic of China (https://gd.huaxia.com/c/2021/08/30/743121.shtml). The second source is 'Liang'an changyong ciyu duizhao biao' [A Comparable List of Words Frequently Used in Mainland China and Taiwan, China] available on a sub-website of the Ministry of Education, Republic of China (https://dict.concised.moe.edu.tw/appendix.jsp?ID=54&page=1). Then, this study will analyse the collected texts of diaries based on this dictionary. Following this, the Results part of this study will calculate the percentage change of word frequency. The dictionary is provided in an appendix file.

4 Results

This essay first analyzes the diaries of Wu Songqing, a Chinese official working for the Guomindang government and moved to the Taiwan region in 1949. As shown in his diaries, the frequency of words used in the Taiwan region increased after the Guomindang Government moved to Taiwan in 1949 (Table 1).

The second case is Shao Yulin, a Chinese official working for the Guomindang government and moved to the Taiwan region in 1949. His diaries only last for two years, and hence it is relatively difficult to figure out the long-term changes. However, it is observed that the ratio of Mainland words to Taiwan words in Shao's diaries during 1953–1954 (1.682) is lower than that in Wu's diaries during 1947–1950 (3.266). This can also reflect the decreasing trend of using Mainland words among Guomindang's officials after their moves to Taiwan (Table 2).

	Mainland	Taiwan	Ratio
1953–1954	249	148	1.682
1953	202	127	1.591
1954	47	21	2.238

Table 2: Word counts and Mainland/Taiwan ratios in the diaries of Shao Yulin.

	Mainland	Taiwan	Ratio
1944–1953	1573	278	5.658
1944	35	21	1.667
1945	N/A	N/A	N/A
1946	31	3	10.333
1947	160	12	13.333
1948	74	5	14.800
1949	72	36	2.000
1950	323	70	4.614
1951	532	98	5.429
1952	241	21	11.476
1953	105	12	8.750

Table 3: Word counts and Mainland/Taiwan ratios in the diaries of Guo Shuzi.

The word counts and Mainland/Taiwan ratios in mainland China and the Taiwan region in the diaries of aforementioned migrants are visualized in Figure 1 below.

5 Discussions

This study further verifies its results in domestic Taiwanese (i.e., non-migrants) for comparison. The first case here is Guo Shuzi, a Taiwanese working in a commercial bank in Taiwan. As showcased in her diaries, the frequency of words used in Mainland China increased after China administered Taiwan Province in 1945. Meanwhile, the frequency of words used in the Taiwan region increased after the Guomindang Government moved to Taiwan (while the Communist Government only administered Mainland China) in 1949 (Table 3).

The second case of 'local Taiwanese' is Wu Xinrong, a Taiwanese working in the politics as well as clinical medicine. As showcased in his diaries, the frequency of words used in Mainland China

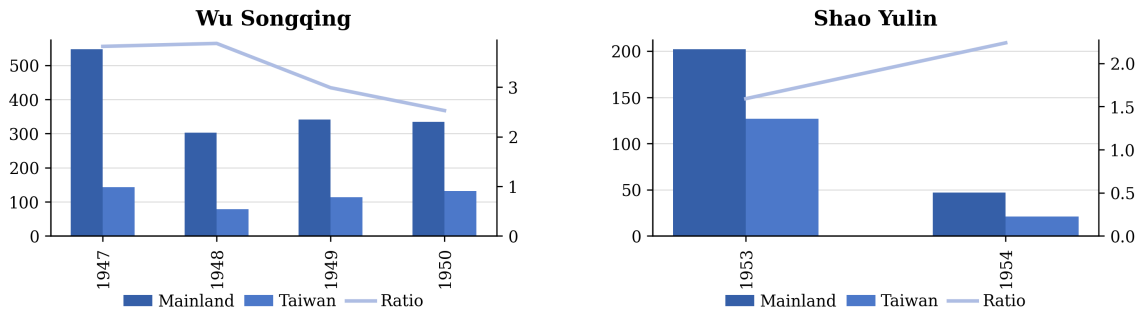

Figure 1: Word counts and Mainland/Taiwan ratios in the diaries of migrants.

Year	Mainland	Taiwan	Ratio
1933–1967	8844	3588	2.465
1933	51	36	1.417
1934	N/A	N/A	N/A
1935	105	35	3.000
1936	107	52	2.058
1937	99	55	1.800
1938	821	301	2.728
1939	713	262	2.721
1940	972	272	3.574
1941	753	260	2.896
1942	851	310	2.745
1943	654	171	3.825
1944	538	190	2.832
1945	470	288	1.632
1946	232	164	1.415
1947	201	148	1.358
1948	105	51	2.059
1949	102	43	2.372
1950	146	65	2.246
1951	143	70	2.043
1952	57	32	1.781
1953	23	10	2.300
1954	N/A	N/A	N/A
1955	95	59	1.610
1956	123	56	2.196
1957	82	35	2.343
1958	157	66	2.379
1959	105	48	2.188
1960	145	71	2.042
1961	154	55	2.800
1962	175	75	2.333
1963	171	70	2.443
1964	168	73	2.301
1965	146	86	1.698
1966	150	71	2.113
1967	30	8	3.750

Table 4: Word counts and Mainland/Taiwan ratios in the diaries of Wu Xinrong.

	Mainland	Taiwan	Ratio
1938–1950	5396	1266	4.262
1938	130	53	2.453
1939	228	95	2.400
1940	229	19	12.053
1941	594	153	3.882
1942	223	19	11.737
1943	471	95	4.958
1944	1154	230	5.017
1945	1105	245	4.510
1946	423	131	3.229
1947	152	46	3.304
1948	411	122	3.369
1949	180	38	4.737
1950	96	20	4.800

Table 5: Word counts and Mainland/Taiwan ratios in the diaries of Ye Shengji.

	Mainland	Taiwan	Ratio
1944–1950	1980	592	3.345
1944	192	36	5.333
1945	689	202	3.411
1946	55	18	3.056
1947	289	141	2.050
1948	373	114	3.272
1949	226	31	7.290
1950	156	50	3.120

Table 6: Word counts and Mainland/Taiwan ratios in the diaries of Yang Jizhen.

increased after the outbreak of the Second Sino-Japanese War in 1937, perhaps because after that 'local Taiwanese' started to turn attention to mainland China. Accordingly, the frequency of words used in Mainland China decreased around 1945. Meanwhile, the frequency of words used in the Taiwan region increased after the Guomindang Government moved to Taiwan (while the Communist Government only administered Mainland China) in 1949 (Table 4).

The third case of 'local Taiwanese' is Ye Shengji, a Taiwanese medical expert. As showcased in his diaries, the frequency of words used in Mainland China increased with the World War II, perhaps because after that 'local Taiwanese' started to turn attention to mainland China (Table 5).

The fourth case of 'local Taiwanese' is Yang Jizhen, a Taiwanese intellectual who worked for the Manchukuo for a short time. As showcased in his diaries, the frequency of words used in Mainland China decreased after the Guomindang Government moved to Taiwan (while the Communist Government only administered Mainland China) in 1949 (Table 6).

The fifth case of 'local Taiwanese' is Lin Xiantang (also known as guanyuan xiansheng), a Tai-

Year	Mainland	Taiwan	Ratio
1927–1955	9754	5561	1.754
1927	234	254	0.921
1928	N/A	N/A	N/A
1929	255	130	1.962
1930	394	205	1.922
1931	325	207	1.570
1932	483	210	2.300
1933	335	238	1.408
1934	413	258	1.601
1935	349	228	1.531
1936	N/A	N/A	N/A
1937	395	208	1.899
1938	272	274	0.993
1939	289	174	1.661
1940	185	96	1.927
1941	316	254	1.244
1942	244	204	1.196
1943	248	193	1.285
1944	345	238	1.450
1945	288	271	1.063
1946	576	247	2.332
1947	921	345	2.670
1948	714	225	3.173
1949	667	195	3.421
1950	235	130	1.808
1951	262	176	1.489
1952	200	99	2.020
1953	298	174	1.713
1954	255	208	1.226
1955	256	120	2.133

Table 7: Word counts and Mainland/Taiwan ratios in the diaries of Lin Xiantang.

wanese politician and activist. As showcased in his diaries, the frequency of words used in Mainland China increased after China re-administrated the Taiwan region in 1945, and decreased after the Guomindang Government moved to Taiwan (while the Communist Government only administered Mainland China) in 1949 (Table 7).

The word counts and Mainland/Taiwan ratios in mainland China and the Taiwan region in the diaries of aforementioned five non-migrants are visualized in Figure 2 below.

6 Conclusion

Overall, we can summarize three turning points. In 1937, the frequency of words used in Mainland China increased after the outbreak of the Second Sino-Japanese War, perhaps because after that 'local Taiwanese' started to turn attention to mainland China (Wu Xinrong and Ye Shengji). In 1945, the available Mainland/Taiwan word-count ratios show different patterns around the end of World War II: an increase from 1944 to 1946 for Guo Shuzi, but a decrease from 1944 to 1946 for Wu Xinrong. In 1949, the frequency of words used in the Taiwan region increased after the Guomindang Government moved to Taiwan in 1949 (Guo Shuzi and Wu Xinrong), while the Communist Government only administered Mainland China (Table 8).

	Migrants from Mainland China to the Taiwan Region	Non-Migrants / 'Local Taiwanese'
1937	Not available	Increased frequency of words used in Mainland China (Wu Xinrong and Ye Shengji)
1945	Not available	Mainland/Taiwan ratios increased from 1944 to 1946 for Guo Shuzi, decreased from 1944 to 1946 for Wu Xinrong, and increased from 1945 to 1946 for Lin Xiantang.
1949	Increased frequency of words used in the Taiwan Region (Wu Songqing)	Increased frequency of words used in the Taiwan Region (Guo Shuzi, Wu Xinrong, Yang Jizhen, and Lin Xiantang)

Table 8: Three turning points in the changing usages of languages in selected diaries.

However, this study also holds certain limitations. First, due to data availability, there are only two migrant diaries available in the 'Knowledge Database of Diaries in Taiwan'. Hence, most samples covered in our study are non-migrants. This limits our understanding on the dynamics between migration and language change. Second, due to material limitations, the limited number of available samples can only support descriptive observations and visualization, while further exploration on correlations and causal relationships remain unfeasible.

In the future, further studies can deepen our scope on the related topics by extending two lines of research directions. First, it will be helpful if more diaries of migrants from Mainland China to the Taiwan region can be available, with which we can further conduct comparative studies on the differential language changes between migrants and non-migrants in the Taiwan region. Second, future study can be based on larger sample sizes and connect the linguistic data to cultural and social factors. This will support in-depth explorations on the correlations and potential causal relationships behind language changes and external environments.

As of now, the findings from this study hold certain implications for both academic research and policy practice. From a theoretical perspective, this study sheds light on the role of geopolitical tension and/or migration on language changes. This can fulfil our knowledge gap caused by the lack of scholarly attention on the linkage between the linguistic changes and migrant populations from Chinese Mainland to the Taiwan region. From a practical perspective, the analysis results can remind policy makers to offer actionable solutions on protecting language diversity and improve language policies in future periods of geopolitical changes.

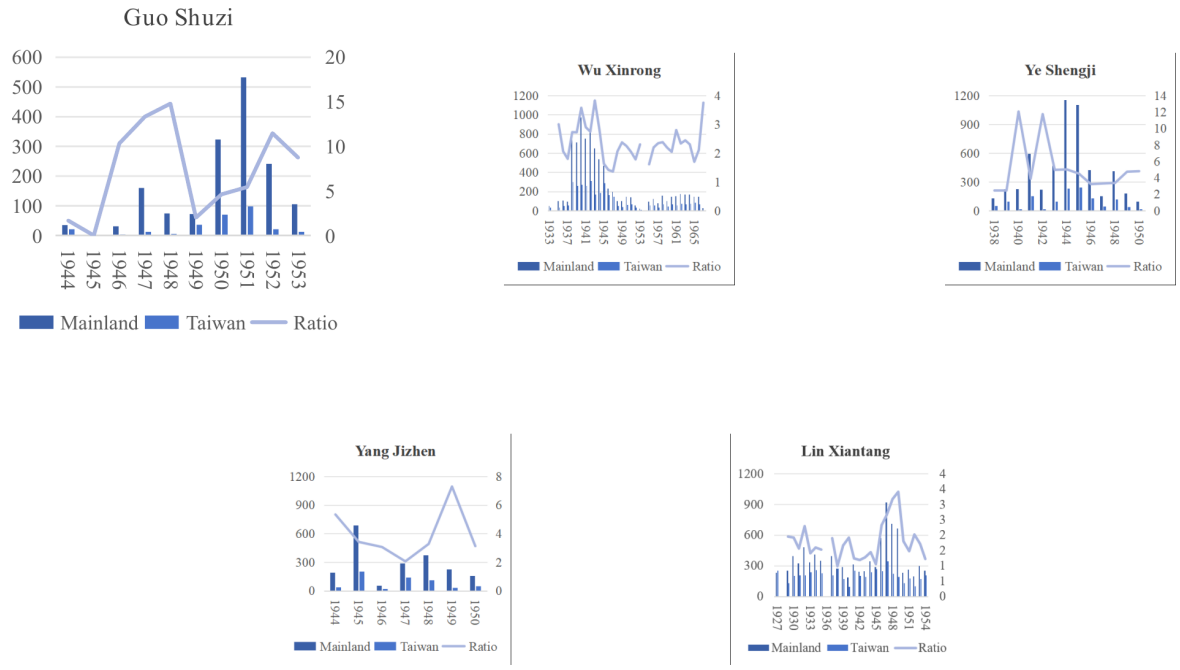

Figure 2: Word counts and Mainland/Taiwan ratios in the diaries of non-migrants.

References

Claude Allibert. 2008. Austronesian migration and the establishment of the malagasy civilisation: Contrasted readings in linguistics, archaeology, genetics and cultural anthropology. *Diogenes*, 55(2):7–16.

Hui-Ching Chang and Richard Holt. 2011. Naming china: Taiwan's national day speeches as identity politics. *Journal of Language and Politics*, 10(3):396–415.

Jean-François Dupré. 2016. Legislating language in taiwan: From equality to development to status quo. *Language Policy*, 15(4):415–432.

Janice Fon and Hui-Ju Hsu. 2007. Positional and phonotactic effects on the realisation of dipping tones in taiwan mandarin. In Carlos Gussenhoven and Tomas Riad, editors, *Tones and Tunes, Volume 2: Experimental Studies in Word and Sentence Prosody*, pages 239–269. Mouton de Gruyter.

Hui-Ju Hsu and John Kwock-Ping Tse. 2009. The tonal leveling of taiwan mandarin: A study in taipei. *Concentric: Studies in Linguistics*, 35(2):225–244.

Jun Huang. 2026. Pivoting academic communities, piloting 'diaspora options': Diaspora academicians of chinese mainland and taiwan region. *Higher Education*. Online first.

Thi Thuy Trang Le. 2025. Navigating multilingual worlds: Vietnamese immigrant children's translanguaging practices and family resilience in taiwan. *Journal of Multilingual and Multicultural Development*. Advance online publication.

Chih-Ying Lin. 2015. The role of gender in taiwan and mainland chinese compliments. In Kate Beeching and Helen Woodfield, editors, *Researching Sociopragmatic Variability: Perspectives from Variational, Interlanguage and Contrastive Pragmatics*, pages 49–71. Palgrave Macmillan.

Chih-Ying Lin, Helen Woodfield, and Wei Ren. 2012. Compliments in taiwan and mainland chinese: The influence of region and compliment topic. *Journal of Pragmatics*, 44(11):1486–1502.

Mei-Chen A. Lin. 2019. Construction of immigrant mothers' language experiences in taiwan: Mothering in one's second language is a 'choice'? *Language Policy*, 18(1):87–105.

Shih Lin. 2018. 'then you can ride the scooter to run away!' gender positioning of marriage-migrants in adult mandarin education in taiwan. *International Journal of Bilingual Education and Bilingualism*, 21(6):617–631.

Ding-Ting H. Liu, Yu-Hsuan Chang, and Paul Jen-Kuei Li. 2025. Unraveling the impact of sociocultural factors on indigenous heritage language proficiency in taiwan. *Language Problems & Language Planning*, 49(1):48–78.

Ding-Ting H. Liu, Yu-Hsuan Chang, Paul Jen-Kuei Li, and Jen-Pin Lin. 2015. Language shift of taiwan's indigenous peoples: A case study of kanakanavu and saaroa. *Journal of Multilingual and Multicultural Development*, 36(7):729–749.

Fang Liu and Xin Liles. 2025. Regional variation in conventional expressions: A comparative study of

mandarin in mainland china and taiwan. *Lingua*, 326:104035.

Yu Liu and Yan Zhao. 2015. English spelling variation and change in newspapers in mainland china, hong kong and taiwan. *English Today*, 31(4):5–14.

Don McNaught. 2019. From mountains to megabytes: The digital revolution of indigenous language education in taiwan. In Ari Sherris and Susan D. Penfield, editors, *Rejecting the Marginalized Status of Minority Languages: Educational Projects Pushing Back Against Language Endangerment*, pages 70–89. Multilingual Matters.

María Cristina Pérez. 2023. The representation of migration in parliamentary settings: Critical cross-linguistics corpus-assisted discourse analyses. *Humanities & Social Sciences Communications*, 10(1):965.

Wei Ren. 2015. Sociopragmatic variation in mainland and taiwan chinese refusals. In Kate Beeching and Helen Woodfield, editors, *Researching Sociopragmatic Variability: Perspectives from Variational, Interlanguage and Contrastive Pragmatics*, pages 72–93. Palgrave Macmillan.

Qian Wang, Hsiao-Wei Chen, and Ying Zhong. 2009. Foreign is not unfamiliar: A translation impact study involving taiwan subjects. *Meta*, 54(2):342–356.

Jennifer M. Wei and Ren-feng Duann. 2019. Who are we? contesting meanings in the speeches of national leaders in taiwan during the authoritarian period. *Journal of Language and Politics*, 18(5):760–781.

Pei-Lin Yang. 2013. A critical discourse analysis of taiwan's national debate on economic ties with china. *Taiwan Journal of Linguistics*, 11(2):83–103.

Jian Zhang. 2026. Regional pragmatic variation in tag questions: A corpus-based comparison of mainland chinese and taiwan chinese. *Corpus Pragmatics*, 10(1):38.