# Domestic Linguistic Distance in China

Seoyeon Lee

Abstract

This research delves into China's linguistic landscape, exploring the intricate relationship between language policies and diversity. Despite being home to a vast population and numerous ethnic groups, China exhibits comparatively low linguistic diversity. Through a comprehensive analysis, this study examines the impact of language policies, focusing on constitutional guarantees, language laws, and educational initiatives. The research utilizes metrics developed by scholars such as Fearon and Greenberg to measure linguistic diversity and assesses the effects of policies on dialects, Putonghua, and minority languages. The findings highlight complex outcomes, including bidialectalism, the vernacularization of Putonghua, and challenges faced by minority languages. This exploration contributes to a deeper understanding of China's linguistic dynamics and their implications for cultural diversity.
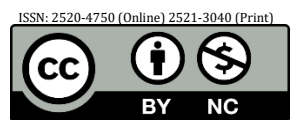
Keywords: *Linguistic Diversity, Language Policies, Putonghua, Minority Languages, Bidialectalism.*

About Author (s)

Seoyeon Lee, Ph.D. Candidate in the Department of Philosophy and Religious Studies, Peking University, Beijing, China.

1. Introduction

China is one of the most populous countries, with a population of approximately 1.4 billion people as of 2023. The country is characterized by a diverse range of ethnic groups, officially recognized as 56 minority ethnicities. Notably, the Han ethnic group constitutes a significant portion of China's population. China is a multicultural nation with a multitude of languages spoken. The official language is Mandarin Chinese (普通话, Putonghua), widely used by the majority of the Chinese population. However, various regional languages are also spoken within China, reflecting the cultural characteristics of different regions and minority ethnic groups. Examples of these regional languages include Cantonese, Hokkien, Uyghur, Mongolian, and others. The linguistic diversity in China reflects both regional distinctions and the cultural uniqueness of minority ethnic groups. Despite the dominance of Mandarin Chinese, these regional languages play a vital role in communication and cultural expression within their respective communities.

Scholars such as Fearon (2003) and Greenberg (1956) have played key roles in empirically measuring ethnic and linguistic diversity. Linguistic distance is often calculated using computational methods that analyze linguistic data and compare the structural and lexical aspects of languages. For example, two languages that share a common origin and have undergone minimal changes over time would have a short linguistic distance. On the other hand, languages that have evolved separately and have significant differences in terms of vocabulary, grammar, and phonetics would have a greater linguistic distance. Both Fearon and Greenberg empirically measured the linguistic diversity and compiled a list of a linguistic classification system designed to categorize languages based on their genealogical relationships. Their indices represent the likelihood that two randomly selected individuals from a population will have distinct mother tongues. As such, it varies between 0 (indicating everyone shares the same mother tongue) and 1 (suggesting no two individuals have the same mother tongue). According to Fearon's Index, China is ranked 150th among 215 countries, while Canada is 65th. It means that China is much less diverse than Canada in terms of its linguistic diversity. In addition, Greenberg's Diversity Index also indicates that China is less linguistically diverse than Canada ranking 89th among 232 countries, while Canada is 74th.

This research aims to investigate the reasons behind China's comparatively low linguistic diversity, despite having the world's largest population, while acknowledging the variation in territories occupied by diverse ethnic groups. Two potential factors are explored. Firstly, countries with low linguistic diversity may adopt and strongly standardize a particular language as the national language. This can lead to the marginalization or lack of preservation of regional or minority languages. Secondly, language policies can significantly impact diversity. Forced suppression of certain languages or a lack of education in minority languages can contribute to a decrease in linguistic diversity. By examining these factors, the research seeks to unfold the complexities surrounding China's linguistic landscape and its relationship with ethnic diversity.

2. Literature Review

The research explores the intricate relationship between China's linguistic diversity and its cultural and ethnic landscape, with a focus on key studies that have contributed to this understanding. Fearon (2003) stands out as one of the pivotal scholars who have empirically measured linguistic diversity within diverse cultural and ethnic contexts. His research method involves calculating linguistic distance through computational analysis, examining the structural and lexical facets of languages. Fearon's work provides valuable insights into the dynamics of linguistic diversity, revealing how languages evolve and relate to cultural

distinctions. By categorizing languages based on genealogical relationships, Fearon's contributions offer a quantitative framework for assessing linguistic variation. Fearon (2003) has been instrumental in empirically measuring linguistic diversity and understanding its implications within diverse cultural and ethnic contexts. Fearon's research method involves calculating linguistic distance through computational analysis, examining the structural and lexical facets of languages. His work provides valuable insights into the dynamics of linguistic diversity and reveals how languages evolve and relate to cultural distinctions. By categorizing languages based on genealogical relationships, Fearon's contributions offer a quantitative framework for assessing linguistic variation.

Fearon proposed utilizing the linguistic distance between the "tree branches" of two languages as a measure, though an imperfect one, of the cultural distance between groups whose primary language is those languages. He regarded this linguistic distance as a proxy for cultural fractionalization. He performed a quantitative analysis to determine the linguistic distances across languages within countries, providing insights into the relationships among these languages. Fearon's approach offered a concrete means of quantifying both language diversity and relationships. In addition to his linguistic research, Fearon expanded his investigation to include various religious and ethnic groups, presenting a comprehensive portrayal of the multifaceted nature of diversity. The relationships between linguistic, cultural, and ethnic identities were better understood as a result.

Fearon's data for the linguistic distance indices covers a wide variety of countries, and the indices for each nation are created using weighted averages of the linguistic distances. For each pair of living languages included in the 26th edition of *Ethnologue*, the linguistic distance between them is calculated as the first step. *Ethnologue* is a recurring scholarly publication that is accessible in both physical and digital formats (https://www.ethnologue.com/). The resource provides statistical statistics and relevant information pertaining to the 7,168 living languages worldwide. Regarded as the most thorough compilation of its kind, it functions as an exhaustive inventory of languages. Its first edition was published in 1951, and SIL International, an American non-profit with a connection to evangelical Christianity, is presently in charge of managing it. *Ethnologue* compiles a comprehensive inventory of living languages, organizing them into categories based on shared linguistic characteristics. When two languages are selected randomly and determined to belong to the same linguistic family, it is generally believed that they share similarities and so demonstrate a reduced linguistic distance.

Expanding on Fearon's groundwork, Desmet *et al.* (2009) further enrich our understanding through the provision of five distinct indices: the Greenberg Index (GI), Ethnolinguistic Fractionalization (ELF), Esteban and Ray Index (ER), Reynal-Querol Index (RG), and Peripheral Heterogeneity (PH). Within this array of indices, the Greenberg Index (GI), initially introduced by Greenberg (1956), assumes a central role. This linguistic classification system is meticulously designed to categorize languages based on their genealogical relationships, thereby contributing to the systematic organization and classification of the vast array of languages spoken globally. The Greenberg Index (GI) operates with a hierarchical structure, wherein languages are grouped into families, families into stocks, and stocks into phyla. This systematic arrangement serves the primary purpose of reflecting the evolutionary history of languages and their connections through descent. While the Greenberg Index (GI) has been widely influential, sparking significant debates within the linguistic community, it has undeniably played a crucial role in shaping subsequent research and discussions on language families and their origins.

Moving to the demographic landscape, approximately 91.6% of China's population identifies as Han (Chinese), with the remaining less than 10% distributed among 55 minority groups, varying in size from larger communities like the Zhuang (around 16 million individuals) to much smaller groups such as the Lhoba (approximately 3000 people) (Mofcom, 2009; National Bureau of Statistics of China, 2011). Linguistic diversity within these 56 nationalities spans a broad spectrum, encompassing languages ranging from 135 to 281 (Kurpaska, 2019). Despite the dominance of the Han ethnic group, linguistic uniformity is not observed. The Han Chinese population speaks a multitude of dialects, distinct to such an extent that many linguists categorize them as a group of related but not mutually unintelligible languages, known as Sinitic (Mair, 1991; Chappell & Li, 2016). These dialects, or languages, are often classified into 7-10 groups, depending on the criteria employed.

3. Methodology and Results

Linguistic diversity is assessed through the methodologies developed by Fearon and Greenberg, who empirically measured linguistic diversity and constructed a linguistic classification system designed to categorize languages based on their genealogical relationships. The indices formulated by Fearon and Greenberg serve as key metrics, representing the likelihood that two randomly selected individuals from a population will have distinct mother tongues. The scale of these indices ranges between 0 (indicating a scenario where everyone shares the same mother tongue) and 1 (suggesting that no two individuals have the same mother tongue). To explore the intricate complexities of linguistic diversity, it becomes imperative to first measure linguistic similarity. This initial step involves utilizing linguistic dendrograms to quantify the abstract concept of linguistic similarity. The language genealogical or genetic classification, as provided in *Ethnologue* (Nakagawa and Sugasawa, 2021), forms the basis for constructing these dendrograms. The quantity $e(i, j)$, representing the count of common edges across languages i and j on a dendrogram, offers insights into the linguistic proximity between them. A large value of $e(i, j)$ indicates that languages i and j may be classified inside a meta-group, signifying significant linguistic proximity. In assessing linguistic similarity, the research aligns with the approach outlined by Fearon (2003) and Desmet *et al.* (2009). This involves employing a specific formula to reveal linguistic similarity, thereby laying the groundwork for a comprehensive analysis of the intricate relationship between linguistic diversity and similarity in the research:

$$\text{similarity}(i,j) = \frac{e(i,j)}{\text{gmax}},$$

where g_{max} is the total number of $g(i)$ for all living languages in the world, $g(i)$ is the generation to which the language i belongs, and $e(i, j)$ is transformed into the proportions of cognates between i and j (normalization to the interval [0, 1]). In other words, g_{max} is the most edges that two languages may share. Languages that exhibit a greater degree of similarity are characterized by a reduced linguistic distance. Consequently, an increase in the measure of similarity between two languages, denoted as similarity(i, j), results in a reduction in the linguistic distance between them, represented as $\tau(i, j)$. Moreover, it is postulated that $\tau(i, j) = \tau(j, i)$ is true for all languages i and j. The measure $\tau(i, j)$ is a standardized measure, meaning that all values of τ_{ij} fall between the range of 0 and 1. Additionally, it is important to note that $\tau(i, i) = 0$ holds true for every i. Using this approach, Fearon (2003) defines the distance between languages i and j to be:

$$\tau(i,j) = 1 - [\text{similarity}(i,j)]^{\delta} = 1 - \left[\frac{e(i,j)}{\text{gmax}}\right]^{\delta} \ for \ all \ i,j \ (i \neq j),$$

where $\delta \in (i,j)$ determines the rate at which the linguistic distance decreases with an increasing number of shared edges.

When language $c(i)$ refers to the major language spoken in country i, the linguistic distance between languages j and $c(i)$ is denoted as $\tau_{j,c(i)}$. Therefore, the measure of domestic linguistic distance, DLD(i), is defined as the population-weighted average of linguistic distances to the major language spoken in country i.

$$\text{DLD}(i) = \sum_{j=1}^{K(i)} sj(i)\, \tau_{j,c(i)}$$

Fearon's Index, a fundamental metric in assessing linguistic diversity, positions China at the 150th rank among 215 countries, contrasting with Canada's 65th position. This ranking differential underscores that, according to Fearon's Index, China exhibits significantly lower linguistic diversity compared to Canada. Despite China's population of 1.4 billion, surpassing Canada's 0.39 billion, and having an additional 200 live indigenous languages, China scores 0.1327, while Canada scores 0.7124. Furthermore, Greenberg's Diversity Index, another crucial metric employed in this research, positions China at the 89th rank among 232 countries, whereas Canada is ranked 74th. According to Greenberg's Index, China scores 0.521, while Canada scores 0.603. Although the results are not remarkable, they yet support the findings of Fearon's Index.

4. Discussion
4.1 China's National Constitution and the Language Law
This study explores why China, despite its vast population, exhibits comparatively low linguistic diversity. One factor examined is the tendency of countries with low linguistic diversity to adopt and standardize a national language, potentially marginalizing regional or minority languages. Article 19 of China's National Constitution guarantees the promotion of the standardization of the lingua franca, Putonghua (普通话). This commitment is further emphasized by the Language Law, effective since 2001. Article 4 of the Language Law asserts the right of all citizens to learn and use the standard spoken and written Chinese language (GOV.cn, 2000). Simultaneously, both the Constitution (Article 19; National People's Congress 2004) and the Language Law (Article 8) ensure that all ethnic groups have the freedom to use and develop their own spoken and written languages (GOV.cn, 2000).

Zhou (2001) outlines the primary objectives of language planning in China since the 1950s. First and foremost is the standardization and popularization of the lingua franca of China, denoted as Putonghua (lit. "common speech"), commonly known as Mandarin in the West, and declared as the standard in 1955. Initially intended to replace local varieties, the idea of replacing them was later abandoned. The second aim is to transition from the traditional classic style to writing in a vernacular style. Classical Chinese (文言文) served as the standard literary language until 1919 when it was replaced by the vernacular language (白话文). Since the 1950s, government documents and newspapers have adopted vernacular writing, although Classical Chinese is still visible in formal language and poetry. The third goal is the design and promotion of a system of Chinese phonetic symbols, specifically the *Hanyu Pinyin* system, the official transcription of Chinese characters into the Latin alphabet since 1958. Initially proposed to replace Chinese characters entirely, it was eventually decided to serve as an auxiliary system. The fourth objective is the simplification of Chinese characters, intending to make learning them easier and reduce illiteracy. The fifth goal involves designing and, if necessary, improving writing systems for minor nationalities. The aim is to establish ethnic and linguistic equality, enabling minorities to better learn Putonghua in the process. The Language Law explicitly

designates Putonghua as the official language used by State organs, as well as the "basic language in education and teaching in schools and other institutions of education," as outlined in Articles 9 and 10 (GOV.cn, 2000). This positions the standard language with the highest status. While regional tongues are permitted in unofficial situations and to some extent in local media, traditional folk arts, or publications, these usages come with certain restrictions, as specified in Article 16 of the Language Law (GOV.cn, 2000).

4.2 Language Policies

The discernible disparity in linguistic diversity between China and Canada can be attributed to China's language policies. China's language policies, exemplified by initiatives like Putonghua Promotion Week (推广普通话宣传周), held annually since 1998, are multifaceted and extensive. This endeavor, occurring during the third week of September, spans large cities, towns, villages, and even ethnic minority regions, incorporating diverse activities to promote Putonghua (Wang and Yuan, 2013; Liang, 2015). Additionally, the implementation of the Putonghua Proficiency Test (普通话水平测试) ensures the fluency of native speakers of Chinese dialects, particularly relevant for professions such as teaching and media presentation. The test's proficiency level is a determinant of occupational qualifications. The concerted efforts extend to rural areas and ethnic minority regions and encompass special training for migrant workers and teachers in these regions (Zhou *et al.*, 2013). Crucially, Article 10 of the Language Law mandates the use of Putonghua as the language of instruction in schools across dialectal and ethnic minority areas, as well as in media (Article 12; GOV.cn, 2000). This widespread application ensures the pervasive influence of the national standard language throughout the country. The *Hanyu Pinyin* Romanization system is a part of the script reform in China, which focuses on standardizing and simplifying Chinese characters. This initiative, designed to reduce illiteracy, initially considered replacing characters with an alphabetic script but ultimately chose *Pinyin* as an auxiliary system. The government also embarked on designing writing systems based on *Pinyin* for minority groups without scripts and improving those with existing systems. Despite these efforts, Chinese characters remain the primary writing system in education and minority areas, coexisting with around thirty writing systems used by various ethnic minorities by 2004 (Spolsky, 2014; Zhou, 2001).

4.3 The Effects of Language Policies in China

The active promotion of the standard language, Putonghua, in China has led to a decline in local dialects, resulting in dialect endangerment. The teaching of Putonghua in schools, combined with rural-to-urban migration, has caused communication issues within families, particularly between younger generations and grandparents. Even in seemingly strong local tongue areas like Canton or Shanghai, dialect endangerment is apparent (Spolsky, 2014). The intensive teaching of Putonghua in private spaces is contributing to its dominance at home, leading to the inevitability of the standard language's prevalence in the future. China's language policy establishes Putonghua as the high-variety language, relegating dialects to low-variety situations (Kurpaska, 2013). Triglossia emerges, notably in rural areas where Putonghua replaces the former dialectal standard. Residents navigate their local dialect, the regional standard (e.g., Cantonese), and Putonghua (Li, 2015), with the latter assuming the role of the superior variety. Bidialectalism prevails due to China's linguistic diversity, with speakers proficient in both their mother tongue and Putonghua. Authorities seek a complementary relationship, with the standard in a more prominent position. To support Putonghua promotion, increased research aids teaching and preserves local dialects. However, the diffusion of Putonghua leads to diversification, termed "local Putonghua" or "non-standard Putonghua." Pronunciation reflects this, with speakers incorporating mother tongue characteristics. Syntactic and lexical features from local dialects contribute to the standard's

absorption of dialectal elements. These gradually influence standard pronunciation in each new edition of references, like the Contemporary Chinese Dictionary, increasingly including entries annotated as of dialectal origin. Despite constitutional guarantees for minority languages, a lack of national law protects minority languages, leaving their rights unobserved. The forceful promotion of Putonghua, driven by the need for intercommunication, endangers minority languages. The language attitudes of minorities contribute to language decline, with Putonghua proficiency enhancing job prospects and quality of life. The spread of Putonghua causes a conflict of loyalties as minorities aim to preserve their culture while aspiring to career opportunities, resulting in a lowered status for minority languages. Inequality persists in treating ethnic minorities, and constitutional language rights face challenges. Pressure to know Putonghua often leads to discrimination. Bilingual education efforts, while imperfect, are evident in ethnic minority regions, emphasizing the need to master both the mother tongue and the national standard (Zhou *et al.*, 2013). Disparity extends to areas beyond education. Minority languages, even those with official status, lack proper use in official documents and government conferences, often presented only in Chinese. The language policy's impact has prompted interest and investigation into minority languages, with efforts to disseminate Putonghua. The introduction and improvement of writing systems for ethnic minorities contribute to self-awareness and cultural heritage preservation.

Conclusions

In conclusion, this research explores the multifaceted dynamics shaping linguistic diversity in China, a country with a rich tapestry of ethnicities and languages. Despite its vast population and ethnic variety, China exhibits comparatively low linguistic diversity, as indicated by indices such as Fearon's and Greenberg's. This paper explores the intricate interplay of constitutional provisions, language laws, and active language policies that have contributed to this unique linguistic landscape. The constitutional emphasis on the standardization of Putonghua, China's official language, and the subsequent Language Law have played a pivotal role in promoting a unified linguistic identity. The standardization efforts, dating back to the mid-20th century, sought to establish Putonghua as the lingua franca while simultaneously preserving regional languages. However, the unintended consequence has been a decline in linguistic diversity, with local dialects facing endangerment. Language policies, exemplified by initiatives like Putonghua Promotion Week and the Putonghua Proficiency Test, underscore the government's commitment to linguistic homogeneity. The mandatory use of Putonghua in schools and media further consolidates its dominance, impacting communication within families and contributing to dialectal standardization. The effects of these policies exhibit subtleties, involving bidialectalism, the vernacularization of Putonghua, and the challenges faced by minority languages. Despite constitutional guarantees for minority languages, the absence of specific national laws poses a threat to their existence, with the forceful promotion of Putonghua leading to a conflict of loyalties among ethnic minorities. In essence, while China's language policies have successfully promoted a unified national language, they have also contributed to a decline in linguistic diversity, especially at the regional and minority levels. The delicate balance between fostering a common linguistic identity and preserving the richness of regional languages remains a challenge. As China continues to evolve, future considerations must weigh the impact of language policies on the diverse linguistic tapestry that has been an integral part of the nation's cultural heritage. The complexities surrounding language, identity, and diversity warrant ongoing exploration and thoughtful policymaking to ensure a harmonious coexistence of linguistic traditions within the broader Chinese cultural mosaic.

References

Chappell, Hilary & Lan, L. (2016). Mandarin and Other Sinitic Languages. In: *The Routledge Encyclopedia of the Chinese Language*. Sin-Wai Chan (ed.). London and New York: Routledge. 605–628.

Desmet, Klaus & Weber, Shlomo & Ortuño Ortin, Ignacio. (2009). Linguistic diversity and redistribution. *Journal of The European Economic Association. 7*. 10.1162/JEEA.2009.7.6.1291.

Fearon, James. (2003). Ethnic and cultural diversity by country*. *Journal of Economic Growth. 8*.

Greenberg, J. H. (1956). The measurement of linguistic diversity. *Language*. 32. 109-115.

GOV.cn. (2000). Law of the People's Republic of China on the Standard Spoken and Written Chinese Language (Order of the President No.37). english1.english.gov.cn/laws/2005-09/19/content_64906 html (accessed 11.01.2018).

Kurpaska, Maria. (2019). The effects of language policy in China A. *Język Komunikacja Informacja*. 14-24. 10.14746/jki.2017.12.1.

Liang, Sihua. (2015). Language Attitudes and Identities in Multilingual China: A Linguistic Ethnography. 10.1007/978-3-319-12619-7.

Mair, V. (1991). What is a Chinese "dialect/topolect"? Reflections on some key Sino-English linguistic terms. *Sino-Platonic Papers*. 29. 1-31.

Mofcom [Ministry of Commerce of the People's Republic of China]. (2009). Population and Ethnic Groups of the People's Republic of China. http://english.mofcom.gov.cn/statistic/charts.shtml.

National Bureau of Statistics of China. (2011). Communiqué of the National Bureau of Statistics of People's Republic of China on major figures of the 2010 Population Census [1] (No. 1). www.stats.gov.cn/english/PressRelease/201902/t20190228_1651335.html.

Spolsky, Bernard. (2014). Language and Public Policy. Language management in the People's Republic of China. *Language* 90 (4), 165-179.

Wáng, Huī & Yuan, Zhongrui. (2013). The promotion of Putonghua (Mandarin Chinese): An overview. 10.1515/9781614512530.27.

Zhou Qingsheng, Wei Dan and Xie Junying. (2013). Language policies and regulations in China: An overview. In: Li Yuming and Li Wei (eds): *The Language Situation in China*, *Volume 1*. De Gruyter Mouton [and] The Commercial Press. 11-25.

Zhou, Youguang. (2001). Language planning of China: Accomplishments and failures. *Journal of Asian Pacific Communication*. 11. 10.1075/japc.11.1.03you.

Cite this article:

Published by