# AN EXTENDED REPLICATION THEORY APPROACH TO TAIWAN MANDARIN SYNTAX

by

Lauren Elizabeth Clark

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Linguistics

at

The University of Wisconsin-Milwaukee

August 2024

ABSTRACT

AN EXTENDED REPLICATION THEORY APPROACH TO TAIWAN MANDARIN SYNTAX

by

Lauren Elizabeth Clark

The University of Wisconsin-Milwaukee, 2024
Under the Supervision of Professor Garry Davis

Previous accounts of Taiwan Mandarin syntax have generally taken a sociolinguistic approach, ascribing the presence of non-Mandarin-like features and patterns to interference from Taiwanese Southern Min or other southern varieties of Chinese. However, this approach has two major flaws. Firstly, it fails to explain the existence of patterns in Taiwan Mandarin that are unique to that dialect and the absence in Taiwan Mandarin of patterns that exist in both Standard Mandarin and Taiwanese Southern Min. Secondly, the focus on presence-absence as a binary has obscured important information about the frequency and distribution of various grammatical constructions. In order to address these shortcomings, I propose to reanalyze Taiwan Mandarin syntax from the perspective of a version of Replication Theory (Heine & Kuteva 2005) that has been modified to incorporate ideas from the field of World Englishes.

Replication Theory is built around the premise that, in language contact situations, speakers notice when grammatical structures in the two languages are conceptually similar and create new usage patterns on the basis of those similarities. What World Englishes contributes to this framework is, firstly, the founder principle (Mufwene 1996), that is, the idea that the structural features of a contact language are dependent on those of the dialects spoken by the people actually involved in the contact situation. Since it is highly unlikely that the Mainlanders

who fled to Taiwan in 1949 spoke Standard Mandarin, the founder principle requires that that dialect not be used as a point of reference in discussions about Taiwan Mandarin. Secondly, the language ecology concept of the feature pool (Mufwene 2001) posits that learners create their own idiolects by selecting features present in the speech of those they interact with. At a community level, this means that the features of all language varieties involved in a contact situation are available for selection, though certain features may be more or less likely candidates depending on how much of the population uses them and whether they are reinforced by- or competing against features from other language varieties. Thus, northern Taiwan, where most of the Mainlanders settled, should have had a stronger Mandarin presence in its feature pool that central or southern Taiwan.

Combining these ideas with Replication Theory, I predict that Taiwan Mandarin will not have Standard Mandarin features that are unusual or absent in other varieties of Mandarin, that diachronic analysis will show that features attributed to interference from Taiwanese Southern Min are in fact replica grammaticalizations and not polysemy copies, and that these features will be more grammaticalized in the speech of those from southern Taiwan than those from the north. Using data from both the existing literature and corpora of spoken Taiwan Mandarin collected in the 1980s, early 2000s, and early 2020s, I argue that this extended Replication Theory better explains the syntax of Taiwan Mandarin than previous approaches.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| 1 | first person |
| 2 | second person |
| 3 | third person |
| AUX | auxiliary |
| CL | classifier |
| COM | comitative |
| COMP | complementizer |
| CONT | continuation |
| COP | copula |
| EXCL | exclusive |
| EXP | experiential |
| INCL | inclusive |
| NEG | negative |
| PFV | perfective |
| PL | plural |
| PROG | progressive |
| Q | question marker |
| RES | resultative |
| SFP | sentence-final particle |
| SG | singular |

# ACKNOWLEDGEMENTS

I wish to thank Prof. Garry Davis for his support during this project. From brainstorming ideas for the prelim to sending me sources to giving feedback on early drafts, you have been there every step of the way to offer advice and encouragement. Thank you also to Zoé, Therese, and Dr. Tang Boyland for being willing to answer my many questions about the specific interpretation of this or that word or phrase in Mandarin. This work is all the stronger for your contributions.

I also wish to thank my parents for letting me use your house as a writing retreat and for knowing when not to ask how the dissertation is going. Above all, I am grateful for my husband, Adam; your endless faith in me and unwavering support mean more than you know.

# Chapter 1: Introduction

## 1.1 Chinese Languages and Terminology

Because of political ideology and a shared orthography, "Chinese" is often referred to as if it were a single language with dialects existing in different parts of China. However, these "dialects" are often mutually unintelligible, making it more appropriate, at least in the field of linguistics, to term them languages. In fact, linguists recognize up to ten different sub-families within the broader Sinitic (Chinese) language family: Mandarin, Xiang, Gan, Wu, Min, Hakka, Yue, Jin, Hui, and Pinghua (Chappell et al. 2007). These sub-families can be grouped areally into Northern varieties (Mandarin and Jin), which exhibit more typologically OV features, and Southern varieties (Xiang, Gan, Wu, Min, Hakka, Yue, Hui, and Pinghua), which adhere more strictly to Sinitic's overall VO profile. This is a particularly relevant distinction in the case of Taiwan Mandarin, as Mandarin and Min, the principle language families involved in the formation of this contact variety, belong to different areal groups.

Also relevant when dealing with different Chinese languages is the plethora of names that usually exist for a single variety. For the sake of clarity, therefore, I will here define exactly what is meant by the various terms used throughout this dissertation. Firstly, there are three types of Mandarin which will be pertinent to the current study: the variety spoken on Taiwan, known colloquially as Guóyǔ 國語 or, as it will be called here, Taiwan Mandarin (TM); the non-standard variety spoken by the Chinese arrivals to Taiwan in 1949, hereafter referred to as Mainlander Mandarin (MM); and the formal standard variety that can be found in textbooks, called Pǔtōnghuà 普通話 (PTH). In the literature, PTH is also at times called Mandarin, Peking

Mandarin, Standard Mandarin, Zhōngwén 中文, or simply Chinese. Secondly, in terms of Min, the specific variety present on Taiwan has been referred to by many names in the literature, including Taiwanese Southern Min, Mǐnnánhuà 閩南話, Taiwanese, Hokkien, Hoklo, Holo, Amoy, Fujianese, or Xiamen/Quanzhou/Zhangzhou Dialect. To distinguish between the language variety and the ethnic group that primarily speaks it, this dissertation will use the term Taiwanese Southern Min (TSM) to refer to the dialect and Hokkien to refer to the people. The label "Taiwanese" will apply to any resident of Taiwan whose forbearers were Chinese and arrived on the island before 1945. By contrast, any residents whose Chinese forbearers arrived between 1945 and 1949 will be referred to as Mainlanders.

### 1.2 Motivation for the Current Study

Though many comparative studies of Mandarin as spoken on mainland China and Taiwan have been conducted, most have focused on phonological differences between the two varieties. However, though they are often subtle, there are definite syntactic differences between these two types of Mandarin. Where they have been noted, these syntactic differences are usually attributed to interference from TSM, the home language of the majority of people on Taiwan. Whether this interference is characterized as imperfect learning or substratum interference, the underlying claim is that syntactic patterns from TSM have been borrowed or copied into Mandarin to form a variety unique to Taiwan. However, there are other theoretical approaches, like Replication Theory (RT), that frame this uniqueness differently, attributing it not to simple copying, but to speakers' creative agency in language contact situations. RT is a model of syntactic change in language contact situations that claims speakers create new grammatical

categories through the process of grammaticalization, taking inspiration from one of the languages in the contact situation without outright copying it (Heine & Kuteva 2005). Certain prominent features of TM syntax, like the *yǒu*+VP pattern, have been evaluated from a RT perspective, but many less obvious features have not. Beginning to address this gap is part of the motivation for the current study. The other part of the motivation relates to the assumptions typically made about which variety of a language was present in a contact situation. In most language contact studies, it is the standard variety of each language that is used as the basis for comparison, even when there is little to no reason to believe that the speakers in the contact situation actually spoke that variety. This has also been the case for studies of TM, which almost universally treat PTH as the language variety brought into contact with TSM, despite ample evidence existing that indicates otherwise. Whether it be personal anecdotes or census records, historical sources show that the Mandarin brought to Taiwan was decidedly non-standard. Because they do not contain a precise description of this non-standard Mandarin, though, any attempt to use this variety as a point of reference for language contact studies must first reconstruct it. In this study, the tools used for this reconstruction come from the field of World Englishes. Though, as the name implies, World Englishes is focused only on explaining variation between the different varieties of English found around the world, the models it has developed to do so can be applied equally well to other language contact situations. Of use in this study are the founder principle and the feature pool. The founder principle asserts that the features found in a contact variety are largely determined by which features are present in the speech of the populations actually involved in the contact situation (Mufwene 1996). The feature pool, which is conceived of as being analogous to a gene pool, offers a framework for predicting

which features are likely to be 'selected' (i.e. used) by a contact variety. Features with a high frequency of use and/or that are extant in all of the language varieties present in a contact situation have a high probability of selection, while features that are infrequently used or are rare among contributing language varieties will most likely not be selected (Mufwene 2020). By applying the feature pool concept to the case of TM, then, we can determine which features were present in the 'founder' variety of Mandarin (i.e. MM) brought to Taiwan in the 20th century, and consequently which features that variety contributed to the feature pool for TM. Once these features are selected, RT offers a framework for understanding and predicting how they change over time. Using data from informal spoken corpora collected in the early 1980s, mid 2000s, and early 2020s, this study argues that a RT approach extended to include these World Englishes concepts offers the best explanation of both historical data about TM syntax and current differences between regional varieties of TM.

### 1.3 Organization of the Dissertation

This remainder of this dissertation will be organized thusly: Chapter 2 will give a brief overview of the sociolinguistic history of Taiwan to provide context for Chapter 3, which discusses the theoretical approaches that have been used in past studies of Taiwan Mandarin and elucidates the extended Replication Theory framework used in the current work. Further details about Replication Theory, the founder principle and feature pool concept, and the reconstructed Mainlander Mandarin variety are also covered in Chapter 3. The methods of data collection and analysis, as well as information about the three corpora used in this study, will be explained in Chapter 4. Chapters 5 through 8 each examine one syntactic construction found in Taiwan

Mandarin — its treatment in the literature and its use in the corpora — and make a case for treating it as evidence for extended Replication Theory. Chapter 9 summarizes the findings presented in Chapters 5 though 8, addresses the weaknesses of the current study, and makes suggestions for future research.

# Chapter 2: A Brief Linguistic History of Taiwan

## 2.1 Introduction

Taiwan is comprised of one large island and several smaller islands located roughly 100 miles off the coast of southeast China. The main island is mostly mountainous on its eastern half and fairly flat on its western half. Its population of approximately 23.6 million people (Central Intelligence Agency [CIA] 2022) is made up of four major ethnic groups, each with their own native language(s): Aborigine, Hokkien, Hakka, and Mainlander. The Aborigines make up the smallest proportion of the population, despite having lived on Taiwan the longest and being the most diverse group, consisting of 16 officially-recognized peoples whose native languages belong to the Austronesian family. Once the only people on Taiwan, they currently account for only about 2% of Taiwan's total population. The other three ethnic groups that make up Taiwan's population are all some variety of Han Chinese, with Hokkiens accounting for 70% of the total population (CIA 2022) and Hakkas and Mainlanders each comprising 14% of the population (Country Reports 2022). While these groups do share an ethnicity in the broadest sense, they are still culturally and linguistically distinct, and may also be distinguished by the timing of their arrival on Taiwan.

## 2.2 Early Migration: Hokkien and Hakka

The earliest Chinese to arrive on Taiwan were the Hokkien, who started crossing the Taiwan Strait from the nearby region that is now Fujian province as early as the 11th century (Encyclopædia Britannica 2021). Since most of these early arrivals were fishermen by trade, the

Hokkien settled mainly along the western coast of Taiwan, often displacing the indigenous population and forcing them to move inland to the small valleys scattered throughout Taiwan's mountainous center (Kuo 2005). Because of this relocation, there was relatively little contact between Southern Min, which the Hokkien spoke, and the various aboriginal languages. While contact between mainland China and Taiwan would continue until 1895, the geographic separation between the Hokkien on the mainland and those on Taiwan led to the development of a uniquely Taiwanese variety of Southern Min, namely, Taiwanese Southern Min (TSM).

Hokkien immigration to Taiwan continued for the next several centuries, but it was not until the last few decades of the Ming Dynasty (1368-1644) that their numbers were great enough to become the majority on the island (Kuo 2005; Han 2017). And, as the Hokkien became the majority population on Taiwan, so too did TSM become the dominant language of the island. It is during this period of rising Hokkien dominance that the Hakka and their eponymous language variety began arriving in number, mostly from what are now Guangdong and Zhejiang provinces (Chiung 2001). The Hakkas' arrival on Taiwan was different from that of the Hokkien in a few key ways. Aside from the obvious differences of their later arrival and lower numbers, the Hakka also settled differently, living, with the exception of a stretch of coast in central Taiwan (Kubler 1981), in the inland valleys alongside the Aborigines (Kuo 2005). This, and longstanding ethnic tension between the Hakka and Hokkien, means that Hakka has historically had much more contact with the aboriginal languages of Taiwan than it has with TSM.

### 2.3 Japanese Imperialism

Further migration between mainland China and Taiwan was cut off before the end of the 17th century due to the seizure of Taiwan by Japan after the first Sino-Japanese war. For the 50 years between 1895 and 1945, Japanese culture and language were aggressively promoted in Taiwan, particularly after the outbreak of the Sino-Japanese war in 1937. Schools, which had previously used the local language as the medium of instruction (Scott & Tiuⁿ 2007), now taught in Japanese, with adults also receiving instruction in the language through targeted education programs (Kubler 1981). Additionally, while publications in Chinese were tolerated in the early years of occupation, Japanese quickly became the dominant language of media and public discourse (Wei 2006). This happened in stages, with the importation of written material from China banned in 1903, and the printing of any material in Chinese and use of any language but Japanese in schools outlawed in 1937 (Kubler 1981).

Attempts were also made to enforce Japanese as the language of private life, with public servants being required to use Japanese at home (Huang 1993) and a "Name-changing Campaign" begun in 1940 to pressure Taiwanese families to adopt Japanese names (Scott & Tiuⁿ 2007). Families who did so were rewarded with a larger portion of rationed food (Golovachev 2007, as cited in Brubaker 2012), no small thing during wartime. So although the Japanese themselves lived largely segregated from the Taiwanese in specially-constructed residential compounds in major cities (Brubaker 2012) and attended high-quality schools that few Taiwanese were allowed entry to (Kuo 2005), their language became an inescapable part of life on Taiwan. Consequently, by 1940, roughly 51% of the population knew some Japanese, with that number rising to 71% by 1944 (Scott & Tiuⁿ 2007), though fluency naturally varied by age

and social class (Brubaker 2012). Despite this, and despite the official attempts to replace local languages in the private sphere, TSM remained the home language of the majority of people on Taiwan (Scott & Tiuⁿ 2007).

### 2.4 Late Migration: Mainlanders

With the end of WWII in 1945, Japan was forced to return Taiwan to China, bringing the island under control of the Kuomintang (KMT) political party and a new set of aggressive language policies. These policies focused on the promotion of Pǔtōnghuà (PTH), a variety of Mandarin which the KMT had created somewhat artificially and implemented in 1932 as the standard language of China (Chen 1954). PTH combined Beijing phonology, northern Mandarin vocabulary, and modern Mandarin syntax (Cheng 1985) to form a generalized sort of Mandarin, and had been in use on the mainland in all government and military business for just over a decade by the time Taiwan again fell under Chinese purview. Unlike TSM's rise to prominence, which was accomplished in a bottom-up fashion through gradual demographic shift, PTH became a prominent language on Taiwan purely through top-down government policy. And while the KMT's promotion of PTH on the mainland did not aim to replace local dialects in informal spheres, the language policies instituted on Taiwan did (Chen 2001).

The first step towards this goal was to forbid all use of Japanese. This change was met with resistance from the educated and upper-class, who continued to use Japanese words and phrases in their speech as both a matter of convenience and a way to signal their social status (Kubler 1981). However, over time, as a new generation grew up in an environment where Japanese was no longer the lingua franca, Japanese ceased to be a commonly-spoken second

language on Taiwan. The second step, teaching everyone PTH, ran into some more substantial problems. To begin with, the government workers and teachers who came to Taiwan between 1945 and 1948 were a definite minority. Consequently, the average Taiwanese's contact with speakers of PTH was quite limited, and exposure to PTH came largely through language classes. Unfortunately, the quality of classroom exposure was generally low, given that there was a severe lack of qualified teachers (Kuo 2005; Tse 1986). Most classrooms were staffed by local teachers who were given only enough training to read the special characters in their new textbooks that allowed them to sound out texts using PTH pronunciation (Kubler 1981). One such teacher characterized the experience as "learn[ing] most of his Mandarin together with [his] students" (Kubler 1981, p. 39).

The situation of Mandarin on Taiwan changed in 1949, when the defeat of the KMT in the Chinese Civil war sent roughly two million government workers, soldiers, and refugees fleeing to Taiwan from mainland China. These newly-arrived "Mainlanders" settled primarily in urban centers on the north of the island, living in the residential compounds left behind by the Japanese (Brubaker 2012). Though often treated as a monolith, this group contained substantial diversity, with census records showing individual from all parts of China and seven of the ten Sinitic language families: Mandarin, Wu, Min, Yue, Xiang, Gan, and Hakka (Kuo 2005). Of these language families, Mandarin was the most strongly represented, with 43% of Mainlanders coming from regions where at least one dialect of Mandarin was spoken natively (Brubaker 2012). However, the majority of Mainlanders, like the Taiwanese, would have learned PTH as a second language. Their version of PTH, therefore, would have had the misalignment with formal standards typical of second-language speakers. The close contact imposed by their new living

situation between these diverse types of speakers makes it almost certain that dialect leveling took place in the PTH the Mainlanders used, resulting in a shared variety that likely differed from standard PTH in a number of key ways (more on this in Chapter 3). It is therefore prudent to draw a distinction between standard PTH and the version used by the Mainlanders, which will be referred to in this work as Mainland Mandarin (MM).

Though there were now, in the 1950s, more speakers of MM on Taiwan, the average Taiwanese still had little to no contact with fluent speakers of MM outside of the classroom and the few contexts in which ordinary citizens interacted with government officials. This, combined with the belief on the part of both the Mainlanders and Taiwanese that they were the more culturally advanced group, meant that tensions between the two populations were high and social interaction was low (Brubaker 2012). Furthermore, shortly before the arrival of most of the Mainlanders, a series of revolts had resulted in martial law being declared on the island. During much of this martial law period, languages other than MM were considered "a threat to national cohesion and unity" and their use was severely restricted (Hsiau 1997, pp. 306-307). Television stations could only show non-MM programming for a fraction of their total broadcast time, children faced fines and various humiliating or painful punishments if they were caught using a language other than MM in school, fluency in MM was required for all government positions, MM was the only language permitted in courthouses, and the military (in which all men over 18 were required to serve) outright banned use of non-MM languages (Brubaker 2012). While these policies were effective in quickly increasing MM proficiency across Taiwan, their harshness meant the relationship between Mainlanders and Taiwanese remained distant.

### 2.5 Shifting Power and Increased Language Contact

It was not until the 1960s, when the power of the Mainlanders within the KMT began to wane, that the social dynamics on Taiwan changed and TSM and MM came into frequent contact. Industrialization and the accompanying urbanization during the 1960s and 70s led to a sharp rise in enrollment at all levels of education, which in turn meant that more children and youths were spending significant portions of their day surrounded by MM (Mo 2000). Parents generally encouraged this, with some even refusing to speak their native language to their infants or toddlers in the hope that using only MM with them would better prepare them for the 'good' schools that would eventually get them into university (Gates 1981). Socially, this led to the younger urban generations on Taiwan growing up in a situation where MM was both the language of inter-ethnic communication and that of home life (Wei 2006). The resulting swell of educated young adults fueled the growth of a middle class on Taiwan that was not only eager to vote, but also to run for public office. By the early 1970s, the number of Taiwanese and Mainlanders in the KMT was approximately equal, and by the late 80s, nearly a third of the KMT Central Committee members were Taiwanese (Brubaker 2012). As a result of this shift in power, martial law was repealed in 1987, along with the punishments for speaking non-MM languages. The Taiwanese continued to gain in political power, making up 70% of the KMT by 1992 and 50% of its ruling body by 1993 (Hsiau 1998). These decades also saw the formation of the Democratic Progressive Party (DPP), whose platform included support for bilingual education and promoting the study of Taiwan's history and culture (Mo 2000). Under pressure from the growing DDP, the KMT changed national educational policy to prohibit discrimination

against non-MM languages (Huang 1997) and to allow for 40-50 minute mother-tongue classes to be offered as extra-curricular options in schools (Mo 2000).

The election of a DDP candidate to the presidency in 2000 added further momentum to the growing public desire for multilingualism. A new curriculum was instituted in schools in 2001 which required all students to spend one hour per week in one of the "native" (i.e. non-MM) language classes that had previously only been extracurriculars (Wei 2006). The regulations limiting non-MM programming in media were lifted (Klöter 2017), and code-mixing of MM and local languages, particularly TSM, began to be used as a rhetorical device in speeches, commercials, daily conversation, and so on (Wei 2008). TSM by no means replaced MM, as over 90% of Taiwan's population still had communicative ability in MM (Tsao 2008), but social consciousness had reoriented towards a more multicultural linguistic identity. Even when the KMT regained power in 2008, multicultural and multilingual ideologies had gained too strong a foothold in Taiwanese society to be fully reversed.

## 2.6 Conclusion

In its history, Taiwan has undergone several shifts in dominant language. Some, like the shift from Austronesian languages to TSM, happened gradually as a result of human migration and the accompanying demographic changes. Others, like the imposition of Japanese and Mandarin, happened abruptly as a result of political changes. Because it was imposed by a minority elite, MM was not a social language for most Taiwanese for several decades. After rapid industrialization and urbanization in the 1960s and 70s led to higher levels of school enrollment, however, MM became a social language for the urban youths who would become

Taiwan's middle class. Increasing involvement of Taiwanese in governance from the 1970s-1990s lead to a relaxing of Mandarin-only policies and an increase in use of TSM in public life. Since 2000, Taiwan has continued to maintain and promote this new multilingual identity and ideology.

# Chapter 3: Theoretical Background

## 3.1 Introduction

This chapter explains the theoretical framework of this study and situates it within the broader language contact field. Some of the contact literature on TM specifically is discussed, as well as more general approaches to language contact, such as Thomason and Kaufmann's (1988) borrowability scale from sociolinguistics, Heine and Kuteva's (2005) Replication Theory (RT), and Mufwene's (1996, 2001) founder principle and feature pool from World Englishes. This chapter ends by proposing an extended RT approach that incorporates World Englishes concepts into the RT framework.

## 3.2 Comparative Studies

Many of the earliest linguistic analyses of TM acknowledged the social factors salient to the Taiwan context, but their actual methodology was more of a direct comparison of linguistic features and patterns, similar to what is used in historical linguistics to trace language genealogies. These studies generally set TM next to PTH, identified places where the TM pattern diverged from the standard model, then attributed those non-standard features to the influence of another language found on Taiwan, typically TSM. Using this approach, researchers have credited various features of TM syntax — such as obligatory use of auxiliary verbs as operators (Cheng 1985) or the omission of the subordinating marker after adjective clauses with definite heads (Kubler 1981) — to contact with TSM, Japanese, Hakka, etc. Underlying this framework, then, are certain assumptions from the field of Second Language Acquisition (SLA),

namely, that adult learners of a language will unavoidably incorporate grammatical patterns of their L1 into their L2. Changes in the frequency of a feature are also sometimes attributed to L1 influence. For example, the use of *yángzi* 樣子 'type/kind' at the end of descriptive sentences is infrequent in PTH, but highly frequent in TSM. Under the influence of TSM, therefore, TM also uses this pattern much more frequently than does PTH (Kubler 1981).

While this approach offers a great deal in terms of explaining non-standard features in contact varieties, it does have a fairly conspicuous weakness: it cannot explain the absence in TM of patterns that are found in both PTH and TSM. For instance, TM does not make an inclusive/exclusive distinction in the first person plural, which both TSM and PTH do.

(1) Comparison of PTH, TSM, and TM: inclusive/exclusive pronoun[1]

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **PTH:** | 我們 / 咱們 | 這 | 個 | 店 | 樣樣 | 都 | 有 |
| | wǒmen/zánmen | zhèi | ge | diàn | yàngyang | dōu | yǒu |
| | 1PL.EXCL/1PL.INCL | this | CL | shop | everything | all | have |
| **TSM:** | 阮 / 咱 | 這 | 個 | 店 | 每樣 | 都 | 有 |
| | góan/lán | chit | keng | tiàm | tàk-hāng | lóng | ū |
| | 1PL.EXCL/1PL.INCL | this | CL | shop | everything | all | have |
| **TM:** | 我們 / *咱們 | 這 | 個 | 店 | 樣樣 | 都 | 有 |
| | wǒmen/*zánmen | zhèi | ge | diàn | yàngyang | dōu | yǒu |
| | 1PL.EXCL/*1PL.INCL | this | CL | shop | everything | all | have |

"Our shop has everything."

(adapted from Cheng 1985, p. 368)

[1] For ease of comparison, the Standard Mandarin characters will be used in all TSM examples, with the exception of pronouns.

TM is also distinct from both TSM and PTH in that it does not double-mark continuation by using both a time adverb and sentence-final particle.

(2) Comparison of PTH, TSM, and TM: marking continuation

| | | | | | |
|---|---|---|---|---|---|
| **PTH:** | 他 | 還 | 在 | 學校 | 呢 |
| | tā | *hái* | *zài* | *xuéxiào* | *ne* |
| | 3.SG | still | located.at | school | CONT |
| **TSM:** | 伊 | 還 | 在 | 學校 | 呢 |
| | i | *iáu* | *tī* | *hàk-hāu* | *leh* |
| | 3.SG | still | located.at | school | CONT |
| **TM:** | 他 | 還 | 在 | 學校 | *呢 |
| | tā | *hái* | *zài* | *xuéxiào* | **ne* |
| | 3.SG | still | located.at | school | *CONT |

"He is still at school."

(adapted from Cheng 1985, p. 371)

Additionally, while the comparative framework can do a great deal to explain why non-PTH-like features of TM are the way they are, it has very little to offer in terms of predictions about which features are most likely to be affected by language contact. To answer this sort of question, scholars typically take a sociolinguistic approach.

### 3.3 Sociolinguistic Approaches

The majority of work on contact linguistics has been done from a sociolinguistic perspective. This is unsurprising, given that the field of contact linguistics is generally considered to have begun with Uriel Weinreich's monograph *Languages in Contact* (1953), in which he claims outright that linguistic outcomes in contact situations are primarily conditioned

by sociocultural variables. Extent and degree of bilingualism, length of contact, geographical and demographic distribution, social factors (such as religion, race, gender, and age), use in different social functions, and political and ideological factors are all identified by Weinreich (1953) as salient factors in a language contact situation. Depending on the situation of these social factors, he claims, linguists can make predictions about the type and extent of linguistic change that will occur in a given contact situation. Thomason and Kaufmann (1988) build on this ideological foundation in their highly influential book *Language Contact, Creolization, and Genetic Linguistics*, asserting that there are no absolute linguistic constraints on the type or degree of contact-induced change, and that linguistic factors play a subordinate role to social factors in determining the outcome of contact-induced change. Instead, they propose a borrowability scale with two broad social circumstances on either end of a spectrum of types and amounts of change: borrowing and shift. In a borrowing situation, native speakers of a language incorporate foreign elements into their speech. When contact is casual and there is little bilingualism, these foreign elements are typically lexical items (Thomason & Kaufman 1988). As contact becomes more intense and lengthy, and the extent of bilingualism increases, the elements borrowed become more structural in nature, involving phonological and syntactic patterns. Such changes, however, often require multiple centuries worth of intense contact to occur (Thomason & Kaufman 1988).

By contrast, in a shift paradigm, changes are usually structural, and occur within the realms of phonology and syntax. In a shift situation, one group learns the language of another, meaning there is widespread (though not necessarily universal) bilingualism within at least the shifting group. If those speakers are fluently bilingual and well-integrated into the target

language (TL)-speaking community before language loss occurs within their L1, then differences in their production of the TL compared to native speakers' will be few and minor (Thomason & Kaufman 1988). However, in cases where the shift occurs too rapidly for the shifting group to become fully bilingual before language loss affects the L1, the shifting group's production of the TL will diverge significantly from that of native speakers (Thomason & Kaufman 1988). In either case of shift, whether or not changes spread throughout the entire TL-speaking community depends on the relative size of the shifting group: if they outnumber the native TL-speakers by a substantial amount, then the TL as a whole is likely to change, regardless of the attitude of the native speakers towards those changes (Thomason & Kaufman 1988). Furthermore, if these changes to the TL include typologically marked contrasts and patterns that the shifting group has carried over from their L1, the grammar of the TL might become more complicated, rather than less, despite grammatical simplification being a more common result of shift situations (Thomason & Kaufman 1988). Of additional interest on this point is Thomason and Kaufman's (1988) claim that these carried-over marked structures are themselves changed in the processes of transfer, frequently manifesting in the TL in ways that are not identical to their use in the shifting group's native language. Unfortunately, they do not provide a mechanism for how such innovations occur, nor why certain structures undergo this process while others do not.

### 3.4 Replication Theory

Replication Theory (RT) proposes an approach to language contact that is substantially different from sociolinguistic models. To begin with, RT limits its consideration to grammatical meanings within the syntactic domain. Thus, any sort of phonological or lexical transfer is

outside the scope of RT. Another major difference is the role of social factors. While both sociolinguistic models and RT assume some level of bilingualism in the speaker community, beyond that, Heine and Kuteva (2005, p. 13) claim that there are "no significant sociolinguistic parameters that regularly correlate with the presence or absence of, or distinctions between, specific types of grammatical replication" in a contact situation. Finally, in contrast to Thomason and Kaufmann's paradigm, RT proposes both constraints on the type of contact-induced innovations that can occur and a mechanism for their creation.

The first constraint is the mechanism of change itself: grammaticalization. Grammaticalization is a unidirectional process by which linguistic forms and constructions gain gradually more grammatical meanings over time. The grammaticalization process has four stages: extension, desemanticization, decategorialization, and erosion. In extension, use of an existing linguistic expression is extended to new contexts, taking on new meanings as it is reinterpreted in the new context. Desemanticization, also sometimes termed "semantic bleaching", involves the loss of specific meaning in favor of more general or grammatical meaning. In decategorialization, morphosyntactic properties associated with the expression's original form are lost. Finally, erosion, also called "phonetic reduction" takes place and the expression loses some of its original phonetic substance (Heine & Kuteva 2005, p. 15). While this is an entirely common process that frequently occurs within a language, it can also be triggered by language contact. In such cases, either ordinary contact-induced grammaticalization or replica grammaticalization will take place.

Figure 1. Types of contact-induced grammaticalization

| | **Ordinary contact-induced grammaticalization** | **Replica Grammaticalization** |
|---|---|---|
| 1 | Speakers notice that in the model language (M) there is a grammatical category Mx. | |
| 2 | They create in the replica language (R) an equivalent category Rx on the basis of the linguistic material available in R. | |
| 3 | They draw on universal strategies of grammaticalization, using construction Ry in order to develop Rx. | They replicate the grammaticalization process they assume occurred in M, using an analogical formula such as [My > Mx] : [Ry > Rx]. |
| 4 | They grammaticalize Ry to Rx. | |

(Heine & Kuteva 2005)

As shown in Figure 1, both types of grammaticalization proceed identically in the initial stages. In stage 1, speakers notice that there is a grammatical category Mx in the model language M. In stage 2, speakers use the linguistic material available in the replica language R to create an equivalent category Rx in R. It is in stage 3 where the two processes diverge. Ordinary contact-induced grammaticalization employs universal strategies of grammaticalization to develop an existing use pattern Ry into the desired new structure Rx. An example of this would be that of dual pronouns in Tayo, a French-based creole that formed in St. Louis, New Caledonia in the 1860s. The main Melanesian languages spoken in St. Louis at that time both have an obligatory semantic category of dual (Mx), which French does not. Presumably to address this perceived gap, speakers created a dual form in Tayo by taking the French word for 'two' *deux* (Ry) and grammaticalizing it into a dual suffix *-de* (Rx) (Corne 1995). Because there is no evidence in the Melanesian languages near St. Louis of 'two' being the source of the dual marker, we cannot say that Tayo speakers replicated a grammaticalization process they observed in the model languages. Instead, they drew on universal strategies of grammaticalization, making this a case

of ordinary contact-induced grammaticalization. Replica grammaticalization, by contrast, occurs when speakers follow the process of grammaticalization they assume took place in the model language to create Mx, introducing an element of agency into the grammaticalization process. An example of this process is the creation of the immediate perfective construction in Irish English. In Irish Gaelic, the immediate perfective aspect (Mx) is encoded using the preposition 'after' in the structure 'X is after Y' where 'Y' may be either an NP or a non-finite VP. Because 'after' is used in both prepositional structures (My) and this perfective structure, speakers can infer the pathway of grammaticalization. Around the late seventeenth century, this grammaticalization process was replicated into Irish English, resulting first in sentences like "He's after the flu" *He just had the flu*, where 'Y' could only be an NP, but grammaticalizing over time to also allow non-finite VPs, like "She's after selling the boat" *She has just sold the boat* (Sullivan 1980 in Heine & Kuteva 2005). Thus, the immediate perfective is not a case of polysemy copying/calquing/loan translation (more on this later), but one of replica grammaticalization. What both of these examples illustrate is that when new syntactic structures develop as a result of language contact — whether via ordinary contact-induced grammaticalization or replica grammaticalization — specific stages of the process are predictable, namely, extension, desemanticization, decategorialization, and, finally, phonological erosion.

Constraints on the selection of features for grammatical replication to some extent follow from the grammaticalization process. After all, syntactic categories and structures that do not exist in the model language cannot serve to trigger replication, and since the new Rx is formed out of already-existing use pattern Ry, the structure of the replica language also places

restrictions on what replications are possible. In this way, model and replica languages impose both individual and joint restrictions on what replications are possible. However, more salient in this regard is the idea of equivalence, as the choice of Ry is not random, but rather made on the basis of its perceived similarity to the model Mx. As Heine and Kuteva (2003, p. 562) put it,

> In order to develop a structure that is equivalent to the one in the model language, speakers choose among the use patterns that are available in the replica language the one that corresponds most closely to the model, frequently one that until then was more peripheral and of low frequency of use, and they activate it—with the effect that a peripheral pattern gradually turns into the regular equivalent of the model, acquires higher frequency of use, and eventually it may emerge as a full-fledged grammatical category.

By use pattern, Heine and Kuteva (2005) mean a recurrent piece of linguistic discourse that is associated with a specific grammatical meaning, but not obligatorily. To illustrate, in German, both *Herbstzeit* and *Zeit des Herbstes* can be used to say 'autumn'. *Herbstzeit* (lit. 'autumn time') follows the nominal compounding pattern, which is a major (i.e. frequent) use pattern in German, while *Zeit des Herbstes* (lit. 'time of the autumn') follows the possessive/genitival pattern, which is a minor (i.e. infrequent) use pattern in German. Neither use pattern can be said to be obligatory, since the other is always a grammatical option. What distinguishes use patterns from grammatical categories is that use patterns are more restricted in their occurrence, usually found only in a particular regional variation, social stratum, register, etc. (Heine & Kuteva 2005). To continue with the German example, while nominal compounding may be the major use pattern in Standard German, in the variety spoken in eastern Belgium, the possessive/genitival *Zeit des Herbstes* pattern has been developed into the major use pattern after the example of

French, which is the majority language in that region (Riehl 2001 in Heine & Kuteva 2005). This transition from minor to major use pattern, Heine and Kuteva (2005) claim, is how the extension stage manifests in most cases of grammatical replication. However, unlike the constraints imposed by the principles of grammaticalization, which are universal, the restrictions related to equivalence are situation-specific and require knowledge of the structure of the language varieties in contact.

This knowledge is also important when it comes to distinguishing grammatical replications from other types of historical or contact-induced changes. For example, grammaticalizing a lexical verb into an auxiliary verb is a fairly common diachronic change cross-linguistically, but such a development may still be considered contact-induced if there is evidence that the grammaticalization was accelerated by language contact. Acceleration is difficult to measure, but can be indicated by synchronic regional variation: if the dialect of the region in closest contact with the model language uses a more grammaticalized form of a feature than dialects from other regions, then contact-induced acceleration has likely occurred (Heine & Kuteva 2010). Take, for instance, comparative and superlative adjective constructions in English. As an Indo-European language, English inherited suffixal inflections *-er* and *-est* and suppletive forms (e.g. good, better, best) to express degrees of comparison. However, these forms have a tendency to be replaced with analytic constructions (e.g. exciting, more exciting, most exciting), especially in the Romance languages. And when looking at regional varieties of English in the 1600s, analytic comparatives that were rarely used in the north had been in use in the south (which was much more in contact with France) for a century (Danchev 1989 in Heine

& Kuteva 2005). These circumstances are strongly suggestive of acceleration, and therefore the change may be views as contact-induced.

By contrast, cross-linguistically unusual features are often more unambiguously the result of language contact. For instance, if both the model and replica language have an unusual feature, but only the languages genetically related to the model language also have that feature, then the existence of that feature in the replica language is almost certainly due to contact-induced transfer (Heine & Kuteva 2005). However, for this transfer to be a case of grammatical replication, there must of course be evidence that it is grammaticalization that accomplished the emergence of the feature in the replica language. Here, knowledge of the pre-contact structures of the involved languages is required. Not only is it necessary that the corresponding structures Mx and Ry existed in each language variety, Ry must necessarily have been less grammaticalized than Mx. Without this gap in degree of grammaticalization, after all, there would be no need for a process to create Rx. Additionally, because of this gap, a replication will first appear in the replica language in a less grammaticalized form (Ry) than its equivalent in the model language (Mx). This is the key difference between grammatical replication and polysemy copying and calquing/loan translation (Heine & Kuteva 2010). Calquing/loan translation involves translating an expression word-for-word from one language into another (e.g. *honeymoon* in English to *luna del miel* in Spanish) and polysemy copying is the imitation of a pattern whereby a single wordform is associated with multiple related meanings (e.g. the use an interrogative pronoun as a relative clause marker). As both of these types of transfer involve imitation, the pattern used in the replica language will be immediately identical to that of the model language, with no restrictions in contexts of use or use of less grammaticalized forms. With grammatical

replication, however, neither of these things are the case. To begin with, as mentioned previously when discussing use patterns, the immediate result of contact is often merely an increase in the frequency of a minor use pattern, rather than the creation of an entirely new pattern (Heine & Kuteva 2005). Additionally, in cases of ordinary contact-induced grammaticalization, where the model language shows no evidence of how the model category (Mx) was developed, there is no polysemy pattern to be copied in the first place. Furthermore, in both ordinary contact-induced and replica grammaticalization, replications start out more contextually constrained than the model construction, as is the case with indefinite articles in Basque. On the basis of the French model, where the numeral 'one' (My) has grammaticalized into the indefinite article (Mx), Basque speakers have begun to grammaticalize their own 'one' (Ry). However, the Basque indefinite article is more contextually restrained than the French model, occurring almost entirely in cases of specific reference and often being treated as optional in cases where it would be obligatory in French (Heine & Kuteva 2003). This cannot be treated as polysemy copying, since the whole polysemy pattern has not, in fact, been copied. Finally, as a process, grammatical replication involves an intermediate stage where the new category (Rx) is less grammaticalized than the corresponding model category (Mx), resulting in ambiguity between the lexical and grammaticalized meanings (Heine & Kuteva 2005). To return to the case of the immediate perfective in Irish English, while both Irish Gaelic and Irish English may currently employ the use pattern 'X is after Y' with non-finite VPs, Irish English for some time could only use an NP 'Y' in this construction (Sullivan 1980 in Heine & Kuteva 2005). That the replica category (Rx) went through a less grammaticalized and more restricted stage before reaching full congruence

with the model category (Mx) is evidence that what happened was grammatical replication, not polysemy copying or calquing.

Grammatical replication, therefore, is a process distinct from other types of grammatical transfer. Rather than copying a pattern, grammatical replication copies a process (or, in the case of ordinary contact-induced grammaticalization, initiates a process). Because this process takes time to play out, diachronic data is necessary to identify replications. It is also why the "spontaneous replications" often produced by bilinguals during the initial period after language contact cannot be considered grammatical replications unless an idiosyncratic innovation is taken up and used regularly by other speakers such that it acquires stability across time (Heine & Kuteva 2010). Though RT considers only the syntactic domain, it is still a useful framework in the field of language contact, as it provides a clear mechanism of change, a set of constraints on the possible candidates for change, and predictions for the outcomes of contact-induced language change.

### 3.5 World Englishes

In analyses of contact varieties from both sociolinguistic and RT perspectives, linguists have often implicitly assumed that speakers in the contact situation spoke the standard variety of their language. However, scholarly work in the field of World Englishes has demonstrated that this approach is not only historically inaccurate, but obscures valuable information about the genesis of linguistic features in contact varieties. Just as variation among regional dialects of American English can largely be explained by identifying which region the British immigrants who settled there came from, so too can otherwise puzzling linguistic features in a contact

variety often be traced back to a non-standard dialect. This idea that the characteristics of a contact variety are largely dependent on the characteristics of the specific dialects spoken by the populations in contact is known as the founder principle (Mufwene 1996).

While the founder principle provides a starting point for explaining the structure of contact varieties, there is still the question of why certain features are selected for retention and others are not. To answer this question, Mufwene (2001) proposed the notion of a feature pool: an abstract collection of all the linguistic features present in a speaker community, from which individual speakers select a subset to make up their idiolect. Circumstances both social and linguistic impact which features are selected, including population structure, overall frequency, representational frequency, and markedness (Lim 2020; Cheshire et al. 2011 in D'Arcy 2020; Mufwene 2020). Population structure has to do with the type of colony an English developed in: settlement, trade, or exploitation (Lim 2020). The relevant type for this study is the exploitation colony. In these colonies, settlers were a minority group that nevertheless ruled over the indigenous population for the purpose of extracting raw materials that could be sent back to England. As a result, English was spread primarily (though of course, not exclusively) via English-medium education, and the dialects most people were exposed to were that of their teachers, who were not necessarily native speakers of English (Lim 2020). Often in these colonies, English became an interethnic lingua franca, meaning that local populations were using English primarily to communicate with each other, rather than with the native-speaker colonists. In such situations, a greater degree of influence is expected from the local language(s), as their features will be more frequent in the feature pool (Mufwene 2020). This is the case for the formation of Irish English. During the genesis of this dialect, frequent (in terms of overall

frequency) features of Irish Gaelic, like the substitution of /hw/ for /w/ in word-initial position, were frequently contributed to the feature pool by bilingual speakers. Because of this frequency, these features were selected for use in Irish English (Mufwene 2020). When Irish English was brought to America, however, this feature was not common among the Englishes spoken there (representational frequency), nor was the population of Irish immigrants large enough to comprise a significant portion of the overall population (overall frequency). Thus, this feature was not selected for use in American English. However, in places where a larger proportion of the European settlers were from Ireland, such as the Midland region, these features were frequent enough within the feature pool to be selected for use in the regional variety. Frequency also plays into Mufwene's (2001) idea of markedness, which he defines relativistically as how common (i.e. frequent), transparent, regular, salient, semantically substantive, etc. a feature is *in its language ecology*. So a feature like numeral classifiers, while uncommon globally (Her et al. 2022), would not be considered marked in the ecologies that gave rise to the Melanesian pidgins, as many of the languages in that region of the world use numeral classifiers (Mufwene 2002). Combined with the founder principle, the feature pool concept is a useful tool for predicting language contact outcomes.

### 3.6 Extended Replication Theory

While ideas from World Englishes may have, due to the inherent focus of the field, been applied only to contact situations involving English, there is nothing in them that is necessarily English-specific. From a strictly theoretical perspective, in fact, the field would perhaps be better termed "colonial languages". In that case, since the history of Mandarin on Taiwan

strongly resembles that of English in Britain's exploitation colonies, there seems ample reason to assume that concepts like the founder principle and feature pool would be eminently applicable to Taiwan Mandarin. These concepts are also far from incompatible with the RT approach, and in fact complement it quite well. For instance, RT places emphasis on having knowledge of the pre-contact structures of the model and replica languages, while the founder principle stresses the need to be specific about which dialects we treat as model and replica. This is particularly important in the case of TM, as PTH, despite being the point of comparison in many accounts of TM, was not the dialect of Mandarin the Mainlanders actually spoke. This is because, firstly, PTH was (as mentioned in Chapter 2) not promulgated as the standard dialect until 1932 (Chen 1999), so when the Mainlanders arrived in Taiwan in 1949, a person who grew up speaking PTH from infancy could have been a maximum of 17 years old. Thus, the vast majority of Mainlanders would have learned PTH as adults, and as adult learners, they would have (unavoidably, from a linguistic perspective) incorporated phonological and grammatical elements from whatever Chinese language was native to them into their production of PTH. Many of these Chinese languages would have been distinctly unlike PTH and the broader Mandarin language family to which it belongs. Looking at the 1956 civilian census records[2], appended in Kuo (2005) with information about the dialect(s) spoken in the various provinces of origin[3], we can see that 65.9% of Mainlanders came from a province where at least some of the population spoke one of the southern Chinese languages natively, with 27.5% coming from provinces where only southern Chineses were spoken natively.

---

[2] i.e. The 270,000 soldiers in the KMT army were excluded from this count.

[3] Note that provincial boundaries of the time do not match modern ones.

Table 1. Population of Mainlander civilians in Taiwan in 1956

| Origin | Number | Proportion | Dialects/languages |
|---|---|---|---|
| **Shandong** | 95,845 | 10.33% | N Mandarin |
| **Liaoning** | 14,084 | 1.52% | N Mandarin |
| **Beijing** | 7,850 | 0.85% | N Mandarin |
| **Tianjin** | 5,293 | 0.57% | N Mandarin |
| **Jilin** | 2,060 | 0.22% | N Mandarin |
| **Liaobei** | 1,773 | 0.19% | N Mandarin |
| **Andong** | 1,623 | 0.17% | N Mandarin |
| **Rehe** | 789 | 0.08% | N Mandarin |
| **Heilongjiang** | 1,046 | 0.11% | N Mandarin |
| **Chahar** | 550 | 0.06% | N Mandarin |
| **Nenjiang** | 479 | 0.05% | N Mandarin |
| **Songjiang** | 387 | 0.04% | N Mandarin |
| **Shuiyuan** | 383 | 0.04% | N Mandarin |
| **Hejiang** | 192 | 0.02% | N Mandarin |
| **Xing'an** | 98 | 0.01% | N Mandarin |
| **Mongolia AR** | 338 | 0.04% | N Mandarin, NW Mandarin, Altaic |
| **Hebei** | 36,124 | 3.89% | N Mandarin, NW Mandarin |
| **Shaanxi** | 6,504 | 0.70% | NW Mandarin |
| **Shanxi** | 5,282 | 0.57% | NW Mandarin |
| **Gansu** | 1,358 | 0.15% | NW Mandarin |
| **Qinghai** | 131 | 0.01% | NW Mandarin |
| **Ningxia** | 88 | 0.01% | NW Mandarin |
| **Anhui** | 44,533 | 4.80% | N Mandarin, SW Mandarin,  E Mandarin |
| **Xinjiang** | 277 | 0.03% | Altaic, Indo-European |
| **Henan** | 41,674 | 4.49% | N Mandarin, SW Mandarin |
| **Tibetan AR** | 16 | 0.00% | NW Mandarin, Tibetan |
| **Xikang** | 313 | 0.03% | SW Mandarin, Tibetan |
| **Sichuan** | 37,363 | 4.02% | SW Mandarin |
| **Guizhou** | 4,545 | 0.49% | SW Mandarin |
| **Yunnan** | 5,716 | 0.62% | SW Mandarin |

| | | | |
|---|---|---|---|
| **Hubei** | 37,802 | 4.07% | SW Mandarin, Hakka Gan, Xiang E Gan |
| **Hunan** | 54,154 | 5.83% | SW Mandarin, Gan, Xiang |
| **Guangxi** | 11,620 | 1.25% | SW Mandarin, Tai, Xiang, N Mandarin, Yue, Hakka |
| **Jiangsu** | 108,327 | 11.67% | E Mandarin, Wu, N Mandarin |
| **Shanghai** | 16,179 | 1.74% | Wu |
| **Zhejiang** | 114,830 | 12.37% | Wu, Min |
| **Jiangxi** | 30,666 | 3.30% | Gan, Hakka, Wu, Min |
| **Fujian** | 142,520 | 15.35% | Hakka, N Min, S Min, E Min, N Mandarin |
| **Guangdong** | 93,431 | 10.06% | Yue, Hakka, S Min |
| **Hainan** | 1,817 | 0.20% | S Min, Hakka, N Mandarin |
| **Unknown** | 219 | 0.02% | Unknown |
| **Total** | **928,279** | **100.00%** | |

(adapted from Kuo 2005, pp. 77-78)

So even though the Mainlanders would have used Mandarin to communicate with each other, only a very small percentage of them can reasonably be expected to have spoken PTH. Instead, they would have spoken either a non-standard Mandarin (e.g. Eastern Mandarin, Southwestern Mandarin, etc.) or an L2 Mandarin influenced by the features of southern Chineses. The contact between these different Mandarin varieties would have created a feature pool containing the distinctly northern features that characterized much of PTH, but also features unique to the southern Chineses. In such situations, the feature pool framework predicts that features which are infrequent within a contributing variety and/or rare among contributing varieties are the least likely to be selected for the output variety. In other words, both decidedly northern and decidedly southern features would not survive the selection process to exist in the new Mandarin. Therefore, by assuming that infrequent features and/or features used only by a minority of Chinese languages would not be selected from the feature pool, we can reconstruct a version of Mandarin that is much closer to what the Mainlanders actually spoke, a version that I have

designated Mainlander Mandarin (MM). The features that will be lost in this sort of dialect leveling can be fairly confidently predicted, as a similar process has taken place on mainland China. In Central China, where Northern Chineses and Southern Chineses meet, various "transitional" dialects can be found that make up a typologically distinct group of Central Chineses, which includes multiple dialects of Mandarin (Szeto 2019). As a feature pool approach would predict, Central Chineses do not have features that are rare in both Northern and Southern Chineses (e.g. 3+ term demonstrative systems are found in 24.6% of Northern Chineses, 12.9% of Southern Chineses, and 0% of Central Chineses), but can have features that are rare in one but very common in another (e.g. post verbal temporal adverbs in VPs are found in 0% of Northern Chineses, 100% of Southern Chineses, and 12.5% of Central Chineses). They also contain features found in both Northern and Southern Chineses at rates intermediate between the two (Szeto 2019). These facts support the validity of reconstructing MM via application of the feature pool concept. Additionally, using a reconstructed MM instead of PTH as the point of reference when discussing TM makes some of the features that puzzled early scholars of the dialect, such as the lack of inclusive/exclusive distinction in the first person plural pronoun, not only likely in the general sense, but entirely predictable in the specific case of TM. Since the inclusive/exclusive distinction is not particularly common across Chineses, it would have been rare in the feature pool of the Mainlanders and therefore very unlikely to have existed in MM. That being the case, only TSM would have been contributing this distinction to the feature pool for TM, making it a feature with low relative frequency and consequently a low probability of selection. Thus, combining the feature pool concept with knowledge about the population demographics of the Mainlanders allows us to reasonably reconstruct a variety for

which we do not have clear historical records. This in turn enables us to more reliably identify cases of grammatical replication, since our knowledge about the pre-contact (or at least time-of-contact) state of the replica language will be more accurate.

Another aspect of TM that may be informed by the idea of a feature pool is that of regional variation. RT proposes that grammaticalization can be accelerated by contact with a language variety further along in that process. Evidence of this contact-induced acceleration can be found by contrasting the version of the replica language spoken in the area of most intense contact with the model language with the version spoken in the area of least intense contact (Heine & Kuteva 2010). In other words, in places where a larger portion of the feature pool is comprised of inputs from the model language, contact-induced grammaticalization can be expected to be the most advanced. In Taiwan, where Mainlanders mainly settled in northern cities — 67.7% in the greater Taipei area alone (Chiung 1999 in Brubaker 2012) — the area of most intense contact with the model language (TSM) would have been the south. Thus, the replications under consideration in this study are expected to be most grammaticalized in the southern variety of TM and least so in the northern variety.

Figure 2. Features under consideration

| | TSM | MM | TM |
|---|---|---|---|
| **(Mei)You + VP** | • *ū* +VP as auxiliary:<br>  • habitual action<br>  • assert (non)occurrence of event<br>  • yes/no response<br>• perfective aspect | • *yǒu-méi-yǒu* + VP as perfective question<br>  • *yǒu* + VP as affirmative answer | • *yǒu* + VP as auxiliary:<br>  • habitual action<br>  • assert (non)existence of state<br>  • completed action<br>  • yes/no response |
| **Shuo** | • Stage 5: Full complementizer<br>• Discourse marker:<br>  • Hearsay<br>  • Counter-expectation | • Stage 1: Quotative<br>• (Stage 2: Semi-complementizer?) | • Stage 4: Complementizer<br>• Discourse marker:<br>  • Hearsay<br>  • Counter-expectation<br>  • Intensifier |
| **Yong** | • Co-verb:<br>  • Instrumental Case<br>  • Main predicate<br>  • Overt marking of verb nominalization<br>• Pro-verb | • Co-verb:<br>  • Instrumental Case | • Co-verb:<br>  • Instrumental Case<br>  • Main predicate<br>• Pro-verb |
| **A-not-A questions** | • Only auxiliaries<br>  • Exception: 'to know', 'good', 'right'<br>• Disjunction<br>  • Exception: copula<br>• Deletion of second syllable of bisyllabic verb in first 'A' | • Main verbs, auxiliaries, and adjectives | • Main verbs, auxiliaries, and adjectives<br>• Deletion of second syllable of bisyllabic verb in first 'A' |

### 3.7 Conclusion

This chapter discussed the strengths and weaknesses of various theoretical approaches to language contact before proposing an extended form of RT that includes concepts from the field of World Englishes. In comparison to previous studies of TM, it is hoped that this extended RT framework will offer a more accurate picture of the language varieties in contact, an explanation

for why certain features are selected for use and others are not, and a better explanation for variation within the contact variety. In the next chapter, the methodology of the present study will be laid out.

# Chapter 4: Methodology

## 4.1 Data Collection

### *4.1.1 Corpora*

Three corpora — referred to in this work as the Old Corpus, NCCU Corpus, and New Corpus — were used in this study. The Old Corpus was created from the contents of the book "Variations of Spoken Standard Chinese Volume 2: A Speaker from Taipei" by Cornelius C. Kubler and George Ho. This book contains four transcripts of informal interviews in Mandarin between George Ho and Taipei residents. Of the five speakers, three were men and two were women. All but one speaker came from- and had parents who came from northern Taiwan, with the single exception coming from the south. Speakers ranged from 24 to 63 years of age and had education levels from elementary to graduate school. Though the specific date of the interviews is not given, the book was published in 1984 as a product of a project launched in 1981, so the interviews must have occurred at some point in that three-year range. With each interview, an annotation is included that clarifies non-standard terms or pronunciations and gives context to references to specific districts, schools, etc. in Taipei that the reader may be unfamiliar with. Because these transcripts were originally written using pinyin, they were re-typed in traditional characters for this study.

The NCCU Corpus was created from a selection from the larger NCCU Corpus of Spoken Taiwan Mandarin (formerly the NCCU Corpus of Spoken Mandarin). The NCCU Corpus of Spoken Taiwan Mandarin is made up of informal face-to-face conversations and has been adding conversations since 2006. Each conversation has been transcribed using traditional

characters and includes the age, sex, and relationship of the participants, as well as the date of recording. The conversations selected for inclusion in the NCCU Corpus used in this study were those collected in 2006 or 2007 between speakers who were 20-25 years old. These limitations were imposed to ensure that the speakers in the different corpora would belong to different social generations. Of the 18 speakers in the NCCU Corpus, 12 were female and 6 were male. No information was given about where the speakers were born or grew up, or what languages they spoke besides TM.

The New Corpus is made up of four Mandarin conversations recorded in 2023 by the researcher and one conversation from the NCCU Corpus of Spoken Mandarin that was recorded in late 2021 between two 20-year-old female friends (referred to as the New NCCU Conversation in this study). The 2023 conversations were conducted over online video call and transcribed using traditional characters. All participants were between 22 and 25 years old, were friends with their interlocutor, and had been born in- and grown up on Taiwan. The region of Taiwan in which they had grown up, as well as any languages they spoke beyond TM, were also recorded. Of the eight speakers, six were female and two were male. Three of the females came from northern Taiwan, one from central, and two from southern. One of the males came from the north, and one from the south. All participants spoke both TM and English, with the speaker from central Taiwan and one speaker each from the north and south also speaking TSM. French and Japanese were also spoken by one or more participants.

### *4.1.2 Recruitment*

Several recruitment methods were used in this study. First, the friend-of-a-friend method

was used, with the researcher asking various acquaintances with connections on Taiwan to share the recruitment blurb — written in Mandarin using traditional characters — with anyone they knew who fit the participant criteria. The recruitment blurb was also posted on the Language Exchange in Taiwan Facebook page and on National Tsing Hua University's student messaging board. Additionally, an assistant professor at National Tsing Hua University was kind enough to add this study to a list that psychology students could participate in for course credit. Interested people reached out to the researcher via the email included in the recruitment blurb and were provided with explanations of any part of the research procedure they were unclear about. If they were still interested in participating after that, they scheduled a day and time for the recording with the researcher and chose the video conferencing platform they wished to use. Microsoft Teams, Zoom, and Google Meet were all used.

### *4.1.3 Recording*

Video calls were conducted entirely in Mandarin both for the ease of participant understanding and to avoid activating the participants' non-TM language modes. After introductions, the researcher explained the recording procedure and prompted the participants to ask questions about any point they did not understand. Participants were then asked their age, where they grew up, and what languages they speak. Next, the consent form was uploaded to the chat and shared using the screen share function. Participants were instructed to read the form carefully and ask questions about if they did not understand something or wanted more information. When both participants indicated that they had finished reading and had no questions, the researcher began recording the meeting and asked the participants if they agreed to

take part in the study. Once both participants confirmed their willingness, the researcher reminded them to converse for at least 30 minutes about whatever they chose, and if at any point they said something that they did not want included in the transcript, all they had to do was leave a message saying so in the chat. Participants were instructed to either leave the meeting when they were finished, or, if they had a question for the researcher, to simply say they were finished. The researcher then turned off their video and muted their microphone and let the participants converse, but stayed by the computer in case any questions arose in the course of the conversation. Once participants left or said they were finished, the recording was stopped and the recorded file saved.

**4.2 Data Analysis**

The newly-recorded conversations were transcribed in traditional characters with all identifying information (names of people, cities, schools, etc. removed and replaced with "name", "city", etc. as appropriate) removed. Due to the large number of homophones in TM, each conversation was checked by a native speaker to ensure that what was transcribed made sense. Features were identified by using the "find" tool and examined in their sentential and discourse context to determine their syntactic category and/or meaning. Raw frequencies by relevant category for each feature were tabulated in a spreadsheet editing program.

To calculate normalized frequencies, the word count of each conversation had to be determined. Text in Chinese characters is written without spaces, so to obtain an accurate word count, rather than simply a character count, the transcripts were saved as plain text files with no speaker labels or non-word turns (e.g. laughter) and run through the Chinese Knowledge

Information Processing (CKIP) group's word segmentation program. These word counts were then used to calculate the normalized frequency (uses per 1,000 words) of each feature by category, conversation, speaker (in the Old and New corpora), and corpus.

Direct comparisons were done using the online log likelihood wizard tool from Lancaster University (https://ucrel.lancs.ac.uk/llwizard.html). Log likelihood was used to measure statistical significance, and log ratio to measure effect size. Because of its genre and participant differences with respect to the other corpora, the Old Corpus was not included in these calculations. The NCCU Corpus and New Corpus were compared in each feature and category of feature, and the regional data from the New Corpus were also subjected to pairwise comparisons.

# Chapter 5: *Yǒu* 有 +VP

## 5.1 Introduction

Easily the most remarked-upon syntactic feature of Taiwan Mandarin is its use of *yǒu*+VP. While no one contests that this pattern is present due to the influence of Taiwanese Southern Min, claims about what it is used to mean vary widely. This chapter will begin by reviewing the existing analyses of *yǒu*+VP in TM, then present the justification for approaching this construction as an example of replica grammaticalization. Section 5.3 will lay out the predicted differences in *yǒu*+VP use over time and across regions according to the extended Replication Theory framework. Finally, the results of the current study will be presented and their fit with the analyses found in the literature and the predictions of the extended RT approach will be discussed.

## 5.2 Literature Review

Scholars writing about TM in the 1980s had varying interpretations of the function of *yǒu*+VP. Kubler (1981) proposes that *yǒu* served as an auxiliary that could indicate completed action, assert existence or non-existence (in the case of *méiyǒu*+VP, *méi* 沒 being the negator of *yǒu* 有) of a state (including those described by adjectives, which behave as stative verbs in all Chineses), serve as a brief 'yes/no' response to a question (*yǒu* if 'yes', *méiyǒu* if 'no'), or act as potential complement.

(1) Completed action and 'yes/no' response

**A:** 你 有 看 到 她 嗎?
nǐ yǒu kàn dào tā ma
2.SG AUX see RES 3.SG Q
"Did you see her?"

**B:** 我 沒 有 看 到 她
wǒ méi yǒu kàn dào tā
1.SG NEG AUX see RES 3.SG
"I didn't see her."

(adapted from Kubler 1981, p. 87)

(2) Assertion of (non)existence

這 樣 比較 有 漂亮
zhè yàng bǐjiǎo yǒu piàoliàng
this type relatively AUX beautiful
"It looks better like this."

(adapted from Kubler 1981, p. 89)

(3) Potential complement

台灣話 你 聽 有沒有?
táiwānhuà nǐ tīng yǒu-méi-yǒu
Taiwanese 2.SG hear AUX-NEG-AUX
"Do you understand Taiwanese?"

(adapted from Kubler 1981, p. 93)

Cheng (1984) presents a similar analysis, classifying *yǒu* in *yǒu*+VP constructions as an optional auxiliary with multiple functions. For the most part, these correspond with the ones set forth by Kubler (1981), but there are two points of difference. Firstly, Cheng (1984) claims *yǒu* is used to mark the contrast between habitual and future action, which Kubler (1981) does not. Secondly,

Cheng (1984) makes no mention of the use of *yǒu* as a potential complement. However, as Kubler (1981) specifies that this usage was found only among less-educated speakers, it is possible that this discrepancy exists because of sampling differences. Li and Thompson (1981), though they were not focused on describing TM in particular, do make one claim about the nature of *yǒu* +VP in their chapter on negation in Mandarin, saying that speakers of southern Chineses (e.g. TSM) tend to use *yǒu* in A-not-A questions and answers as though it were the perfective aspect marker. While this is understandably a less nuanced and comprehensive analysis than those presented by Kubler (1981) and Cheng (1984), its does attest to the widespread and frequent non-standard use of *yǒu* in Mandarin by speakers from the south of China.

Later works would show more variation in their classification of the role of *yǒu* in *yǒu*+VP. Wang et al. (2007) focus exclusively on the use of *méiyǒu* as a 'no' response, classifying it as a discourse marker due to its use not only as a negative response to a question or statement, but also as a mechanism of self-correction, a means to introduce a correction or clarification, or a socially-appropriate response to praise or thanks. Liu (2011) argues that *yǒu* is a realis mood marker that occurs with four types of verbs: activity, achievement, accomplishment, and stative. In many ways, this claim overlaps with earlier accounts of *yǒu*+VP, as the first three verb types are all ones that may be completed, and stative verbs fall under the 'assertion' use of *yǒu* identified by Kubler (1981) and Cheng (1984). However, Liu's (2011) analysis of *yǒu* +VP as a realis construction means that it cannot co-occur with future tense, imperative mood, or conditional mood, which is not the case for earlier analyses. On the

contrary, one of the specific examples Kubler (1981) cites utilizes *yǒu*+VP in a conditional mood.

(4) Conditional mood *yǒu*+VP

| 買 | 一百 | 塊 | 以上 | 才 | 有 | 送 |
|---|---|---|---|---|---|---|
| mǎi | yìbǎi | kuài | yǐshàng | cái | yǒu | sòng |
| buy | 100 | piece | above | only.then | AUX | deliver |

"Only if you buy 100 dollars worth or more do we deliver."

(adapted from Kubler 1981, p. 90)

While diachronic changes in TM could explain this difference, other more recent accounts also disagree with Liu's (2011) claim about the nature of *yǒu*+VP. Han (2017) treats *yǒu* as an auxiliary verb with the functions identified by Kubler (1981) and Cheng (1984): acting as an affirmative answer to a 'yes/no' question, affirming the existence or occurrence of a quality or event, and denoting habitual action. Ultimately, in contrast to Liu (2011), Han (2017) concludes that *yǒu*+VP "has nothing to do with aspect or tense" (p. 48). This stance is similar to that of Li (2019), who also considers *yǒu* an auxiliary rather than an aspect marker. Li's (2019) proposed functions of *yǒu*, however, are fewer in number and depend on the characteristics of the VP complement. If the VP denotes a non-specific action, then *yǒu* attests to at least one occurrence of that action. However, if the verb denotes a specific action, then *yǒu* attests to the boundedness of that action. While the latter may seem to be identical to that of a perfect or perfective aspect marker, Li (2019) presents evidence that there are in fact subtle semantic differences between *yǒu* and the perfective marker *le* and perfect marker *guo*. This is not to say, however, that all recent accounts disagree with Liu's (2011) analysis of *yǒu*+VP as a realis construction. In fact, Collart and Su (2022) argue that *yǒu*+VP is a marker of assertive modality, a classification that

encompasses the notion of realis mood. They base this analysis on the frequent co-occurrence of *yǒu*+VP (1) in embedded clauses where the matrix verb has a factive meaning, (2) with deictic past time and assertive attitudinal adverbs, (3) with the progressive and experiential aspect markers *le* and *guo*, and (4) with the copula *shì*. Like Liu (2011), Collart and Su (2022) also point to the absence of *yǒu*+VP in imperative sentences and the conditional mood to support their analysis, though, as mentioned before, there is reason to be skeptical of the claim that *yǒu*+VP is never used in the conditional mood.

Ultimately, while newer analyses of the *yǒu*+VP construction have some overlap with early accounts of TM, they generally contradict those accounts on at least one point. Additionally, more recent analyses are also often in direct conflict with each other. This study will therefore sort tokens of *yǒu*+VP based primarily on their interpretation in context, using the broad categories of 'affirmation', 'completion', 'habitual' and 'yes/no' to cover all the proposed meanings of *yǒu*. Within the affirmation, completion, and habitual categories, *yǒu*+VP tokens will be further grouped by the semantic class of their co-occurring verb: stative (including adjectives), communication, perception, action, auxiliary, and modal. In this way, I hope to avoid bias that might come from a premature commitment to one of the existing schemas.

### 5.3 Extended RT Approach

Regardless of what, exactly, the current function of *yǒu*+VP in TM is, there is ample reason to treat it as a case of replica grammaticalization. For a feature to be considered as such, there must be evidence that a grammaticalization process that can be seen or at least inferred in the model language (TSM) has been imitated in the replica language (TM) to create a new

grammatical category or use pattern. In TSM, the cognate to *yǒu*, *ū*, is used not only as a main verb to indicate existence or possession, but also as a perfective aspect marker and an auxiliary verb to express habitual action or affirm the occurrence of an event (Han 2017, Li 2019). *Ū*+VP, therefore, has long been a perfectly grammatical, and in fact quite frequent, construction in TSM. And because *ū* is used in TSM in both its fully lexical and fully grammatical forms, speakers are able to infer the grammaticalization pathway that resulted in the aspectual and auxiliary uses.

Not all of these grammaticalized uses would have existed in MM, however. In the PTH of the mid-20th century, *yǒu*+VP was not grammatical, as *yǒu* functioned only as a main verb that expressed either existence or possession of an object and therefore necessarily took an NP complement. However, *méi(yǒu)*+VP was grammatical as the negative counterpart of the perfective VP-*le*, appearing in both statements and A-not-A questions (Li & Thompson 1981).

(5) *Méi*(*yǒu*)+VP in PTH

| 我 | 沒(有) | 看 | 見 | 你 |
|---|---|---|---|---|
| wǒ | méi(yǒu) | kàn | jiàn | nǐ |
| 1.SG | NEG(AUX) | see | RES | 2.SG |

"I didn't see you."

(adapted from Li & Thompson 1981, p. 417)

| 我 | 寫 | 錯 | 了 | 那 | 個 | 字 | 沒(有)? |
|---|---|---|---|---|---|---|---|
| wǒ | xiě | cuò | le | nà | ge | zì | méi(yǒu) |
| 1.SG | write | wrong | PFV | that | CL | character | NEG(AUX) |

"Did I write that character wrong?"

(adapted from Li & Thompson 1981, p. 433)

Speakers of southern Chineses, Li and Thompson (1981) claim, extended this pattern to affirmative uses, employing *yǒu*+VP when speaking Mandarin to mark the perfective aspect in

questions and answers. Under the influence of such speakers, other Mandarin speakers began to accept *yǒu-méi-yǒu*+VP as a legitimate alternative to VP-*le méiyǒu*, and *yǒu*+VP as a valid affirmative answer thereto. Given that approximately half of the Mainlanders originally hailed from the south of China (Kuo 2005), *yǒu*+VP would have been frequent enough in the feature pool to be selected for MM. In turn, MM would have contributed *yǒu*+VP to the feature pool for TM. With this contribution being reinforced by TSM doing the same at high frequency, it is no wonder that *yǒu*+VP was selected for use in TM. Once selected, the gap in degree of grammaticalization between TSM and MM motivated grammatical replication, and the grammaticalization pathway evident in TSM made replica grammaticalization possible. That it is the grammaticalization process that was copied into TM, rather than the polysemy associated with *ū*, is evidenced by the fact that, while the recent literature does not agree on the exact status of preverbal *yǒu* in TM (itself suggestive that the feature is in the process of changing), it does agree that, unlike *ū* in TSM, *yǒu* in TM is not used as a perfective aspect marker.

### 5.4 Predictions

Replication Theory predicts that for some time after initial contact between MM and TSM, there would be a good deal of individual variation in the use of *yǒu*+VP. It is only after a use patterns has spread within the speaker community and become a stable feature of the new language variety that the process of grammaticalization will begin. For *yǒu*+VP in TM, the results of grammatical acceptability surveys indicate that stability (in the form of general agreement on what constitutes correct usage) was first established among young people in the 1990s (Han 2017). At that point, the expectation is that replicated features, such as *yǒu*+VP, will

be less grammaticalized in the replica language (TM) than in the model language (TSM). For *yǒu*+VP, that would likely manifest as regular auxiliary use with a more restricted class of verbs than *ū* is used with in TSM. *Yǒu* would not be used as a perfective aspect marker, and, given the commentary by Kubler (1981) and Han (2017), *yǒu* used as a potential complement would also be absent, as its correlation with lower levels of education make it unlikely to have spread widely throughout a community that saw massively increased school enrollment in the 1960s and 70s (Mo 2000). Between the NCCU Corpus and New Corpus, the auxiliary uses of *yǒu* are predicted to increase in frequency and to expand the classes of verbs with which they co-occur. Eventually, though the literature indicates that this has not yet happened (and the existence in Mandarin of a competing construction may preclude it from doing so), *yǒu* may grammaticalize into a perfective aspect marker to mirror TSM *ū*.

When Replication Theory is expanded to include feature pool considerations, additional predictions can be made concerning regional differences. The South is expected to have the highest degree of grammaticalization: few Mainlanders settled there, so a greater portion of the feature pool would have consisted of model (TSM) constructions. The Central region, as another area where few Mainlanders settled, is also expected to use *yǒu*+VP in more grammaticalized ways, i.e. more frequently and with more verb classes. Compared to the South and Central regions, the North is expected to show less grammaticalization of the *yǒu*+VP construction.

### 5.5 Results

Constructions were counted as *yǒu*+VP tokens if they were of the form *yǒu*+VP, *yǒu-méi-yǒu*+VP, *méiyǒu*+VP (where the denoted meaning was not perfective), or (*méi*)*yǒu* (if serving as

a response to an earlier piece of dialogue). Because they are often used to affirm or deny the truth of entire propositions, 'yes/no' *yǒu*+VP tokens were not included in the analysis of VP complement verb class.

### *5.5.1 Old Corpus*

The Old Corpus contained 30 tokens of the *yǒu*+VP construction. The most common meaning denoted by *yǒu*+VP was completion, encompassing half of all uses. Other uses were relatively uncommon, with the habitual meaning used only twice: once by the Clerk, and once by the Interviewer when he echoed the Clerk's statement.

Table 2. Old Corpus: Raw frequency of *yǒu*+VP by speaker

| | **Professor** | **Student** | **Clerk** | **Housewife** | **Interviewer** | **Total** |
|---|---|---|---|---|---|---|
| **Assertion** | 1 | 1 | 2 | 1 | 3 | **8** |
| **Completion** | 2 | 1 | 4 | 1 | 7 | **15** |
| **Habitual** | 0 | 0 | 1 | 0 | 1 | **2** |
| **Yes/No** | 0 | 0 | 0 | 2 | 2 | **4** |
| **Unclear** | 0 | 1 | 0 | 0 | 0 | **1** |
| **Total** | **3** | **3** | **7** | **4** | **13** | **30** |

The interview with the Clerk also had the highest normalized frequency for *yǒu*+VP use at 6.02 uses per thousand words, with other interviews falling between 1.56 and 3.76. The overall frequency of *yǒu*+VP in the Old Corpus was 3.26 uses per thousand words.

Table 3. Old Corpus: Normalized frequency of *yǒu*+VP by interview

| | Professor | Student | Clerk | Housewife | Avg. |
|---|---|---|---|---|---|
| **Assertion** | 0.39 | 0.77 | 1.85 | 0.54 | **0.89** |
| **Completion** | 1.17 | 1.16 | 2.32 | 2.15 | **1.70** |
| **Habitual** | 0.00 | 0.00 | 0.93 | 0.00 | **0.23** |
| **Yes/No** | 0.00 | 0.00 | 0.93 | 1.07 | **0.50** |
| **Total** | **1.56** | **1.93** | **6.02** | **3.76** | |

Easily the most frequent verb class to appear in the VP of *yǒu*+VP was action, representing 68.0% of all uses and making up the majority of uses within each meaning category. Perception and communication verbs were also used, though only when *yǒu* carried a completion meaning. There was a single instance of auxiliary *huì* 會 'will', but no modal or stative verbs. In terms of the number of unique verbs used, the action class was again the highest, with 11 different verbs appearing in the *yǒu*+VP construction. The most frequent of these was, *qù* 去 'to go', which was used 4 times.

Table 4. Old Corpus: Verb classes used in the *yǒu*+VP construction

| | Assertion | Completion | Habitual | Total | Percentage |
|---|---|---|---|---|---|
| **Action** | 7 | 8 | 2 | **17** | **68%** |
| **Auxiliary** | 1 | 0 | 0 | **1** | **4%** |
| **Communication** | 0 | 4 | 0 | **4** | **16%** |
| **Modal** | 0 | 0 | 0 | **0** | **0%** |
| **Perception** | 0 | 3 | 0 | **3** | **12%** |
| **State** | 0 | 0 | 0 | **0** | **0%** |
| **Total** | **8** | **15** | **2** | **25** | **100%** |

*5.5.2 NCCU Corpus*

There were 236 uses of *yǒu*+VP in the NCCU Corpus, the majority of which were a 'yes/no' response (94) or carried a completion meaning (89). In all conversations except NCCU #2, in which *yǒu*+VP was used 17 times with an assertion meaning, assertion uses were relatively infrequent. The habitual meaning was also rarely employed, occurring just 7 times in the corpus and not at all in almost half of the conversations.

Table 5. NCCU Corpus: Raw frequency of *yǒu*+VP

| | **NCCU1** | **NCCU2** | **NCCU3** | **NCCU4** | **NCCU5** | **NCCU6** | **NCCU7** | **NCCU8** | **Total** |
|---|---|---|---|---|---|---|---|---|---|
| **Assertion** | 1 | 5 | 17 | 3 | 7 | 1 | 5 | 2 | **41** |
| **Completion** | 2 | 16 | 9 | 12 | 8 | 9 | 16 | 17 | **89** |
| **Habitual** | 0 | 0 | 1 | 1 | 0 | 2 | 2 | 1 | **7** |
| **Yes/No** | 10 | 6 | 10 | 12 | 5 | 16 | 27 | 8 | **94** |
| **Unclear** | 0 | 0 | 2 | 0 | 1 | 1 | 0 | 1 | **5** |
| **Total** | **13** | **27** | **39** | **28** | **21** | **29** | **50** | **29** | **236** |

Normalized frequencies ranged from 3.87 to 12.30, with over half of the conversations falling between 6.37 and 10.57 uses per thousand words.

Table 6. NCCU Corpus: Normalized frequency of *yǒu*+VP

| | **NCCU1** | **NCCU2** | **NCCU3** | **NCCU4** | **NCCU5** | **NCCU6** | **NCCU7** | **NCCU8** | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|
| **Assertion** | 0.30 | 1.40 | 4.61 | 0.57 | 2.29 | 0.22 | 1.23 | 0.70 | **1.42** |
| **Completion** | 0.60 | 4.49 | 2.44 | 2.28 | 2.62 | 1.98 | 3.94 | 5.95 | **3.04** |
| **Habitual** | 0.00 | 0.00 | 0.27 | 0.19 | 0.00 | 0.44 | 0.49 | 0.35 | **0.22** |
| **Yes/No** | 2.98 | 1.68 | 2.71 | 2.28 | 1.64 | 3.52 | 6.64 | 2.80 | **3.03** |
| **Unclear** | 0.00 | 0.00 | 0.54 | 0.00 | 0.33 | 0.22 | 0.00 | 0.35 | **0.18** |
| **Total** | **3.87** | **7.58** | **10.57** | **5.31** | **6.87** | **6.37** | **12.30** | **10.16** | |

No one verb class made up the majority of VPs, though action was the closest at 46.7%. The communication, perception, and stative classes combined accounted for 47.5% of uses, with

modal and auxiliary classes present, but rare. Within the assertion meaning category, action and stative verbs were equally common, accounting for approximately two-thirds of all assertion uses. Modal and perception verbs were used at roughly equal rates, with auxiliary tokens comprising only a single use of *huì* 會 'will'. The most-used verbs were modal *yào* 要 'to want' and action *dài* 戴 'to wear', with 5 uses each. In the completion category, only the action, communication, and perception verb classes were represented, with action making up almost half of the uses, followed by communication at roughly one third and perception at nearly one fifth. The most commonly-used verbs with the completion meaning were *jiǎng* 講 'to say' (14), and *kàn* 看 'to see' (10). All verbs used with the habitual meaning belonged to the action verb class.

Table 7. NCCU Corpus: Verb classes used in the *yǒu*+VP construction

| | **Assertion** | **Completion** | **Habitual** | **Total** | **Percentage** |
|---|---|---|---|---|---|
| **Action** | 14 | 43 | 7 | **64** | **46.7%** |
| **Auxiliary** | 1 | 0 | 0 | **1** | **0.7%** |
| **Communication** | 0 | 29 | 0 | **29** | **21.2%** |
| **Modal** | 7 | 0 | 0 | **7** | **5.1%** |
| **Perception** | 5 | 17 | 0 | **22** | **16.1%** |
| **State** | 14 | 0 | 0 | **14** | **10.2%** |
| **Total** | **41** | **89** | **7** | **137** | **100.0%** |

*5.5.3 New Corpus*

The New Corpus contained 217 *yǒu*+VP tokens, the largest share of which (87) were those which denoted a completion meaning. The assertion and 'yes/no' response categories each accounted for roughly one fourth of the total uses at 61 and 51 tokens, respectively. Habitual uses were rare, but did appear at least once in 4 of the 5 conversations.

Table 8. New Corpus: Raw frequency of *yǒu*+VP

| | Convo. #1 | Convo. #2 | Convo. #3 | Convo. #4 | New NCCU | Total |
|---|---|---|---|---|---|---|
| **Assertion** | 9 | 9 | 16 | 7 | 20 | **61** |
| **Habitual** | 1 | 2 | 2 | 0 | 2 | **7** |
| **Completion** | 13 | 8 | 21 | 12 | 33 | **87** |
| **Yes/No** | 4 | 0 | 11 | 13 | 23 | **51** |
| **Unclear** | 3 | 1 | 2 | 3 | 2 | **11** |
| **Total** | **30** | **20** | **52** | **35** | **80** | **217** |

Normalized frequencies were fairly similar across all conversations, ranging from 3.80 to 8.22 uses per thousand words. The conversations with the highest normalized frequencies — Conversations #3 and #4, and the New NCCU conversation — were the ones which used the 'yes/no' meaning of (*méi*)*yǒu* most often. Only Conversation #2 contained no examples of the 'yes/no' use.

Table 9. New Corpus: Normalized frequency of *yǒu*+VP

| | Convo. #1 | Convo. #2 | Convo. #3 | Convo. #4 | New NCCU | Avg. |
|---|---|---|---|---|---|---|
| **Assertion** | 1.14 | 1.76 | 1.98 | 1.36 | 2.06 | **1.66** |
| **Habitual** | 0.13 | 0.39 | 0.25 | 0.00 | 0.21 | **0.19** |
| **Completion** | 1.65 | 1.57 | 2.59 | 2.33 | 3.39 | **2.30** |
| **Yes/No** | 0.51 | 0.00 | 1.36 | 2.52 | 2.36 | **1.35** |
| **Unclear** | 0.38 | 0.20 | 0.25 | 0.58 | 0.21 | **0.32** |
| **Total** | **3.80** | **3.92** | **6.42** | **6.78** | **8.22** | |

The verb class that appeared most often in the VP of *yǒu*+VP constructions was action, representing 37.9% of all uses. The stative, perception, and communication classes accounted for most of the remaining uses, each occurring at a rate of approximately 18%. Auxiliary and modal verbs were used less frequently overall, occurring at rates of 2.0% and 5.2%, respectively.

Within each meaning category, there was a different distribution pattern with respect to

the verb class of the VP complement. In the assertion meaning, the stative class was the most common, making up 45.9% of all uses, and the vast majority of these stative verbs were adjectives. Assertion also was the only meaning category to contain instances of auxiliary verbs in the VP. Rarest in this category, however, were communication class verbs, with only a single use of *shuō* 說 'to say' attested. By contrast, in the completion category, communication verbs made up 28.2% of all *yǒu*+VP uses, with 4 unique verbs used, the most frequent of which was *jiǎng* 講 'to say'. Almost twice as common within the completion category, however, were action verbs, at 49.4% of uses. Twenty-nine unique action verbs were used, with the most common, *qù* 去 'to go', occurring 9 times. In the habitual meaning category, action and perception verbs were roughly equal both in terms of frequency and variety, with only perception *kǎolǜ* 考慮 'to consider' used more than once.

Table 10. New Corpus: Verb classes used in the *yǒu*+VP construction

| | **Assertion** | **Completion** | **Habitual** | **Total** | **Percentage** |
|---|---|---|---|---|---|
| **Action** | 13 | 42 | 3 | **58** | **37.9%** |
| **Auxiliary** | 3 | 0 | 0 | **3** | **2.0%** |
| **Communication** | 1 | 24 | 0 | **25** | **16.3%** |
| **Modal** | 6 | 2 | 0 | **8** | **5.2%** |
| **Perception** | 10 | 14 | 4 | **28** | **18.3%** |
| **State** | 28 | 3 | 0 | **31** | **20.3%** |
| **Total** | **61** | **85** | **7** | **153** | **100.0%** |

As it did not include information about where the participants grew up, the New NCCU conversation was excluded from the analysis of *yǒu*+VP use by region. Of the remaining 137 *yǒu*+VP tokens, 54 were used to denote the completion meaning. The next most common use

was assertion at 41 tokens, then 'yes/no' at 28. Habitual *yǒu* was present in every region, but most common in the North, which accounted for 3 of the total 5 habitual uses.

Table 11. New Corpus: Raw frequency of *yǒu*+VP by region

| | North | Central | South | Total |
|---|---|---|---|---|
| **Assertion** | 20 | 11 | 10 | **41** |
| **Habitual** | 3 | 1 | 1 | **5** |
| **Completion** | 20 | 14 | 20 | **54** |
| **Yes/No** | 13 | 5 | 10 | **28** |
| **Unclear** | 4 | 2 | 3 | **9** |
| **Total** | **60** | **33** | **44** | **137** |

In terms of normalized frequency, the single speaker from the Central region was something of an outlier, using the *yǒu*+VP construction at a rate of 8.04 uses per thousand words, as opposed to the 4.66 and 4.75 of the North and South, respectively. The bulk of this difference occurred in the assertion and completion categories.

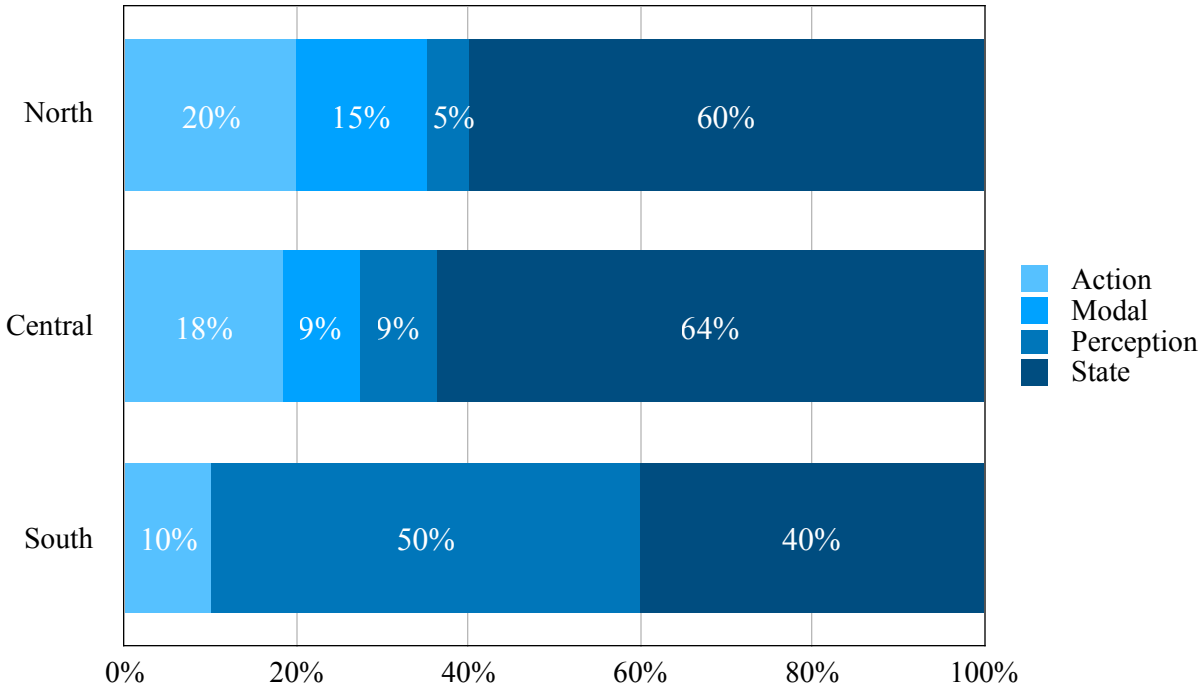
Table 12. New Corpus: Normalized frequency of *yǒu*+VP by region

| | North | Central | South | Avg. |
|---|---|---|---|---|
| **Assertion** | 1.55 | 2.68 | 1.08 | **1.8** |
| **Habitual** | 0.23 | 0.24 | 0.11 | **0.2** |
| **Completion** | 1.55 | 3.41 | 2.16 | **2.4** |
| **Yes/No** | 1.01 | 1.22 | 1.08 | **1.1** |
| **Unclear** | 0.31 | 0.49 | 0.32 | **0.4** |
| **Total** | **4.66** | **8.04** | **4.75** | |

With respect to verb class, the North had the most consistently diverse usage, using 4 verb classes with assertion and completion *yǒu*, and two with habitual. The Central and South regions used either 3 or 4 verb classes with assertion and completion *yǒu*, and only one with habitual. In terms of distribution, the North and Central regions were fairly alike within the

assertion meaning, with stative verbs making up roughly 60% of uses, followed by action verbs at around 20%. The South, by contrast, used perception verbs 50% of the time when *yǒu*+VP carried an assertion meaning and stative verbs only 40% of the time, with action verbs accounting for the remaining 10%. Only the perception class of the Southern speakers saw any single verb used more than twice, with *xǐhuān* 喜歡 'to like' accounting for 4 of the total 5 uses within that category.

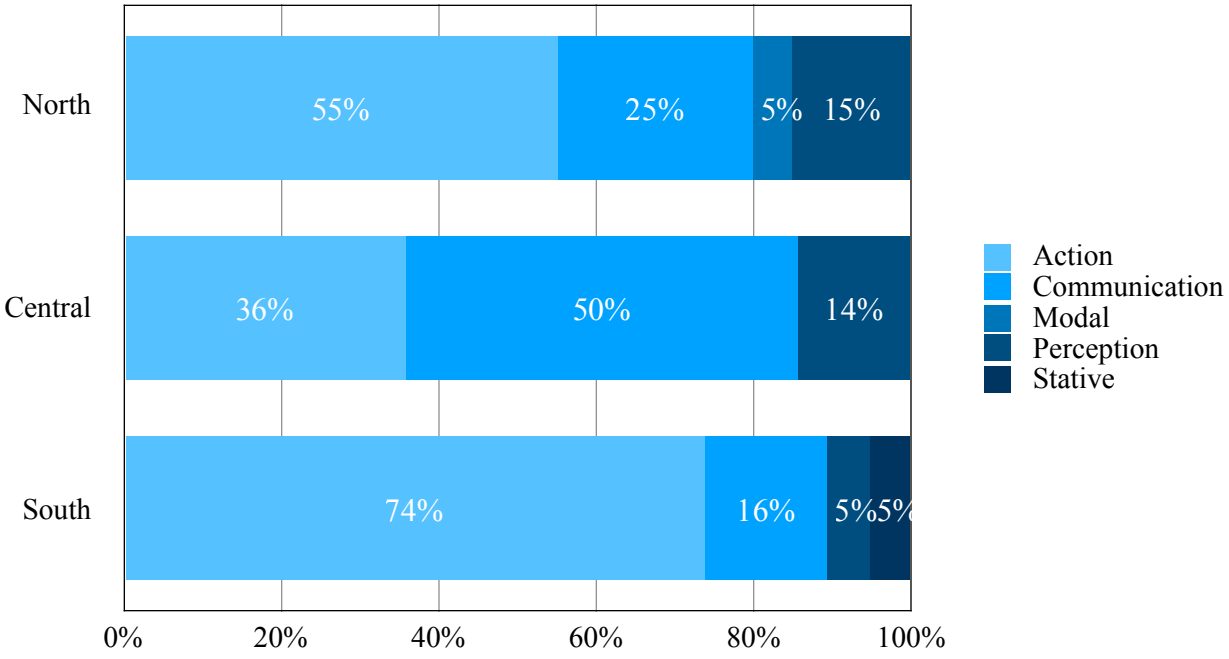
Figure 3. New Corpus: Verb classes used in assertion *yǒu*+VP by region




Within the completion meaning, the North and South patterned similarly, with action verbs making up the majority of uses, followed by communication. The specific action verb used most often differed between the regions, with *gěi* 給 'to give' most common in the North and *qù* 去 'to go' in the South, but both regions used the same communication verb, *jiǎng* 講 'to say' more frequently than any other. This was also one of the most frequently-used

communication verbs in the Central region, where communication verbs occurred in half of the completion *yǒu*+VP constructions.

Figure 4. New Corpus: Verb classes used in completion *yǒu*+VP by region




The North was the only region to use more than one verb class with habitual *yǒu*, using both action *dǎqiú* 打球 'play sports' once and perception *kǎolǜ* 考慮 'to consider' twice. Both the Central and South used an action verb in their singular habitual tokens.

### *5.5.4 Comparison of Corpora*

While the normalized frequency of the *yǒu*+VP construction more than doubled between the Old Corpus and the NCCU corpus, it fell slightly, but significantly, between the NCCU Corpus and the New Corpus (log likelihood 8.74 (critical value 3.84), log ratio -0.41). In fact, all meaning categories except assertion decreased between the NCCU and New corpora. The sharpest decline was in 'yes/no' uses (log ratio -1.13), which was also the only category where

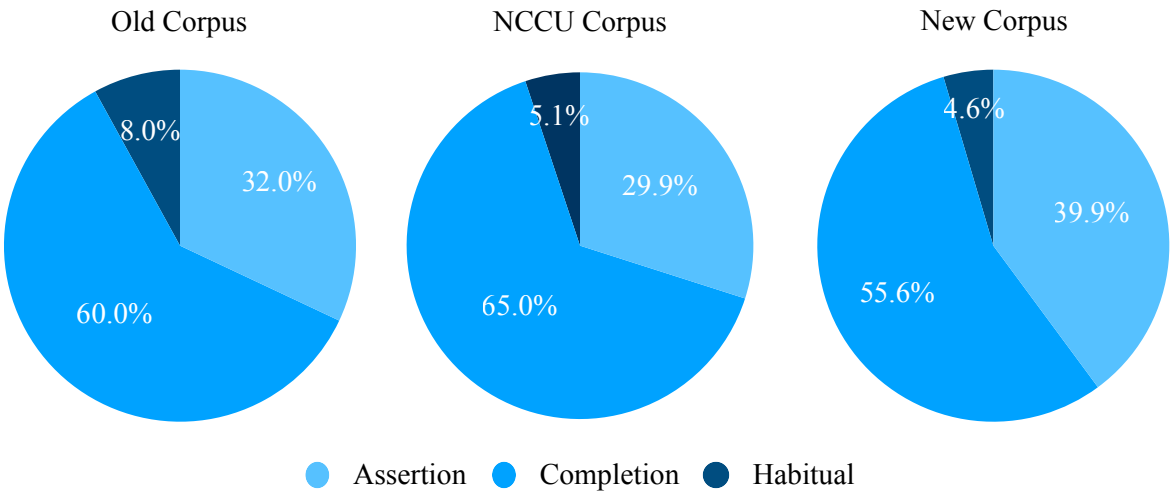
the difference was statistically significant (LL 21.22).

Table 13. Comparison of *yǒu*+VP use in the NCCU Corpus and New Corpus

| | NCCU Corpus | | New Corpus | | Difference | |
|---|---|---|---|---|---|---|
| | *raw freq.* | *per thousand* | *raw freq.* | *per thousand* | *log likelihood* | *log ratio* |
| Assertion | 41 | 1.35 | 61 | 1.70 | 1.30 | 0.33 |
| Habitual | 7 | 0.23 | 7 | 0.19 | 0.10 | -0.24 |
| Completion | 89 | 2.93 | 87 | 2.42 | 1.61 | -0.28 |
| **Yes/No** | 94 | 3.09 | 51 | 1.42 | **21.22** | **-1.13** |
| **Total** | 231 | 7.60 | 206 | 5.73 | **8.74** | **-0.41** |

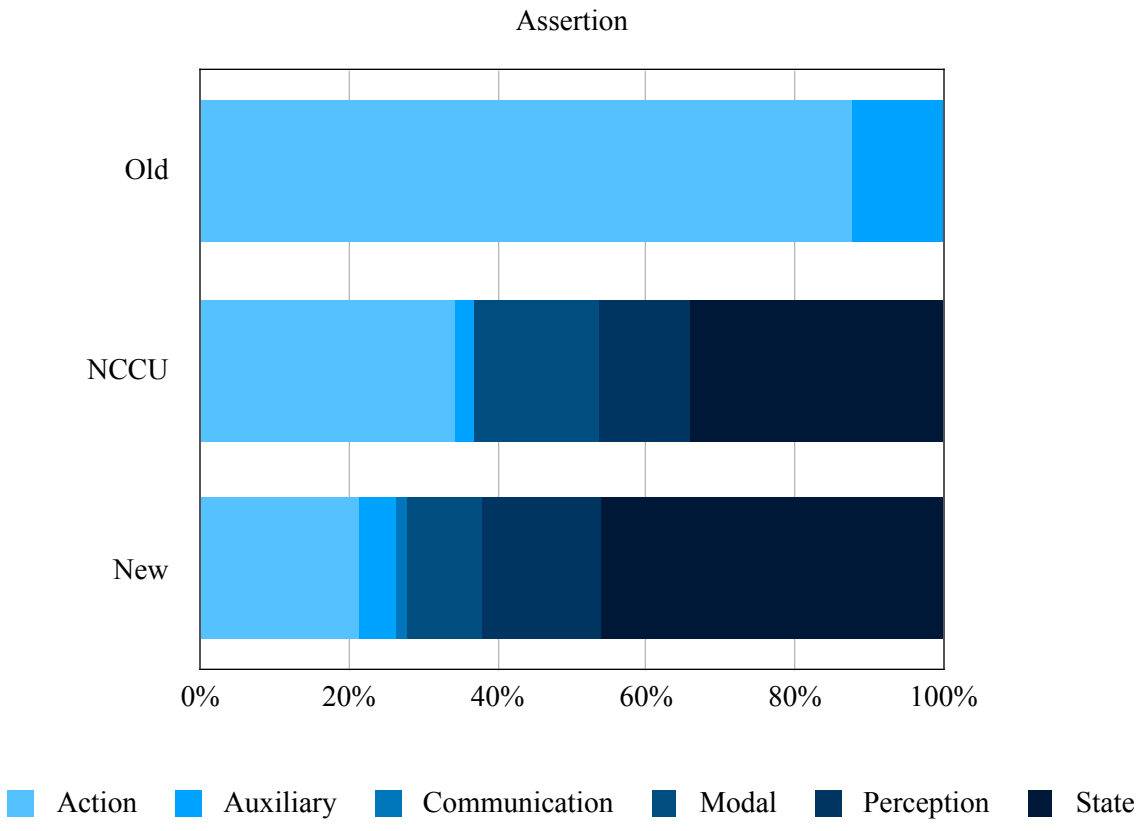
The 'yes/no' meaning category also varied the most widely in use between the three corpora, being relatively infrequent in the Old Corpus at 13.3% of uses, the most common use in the NCCU Corpus at 39.8%, and third most common in the New Corpus at 23.5%. When looking only at instances in which *yǒu*+VP had an overt VP, however, a pattern emerges whereby the completion meaning accounts for over half of all uses, assertion roughly one third, and habitual the remaining 5% or so.

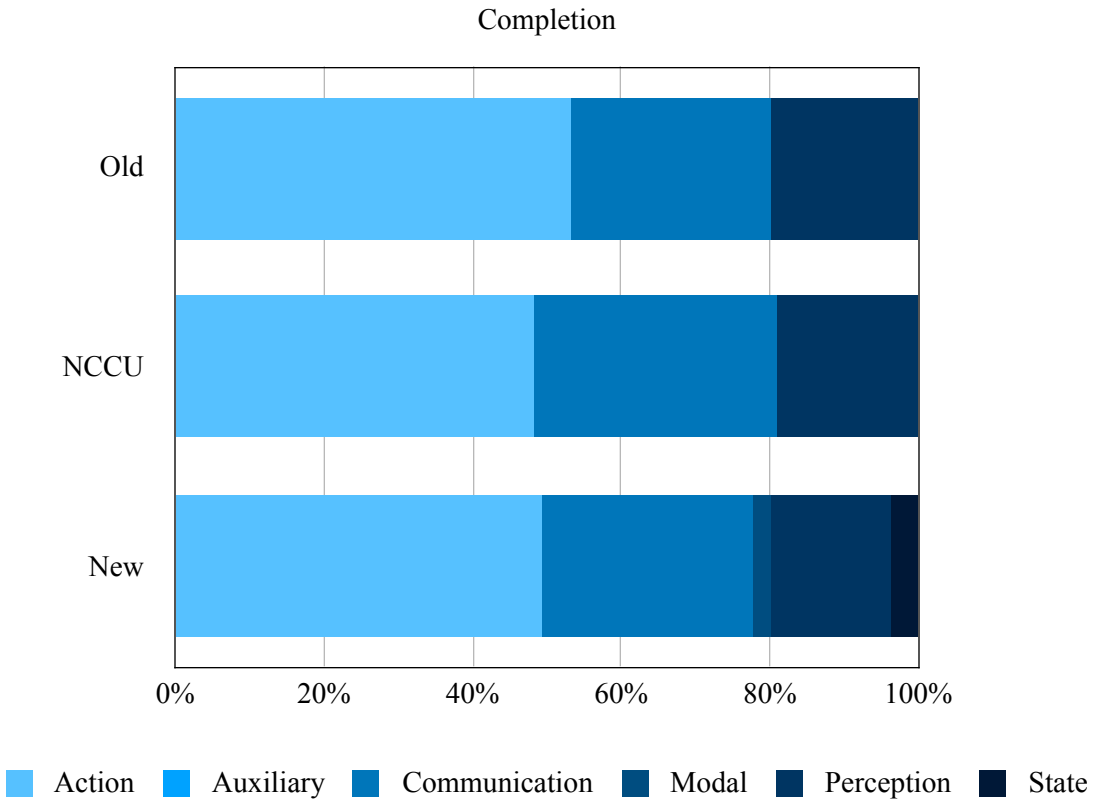
Figure 5. Comparison of *yǒu*+VP use across corpora



Verb classes used within the meaning categories showed a clear trend of ever-increasing variety. The assertion meaning, which only occurred with action and auxiliary verbs in the Old Corpus, expanded to include stative, modal, and perception verbs in the NCCU Corpus and added communication verbs in the New Corpus. It also shifted from being a majority action to a majority stative category. With the completion meaning, both the Kubler and NCCU corpora used action, communication, and perception verbs, with the New Corpus additionally using modal and stative verbs. In the habitual category, the New Corpus was unique in being the only corpus to have two verb classes appear: action and perception.

Figure 6. Verb classes used in *yǒu*+VP across corpora






The regional comparison revealed very little difference — none of it significant — between the North and South in terms of how often *yǒu*+VP was used. Between the Central region and the other two, however, there were several statistically significant differences. In the North-Central comparison, the Central speaker used *yǒu*+VP at significantly higher rates overall (LL 5.71, log ratio 0.80) and to denote the completion meaning (LL 4.76, log ratio 1.13) than did the Northern speakers.

Table 14. Comparison of *yǒu*+VP use between North and Central regions

| | North | | Central | | Difference | |
|---|---|---|---|---|---|---|
| | *raw freq.* | *per thousand* | *raw freq.* | *per thousand* | *log likelihood* | *log ratio* |
| Assertion | 20 | 1.55 | 11 | 2.68 | 1.98 | 0.79 |
| Habitual | 3 | 0.23 | 1 | 0.24 | 0.00 | 0.06 |
| **Completion** | 20 | 1.55 | 14 | 3.41 | **4.76** | **1.13** |
| Yes/No | 13 | 1.01 | 5 | 1.22 | 0.12 | 0.27 |
| **Total** | 56 | 4.35 | 31 | 7.55 | **5.71** | **0.80** |

In the Central-South comparison, the overall use of the Central speaker was again higher to a statistically significant degree (LL 4.88, log ratio -0.77), as was the use of the assertion meaning (LL 4.26, log ratio -1.31).

Table 15. Comparison of *yǒu*+VP use between Central and South regions

| | Central | | South | | Difference | |
|---|---|---|---|---|---|---|
| | *raw freq.* | *per thousand* | *raw freq.* | *per thousand* | *log likelihood* | *log ratio* |
| **Assertion** | 11 | 2.68 | 10 | 1.08 | **4.26** | **-1.31** |
| Habitual | 1 | 0.24 | 1 | 0.11 | 0.32 | -1.18 |
| Completion | 14 | 3.41 | 20 | 2.16 | 1.67 | -0.66 |
| Yes/No | 5 | 1.22 | 10 | 1.08 | 0.05 | -0.18 |
| **Total** | 31 | 7.55 | 41 | 4.42 | **4.88** | **-0.77** |

Ultimately, there was no clear pattern that emerged when comparing regions. Whether in terms of total use, or use to denote any specific meaning, differences between the North and South were neither large nor significant, and differences between the Central and other regions were not consistent.

## 5.6 Discussion

### *5.6.1 Classifying Yǒu+VP*

Broadly, existing claims about the syntactic nature of *yǒu* in *yǒu*+VP align with one of two options: assertive mood marker or auxiliary verb.[1] To be an assertive mood marker, *yǒu*+VP must only occur in contexts where what the speaker is saying is a thing they can know to be true. In other words, it may not co-occur with future tense, appear in imperatives, or be used with conditional mood. But although they are infrequent, there are examples of *yǒu*+VP occurring in future and conditional contexts in the New Corpus.

(6) Future *yǒu*+VP

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **A:** | 你 | 明天 | 有 | 要 | 回 | 家 | 嗎? |
| | nǐ | míngtiān | yǒu | yào | huì | jiā | ma |
| | 2.SG | tomorrow | AUXASSERTION | will | return | home | Q |
| | "Will you return home tomorrow?" | | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| **B:** | 我 | 等下 | 就 | 回 | 家 |
| | wǒ | děngxià | jiù | huì | jiā |
| | 1.SG | later | then | return | home |
| | "I'll return home later." | | | | |

(New NCCU, lines 370-371)

[1] As Wang et al. (2007) focus only on the use of *mèiyǒu* to mean 'no' and not any other manifestations of the *yǒu*+VP construction, their discourse marker classification is set aside here.

(7) Conditional *yǒu*+VP

| 如果 | 那 | 一 | 天 | 的 | 操作 | 沒 | 有 | | 超 |
|---|---|---|---|---|---|---|---|---|---|
| rúguǒ | nà | yī | tiān | de | cāozuò | méi | yǒu | | chāo |
| if | that | one | day | DE | operations | NEG | AUX$_{\text{ASSERTION}}$ | | extremely |

| 累 | 到 | 時間, | 就 | 是 | 我們 | 約 | 四 | 個 | 小時 |
|---|---|---|---|---|---|---|---|---|---|
| lèi | dào | shíjiān | jiù | shì | wǒmen | yuē | sì | ge | xiǎoshí |
| tired | arrive | time | then | is | 1.PL | arrange | four | CL | hour |

"If you are not super tired by the time of that day's operations, we will give you four hours."

(Conversation #3, line 84)

So while there are no examples in the present data of *yǒu*+VP used in imperative statements, there are enough instances of the construction occurring in irrealis contexts to justify discarding the claim that *yǒu* is an assertive mood marker.

For *yǒu* to be considered an auxiliary, it must be shown to contribute a functional meaning to the verb following it, rather than its own lexical meaning. This is clearly the case with *yǒu*+VP, as neither the possessive or existential 'to have' meanings of lexical *yǒu* are accessible in this construction.

(8) Auxiliary *yǒu*+VP

(about a past teacher)

| 他 | 至少 | 有 | 給 | 講義 |
|---|---|---|---|---|
| tā | zhìshǎo | yǒu | gěi | jiǎngyì |
| 3.SG | at least | AUX$_{\text{COMPLETION}}$ | give | handout |

"At least he gave handouts."

(Conversation # 4, line 103)

| 我 | 目前 | 還是 | 有 | 考慮 | 回 | 去 | [City B] |
|---|---|---|---|---|---|---|---|
| wǒ | mùqián | háishì | yǒu | kǎolǜ | huí | qù | |
| 1.SG | currently | still | $AUX_{HABITUAL}$ | consider | return | go | [City B] |

"I'm currently still considering returning to [City B]."

(Conversation #2, line 14)

The present data therefore supports the analysis of *yǒu* in *yǒu*+VP constructions as a multifunctional auxiliary verb.

*5.6.2 Evaluating Predictions*

The predictions for the use of *yǒu*+VP in the Old Corpus were that the construction would show substantial individual variation in its frequency of use, the types of meaning associated with that use, and the verb classes present in the VPs. By the time of the NCCU Corpus, it is expected that a community norm would have formed, meaning speakers would be more similar in how and how often they use *yǒu*+VP. Use would in general be more frequent than in the Old Corpus and more verb classes would be found in the VPs, but the construction would still be more restricted than its counterpart in TSM, not occurring with all verb classes and not used as a perfective marker. It is also unlikely that the potential complement use noted by Kubler (1981) would be present, given its association with a lack of education. Between the NCCU Corpus and the New Corpus, another increase in use was expected, along with a broadening of the verb classes used in the construction. When separated by geographical region, the New Corpus data was predicted to show speakers from the North using *yǒu*+VP less and with fewer verb classes than those from the South and Central regions.

The majority of these predictions regarding inter-corpora behavior were strongly supported by the data. The Old Corpus did see a great deal of individual variation in terms of frequency of use — the Clerk used *yǒu*+VP 3.86 times more often than the Professor — and while all speakers used *yǒu*+VP most often to communicate an assertion or completion meaning, three of the five speakers demonstrated unique distribution patterns otherwise. The one area in which there was little individual variation was verb classes, since action verbs comprised the clear majority in every meaning category. These facts align with what would be expected under extended RT.

Comparing these results to the NCCU Corpus, there was, as predicted, an increase in the use of *yǒu*+VP — normalized frequency grew from 3.26 to 7.60 — as well as in its regularity of use — NCCU 7 used *yǒu*+VP only 3.18 times more often than NCCU 1. Additionally, speaker behavior in terms of the meaning categories patterned more regularly than in the Old Corpus. All speakers in the NCCU Corpus used *yǒu*+VP in 'yes/no' contexts, and over half in habitual contexts as well. In terms of the verb classes used in the VP, the predicted increase was slight, but present. The assertion meaning added modal, perception, and state verbs, and the completion meaning saw an increased percentage of communication verbs used. Also accurately predicted by the extended RT framework was the lack of *yǒu*+VP tokens denoting perfective aspect or acting as potential complements.

When evaluating the New Corpus with respect to the NCCU Corpus, as predicted, both regularity of use — New NCCU using *yǒu*+VP only 2.16 times more often than Conversation #1 — and variety of verb classes used within meaning categories increased. The assertion category contained instances of every verb class, the completion category expanded to include modal and

stative verbs, and the habitual meaning was used for the first time with a non-action verb. In terms of overall frequency, however, *yǒu*+VP declined between the NCCU Corpus and the New Corpus to a small but significant degree. While the bulk of this difference came from a drop in the use of *yǒu*+VP in 'yes/no' contexts, two out of the other three categories also saw a decrease, albeit not by a statistically significant amount. Although the New Corpus still used *yǒu*+VP more frequently than the Old Corpus (6.03 uses per thousand words compared to 3.26), this decline in use from the levels seen in the NCCU Corpus is unexpected. However, a potential explanation for this can be found in the regional data.

The predictions for the regional comparison were that the North would be more MM-like in its behavior than the other regions, using *yǒu*+VP less frequently overall and with fewer verb classes. However, when compared to the South, which was expected to use *yǒu*+VP most frequently and with the most verb classes, the North actually had very similar rate of use across all meaning categories/use patterns. This suggests that dialect leveling has occurred for this feature, which is unsurprising considering how aware speakers are of it as unique to TM (it is the one non-phonological difference between TM and the Mandarin spoken on mainland China that is referenced on websites for Mandarin language learning programs in Taiwan). Reference to dialect leveling can also explain why there was a slight decrease in use between the NCCU Corpus and New Corpus: when speakers of different dialects attempt to converge their speaking styles, they use the distinctive features of their dialects less frequently (Trudgill 1986). Conceivably, for speakers from the South, this would have involved using *yǒu*+VP less often. And while the Central data, which showed both the highest frequency of use and the most variety of verb classes in the VP complement, might seem to contradict this conclusion, closer

examination of the feature pool for that region offers an alternative explanation for this aberration. Although in this study Hakka has not been considered a major contributor to TM due to its relative rarity among the population, responses to the 2010 population and housing census showed that, in the northern part of Central Taiwan, over half of the population used Hakka at home (DGBAS 2010). And while Hakka grammar is distinct from TSM in many ways, when it comes to *yǒu*+VP, there is a substantial degree of overlap. Like all Chineses, Hakka can use *yu*$^{1}$ (cognate of *yǒu*) to indicate possession or existence, and in common with TSM, *yu*$^{1}$ can also be used as an auxiliary to indicate completion of the complement VP, which can contain a verb of any class (Hashimoto 1973; Chappell & Lamarre 2005). Consequently, in the feature pool for the Central region, use of the *yǒu*+VP construction would have been reinforced by virtue of being present in multiple contributing varieties (see Cheshire et al. 2011 in D'Arcy 2020) in a way that it was not in the North or South. That the TM of the Central region had the most frequent and varied use of *yǒu*+VP is therefore still congruent with the extended RT approach.

### 5.7 Conclusion

In this chapter, *yǒu*+VP was presented as an example of replica grammaticalization in Taiwan Mandarin. The literature on this construction was surveyed, and the two leading claims about the syntactic category of *yǒu* — auxiliary verb or affirmative mood marker — were evaluated. Based on the corpora analyzed in this work, it was determined that *yǒu*+VP is an auxiliary verb construction used to convey notions of assertion, completion, or habitual action, and to act as a 'yes/no' response to a question or statement. *Yǒu* as an auxiliary was shown to have become more regular (i.e., more consistent between speakers in its rate of use) over time

and to have increased in the number of verb classes it co-occurs with across all meaning categories. While some details of diachronic frequency changes and regional use differences appeared to contradict the predictions of extended RT, consideration of other contact phenomena and a more detailed examination of the contributors to regional feature pools showed otherwise.

# Chapter 6: Complementizer *Shuō* 說

## 6.1 Introduction

While a less well-known characteristic of Taiwan Mandarin than *yǒu*+VP, *shuō* used as a complementizer is still a noteworthy feature of the dialect. This chapter will begin by reviewing the existing work on complementizer *shuō* in TM and presenting the justification for treating it as an example of replica grammaticalization. Then, the results of each of the corpora individually, comparisons between the corpora, and comparisons of speakers from different regions of Taiwan will be given. Finally, the fit of these results with the extended Replication Theory used in this work will be discussed.

## 6.2 Literature Review

Early accounts of the non-lexical use of *shuō* in TM proposed a range of analyses, with Kubler and Ho (1984, p. 43) dismissing it as a "filler that gives the speaker time to think about what he will say next" while Cheng (1985) describes it as clause-initial complementizer. Over time, the idea that *shuō* can act as a complementizer has become the accepted interpretation, though there is still disagreement over the precise contexts in which it can do so. S. Huang (2003), for instance, claims that *shuō* acts as a complementizer to introduce *de dicto* complements only, while J. Huang (2021) claims instead that it introduces non-referential CPs. Chappell (2008, 2015) argues that there is a grammaticalization cline for 'verbs of saying' into complementizers, and that *shuō* is at different positions on that cline in different Sinitic languages and dialects. Of the proposed explanations for the use of *shuō* as a complementizer,

this is the only one that takes into account both a synchronic and diachronic perspective. Therefore, this is the classification system through which the present data will be analyzed to evaluate the fit of the extended RT framework.

Chappell's proposed grammaticalization cline for 'verbs of saying' in Sinitic is comprised of five stages. Stage 1 is represented by the quotative construction, in which the SAY verb introduces either direct or indirect discourse.

$(NP_{Subject})(PP_{Addressee})\ V_{Quotative} : [QUOTATION]$

(Chappell 2008)

Notably, at this stage, *shuō* is not yet acting as a complementizer, since it is the only verb in the main clause. However, this use does begin the extension process that is the first step in grammaticalization, leading to Stage 2, in which *shuō* is still used in a quotative fashion, but in this case as the $V_2$ in a serial verb construction that has another communication verb (e.g. 'tell', 'ask', 'write') in the $V_1$ position.

$(NP_{Subject})\ V_1\ (X)\ V_{2[Semi\text{-}complementizer]} : [QUOTATION]$

(Chappell 2008)

*Shuō* in this construction is interpreted as a linking device that introduces the subsequent clause, but still retains a sense of its lexical meaning. It is thus categorized as a semi-complementizer, rather than a full complementizer. In Stage 3, the class of possible $V_1$ verbs is extended to include verbs of cognition/perception and *shuō* can thus be considered a full complementizer.

$(NP_{Subject})\ V_1\ COMPLEMENTIZER_{(SAY)} : [CLAUSE]$

(modified from Chappell 2015)

Used this way, *shuō* is both desemanticized, as the "say" interpretation is no longer available, and decategorialized, as it is no longer able to take aspectual markers. This syntactic structure remains the same in Stage 4 or 5; the difference is solely in the verb classes that may fill the $V_1$ position. In Stage 4, emotion and stative verbs become permissible, and in Stage 5, modal verbs and even the lexical "say" may co-occur with the SAY complementizer.

### 6.3 Extended RT approach

Grammaticalization of *shuō* into a complementizer is not uncommon among Chineses, having occurred to some extent in at least half of the Chinese language families. It is therefore well-attested as an internally-motivated process. However, as mentioned previously, internally-motivated processes can still be considered contact-induced if there is evidence that they have been sped up by contact with a language that is further along in that process. This is case for *shuō* in TM. At the time of contact between MM and TSM, *shuō* was only acceptable as a fully lexical verb in PTH, in other words, Stage 0 on the grammaticalization cline (Chappell 2015). However, given that Li and Thompson (1981) indicate that *shuō* was used as a Stage 1 quotative in informal Mandarin, and that roughly half of the Mainlanders came from provinces where at least one of the local Chineses used *shuō* at a Stage 2 level or higher (Kuo 2005, Chappell 2015), we can reasonably conclude that *shuō* in MM (replica language) would likely have been used as a Stage 1 quotative and Stage 2 semi-complementizer. By contrast, TSM (model language) has long used its functional equivalent to *shuō*, *kóng*, as a complementizer that could "take any statement" (Cheng 1985, p. 366), thus putting it at Stage 5. Important to note here is that while *kóng* in TSM is used similarly to *shuō* in Mandarin, its actual cognate is *jiǎng* 講, another verb

meaning ‘to say’. That *shuō* grammaticalized in TM, rather than *jiǎng*, would therefore be surprising under a borrowing paradigm, but is expected under RT, as its notion of equivalence is much more concerned with patterns of use than it is word-for-word translation.

(1) *Kóng* used as a Stage 5 complementizer in TSM

| 恁 | 厝叔仔 | 共 | 我 | 講 | 講， | 我 |
|---|---|---|---|---|---|---|
| Lín | ban-chek-à | kah | goá | kóng | kóng, | goá |
| 2SG.PL | youngest.uncle | COM | 1SG | say | $\text{COMP}_{\text{SAY}}$ | 1.SG |

| 還也 | 有 | 做 | 善事 | 啦. |
|---|---|---|---|---|
| oân-á | ū | chò | siān-sū | là. |
| also | PFV | do | good deed | SFP |

“Your youngest uncle told me that I had also done some good deeds.”

(*Fate* 77-78 in Chappell 2015)

This disparity in the degree of grammaticalization between the two languages means that the conditions for grammatical replication were in place for *shuō* when contact was made. Additionally, the fact that *kóng* could be used both as lexical verb and complementizer in the model language (TSM) means that the grammatical replication could be of the replica grammaticalization type. That this is indeed the case and non-lexical *shuō* in TM is not the result of polysemy copying is evidenced by two facts. First, TM complementizer *shuō* has not yet reached the Stage 5 level found in TSM. Second, there is a grammaticalized use of *shuō* in TM that do not exist in TSM. In their study on informal written and spoken TM, Wang et al. (2003) found that in addition to its use as a sentence-medial complementizer, *shuō* could be used as a sentence-initial marker of hearsay, sentence-final marker of counter-expectation, or sentence-final intensifier. While the sentence-initial hearsay marker and sentence-final counter-

expectation marker uses are also found in TSM, the sentence-final intensifier use is unique to TM.

(2) Hearsay marker *shuō*

**A:** 最近 我 常 熬夜 耶，痘痘 都
zuìjìn wǒ cháng áoyè yé dòudòu dōu
recently 1.SG often burn the midnight oil SFP acne all
冒 出來 了，怎麼辦 呢?
mào chūlái le zěnmebàn ne
appear come PFV how to do Q
"Recently, I have often been burning the midnight oil. Acne has appeared on my face. What should I do?"

**B:** SK-II 啊
SK-II a
SK-II SFP
"SK-II!"

**A:** SK-II， 說 每天 只 睡 一 個 小時，
SK-II shuō měitiān zhǐ shuì yī gè xiǎoshí
SK-II SAY every day only sleep one CL hour
你 相信 嗎?
nǐ xiāngxìn ma
2.SG believe Q
"SK-II. It is said that you only need to sleep for one hour with SK-II; do you believe it?"

(3) Counter-expectation marker *shuō*

| 我 | 怎麼 | 沒 | 注意 | 過 | 她 | 有 | 虎牙 | 說 |
|---|---|---|---|---|---|---|---|---|
| wǒ | zěnme | méi | zhùyì | guò | tā | yǒu | hǔyá | shuō |
| 1.SG | why | NEG | notice | EXP | 3.SG | have | fang | SAY |

"Why didn't I notice that she has a sharp tooth?"

(4) Intensifier *shuō*

*(B tells A that a teacher teaches very well. After hearing B's utterance, A, who is not taking the teacher's course, expresses her regret and desire to audit the course.)*

| **B:** | …而且 | 他 | 上 | 的 | 真的 | 很 | 有 | 內容。 |
|---|---|---|---|---|---|---|---|---|
| | érqiě | tā | shàng | de | zhēnde | hěn | yǒu | nèiróng |
| | and | 3.SG | teach | DE | really | very | have | content |

"And his teaching is very informative."

| **A:** | …其實 | 我 | 好 | 想 | 去 | 旁聽 | 說。 |
|---|---|---|---|---|---|---|---|
| | qíshí | wǒ | hǎo | xiǎng | qù | pángtīng | shuō |
| | actually | 1.SG | very much | want | go | observe | SAY |

"Actually, I want to be an auditor in his class."

(adapted from Wang et. al 2003)

For these reasons, *shuō* in TM should be treated as an example of replica grammaticalization rather than polysemy copying.

### 6.4 Predictions

For the first few decades after contact, use of *shuō* is expected to see significant individual variation, with some speakers hardly using it at all, or only at a Stage 0 or 1 level, while others use it frequently as a Stage 3 or 4 complementizer or discourse marker. Once TM has been established as a distinct dialect, however, we would expect to see *shuō* used more often

and more consistently as a complementizer over time. Initially, this would appear in the NCCU Corpus as an increase in Stage 1 and 2 tokens, with the occasional Stage 3 token. As successive generations grow up speaking TM as the community language, Stage 3 tokens will become more common, with Stage 4 tokens starting to appear more frequently. *Shuō* may also be used more often as a discourse marker during this stage of development, and fully lexical (i.e. Stage 0) used of *shuō* will gradually become less frequent. This is the state of affairs expected in the New Corpus.

With respect to regional variation in the New Corpus, because grammaticalized *shuō* is a less salient feature of TM than *yǒu*+VP, it is not predicted to have undergone the same dialect leveling process. The North and South are therefore predicted to show different patterns of use, with the North expected to see the bulk of its *shuō* tokens in Stages 0-2 and the South showing a greater tendency to use *shuō* as a full complementizer and discourse marker. Because Hakka, like MM, does not use its 'say' verb, as a full complementizer, but only as a Stage 2 semi-complementizer, it is expected that the Central region will follow the distribution pattern of the North, but with a higher frequency of use.

### 6.5 Results

*Shuō* tokens that occurred in contexts outside the scope of the verb → complementizer grammaticalization process, such as in a set adverbial phrase (e.g. *duì X lái shuō* 對X來說 'as for X') or compound word (e.g. *xiǎoshuō* 小說 'novel'), were excluded from analysis.

*6.5.1 Old Corpus*

The Old Corpus contained a total of 68 *shuō* tokens. Among these, only one use of *shuō* as a discourse marker was recorded, with the remaining 67 tokens falling on Chappell's grammaticalization cline. Over half of all *shuō* tokens were Stage 0 lexical verbs (39), with just under a third occurring as Stage 1 quotatives (19). The remaining tokens were spread across Stages 2 through 4. In terms of individual speakers, the Student used the greatest number of *shuō* tokens at 24, and the Clerk the fewest at only 1. The Clerk was also the only interviewee to not use *shuō* as a complementizer.

Table 16. Old Corpus: Raw frequency of *shuō* by speaker

| | **Professor** | **Student** | **Clerk** | **Housewife** | **Interviewer** | **Total** |
|---|---|---|---|---|---|---|
| **Stage 0** | 8 | 15 | 0 | 5 | 11 | **39** |
| **Stage 1** | 2 | 7 | 1 | 4 | 5 | **19** |
| **Stage 2** | 0 | 0 | 0 | 0 | 1 | **1** |
| **Stage 3** | 0 | 2 | 0 | 1 | 0 | **3** |
| **Stage 4** | 4 | 0 | 0 | 0 | 1 | **5** |
| **Discourse** | 1 | 0 | 0 | 0 | 0 | **1** |
| **Total** | **15** | **24** | **1** | **10** | **18** | **68** |

In terms of normalized frequency, no pattern of use held true for all interviews. The overall normalized frequency ranged from 7.00 to 11.56 uses per thousand words, with *shuō* occurring most frequently as a Stage 0 lexical verb in all interviews except 'the Clerk', where it was used only as a Stage 1 quotative. The most grammaticalized uses of *shuō* occurred in the interview with 'the Professor', in which Stage 4 *shuō* was used at a rate of 1.56 uses per thousand words and *shuō* was used once as a discourse marker.

Table 17. Old Corpus: Normalized frequency of *shuō* by interview

| | **Professor** | **Student** | **Clerk** | **Housewife** | **Avg.** |
|---|---|---|---|---|---|
| **Stage 0** | 3.89 | 7.32 | 0.00 | 5.36 | **4.14** |
| **Stage 1** | 0.78 | 3.47 | 0.94 | 3.22 | **2.10** |
| **Stage 2** | 0.39 | 0.00 | 0.00 | 0.00 | **0.10** |
| **Stage 3** | 0.00 | 0.77 | 0.00 | 0.54 | **0.33** |
| **Stage 4** | 1.56 | 0.00 | 0.00 | 0.54 | **0.52** |
| **Discourse** | 0.39 | 0.00 | 0.00 | 0.00 | **0.10** |
| **Total** | **7.00** | **11.56** | **0.94** | **9.66** | |

*6.5.2 NCCU Corpus*

The NCCU Corpus contained 313 *shuō* tokens, over half of which were Stage 1 quotatives (177), followed by Stage 0 lexical verbs at just over a quarter (88). At least one Stage 2 semi-complementizer *shuō* was used in every conversation, with a corpus total of 39 tokens. Stage 3/4 full complementizer uses were much less common, with only 8 in the whole corpus. One discourse marker use of *shuō* was recorded. Individual conversations ranged from having 19 to 81 *shuō* tokens, although the majority had between 25 and 45. In all conversations, *shuō* was used most often at Stage 1. Stage 2 *shuō* was used 2 or 3 times in most conversations, with one conversation having 6 such tokens and two conversations having 10 each. Three of the eight conversations used a Stage 3/4 *shuō* at least once, with one conversation using *shuō* as a complementizer 4 times.

Table 18. NCCU Corpus: Raw frequency of *shuō*

| | NCCU1 | NCCU2 | NCCU3 | NCCU4 | NCCU5 | NCCU6 | NCCU7 | NCCU8 | Total |
|---|---|---|---|---|---|---|---|---|---|
| **Stage 0** | 5 | 11 | 10 | 7 | 9 | 24 | 5 | 17 | **88** |
| **Stage 1** | 11 | 13 | 15 | 42 | 15 | 44 | 16 | 21 | **177** |
| **Stage 2** | 3 | 3 | 2 | 2 | 3 | 10 | 10 | 6 | **39** |
| **Stage 3** | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 0 | **5** |
| **Stage 4** | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | **3** |
| **Discourse** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | **1** |
| **Total** | **19** | **29** | **27** | **51** | **27** | **81** | **35** | **44** | **313** |

Normalized frequencies saw a similar distribution, with most conversations using *shuō* between 5 and 10 times per thousand words. Two conversations saw more frequent use at 15.41 and 17.80 uses per thousand words. Stage 1 was again the most frequent type of *shuō* at an average of 5.60 uses per thousand words, followed by Stage 0 at an average of 3.00. Stage 2 *shuō* averaged 1.30 uses per thousand, with individual conversations using it between 0.38 and 2.46 times per thousand words. No conversation used Stage 3/4 *shuō* more than 0.74 times per thousand words.

Table 19. NCCU Corpus: Normalized frequency of *shuō*

| | NCCU1 | NCCU2 | NCCU3 | NCCU4 | NCCU5 | NCCU6 | NCCU7 | NCCU8 | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| **Stage 0** | 1.49 | 3.09 | 2.71 | 1.33 | 2.95 | 5.27 | 1.23 | 5.95 | **3.00** |
| **Stage 1** | 3.28 | 3.65 | 4.07 | 7.97 | 4.91 | 9.67 | 3.94 | 7.36 | **5.60** |
| **Stage 2** | 0.89 | 0.84 | 0.54 | 0.38 | 0.98 | 2.20 | 2.46 | 2.10 | **1.30** |
| **Stage 3** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.44 | 0.74 | 0.00 | **0.15** |
| **Stage 4** | 0.00 | 0.28 | 0.00 | 0.00 | 0.00 | 0.22 | 0.25 | 0.00 | **0.09** |
| **Stage 5** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.00** |
| **Discourse** | 0.00 | 0.28 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.04** |
| **Total** | **5.66** | **8.14** | **7.32** | **9.68** | **8.84** | **17.80** | **8.61** | **15.41** | |

*6.5.3 New Corpus*

Of the 403 *shuō* tokens found in the New Corpus, nearly 70% occurred in two conversations: the New NCCU conversation (118) and Conversation #3 (161). The remaining three conversations had fairly similar numbers of *shuō* tokens, ranging from 38 to 47. Patterns of use, however, were similar across all conversations, with Stage 1 uses being the most common, followed by Stage 0 and Stage 2 in that order. All conversations also used a Stage 3 complementizer at least once. Two conversations used a Stage 4 complementizer multiple times, but only one conversation used *shuō* as a discourse marker.

Table 20. New Corpus: Raw frequency of *shuō*

| | **Convo. #1** | **Convo. #2** | **Convo. #3** | **Convo. #4** | **New NCCU** | **Total** |
|---|---|---|---|---|---|---|
| **Stage 0** | 4 | 15 | 26 | 17 | 26 | **88** |
| **Stage 1** | 28 | 20 | 123 | 23 | 68 | **262** |
| **Stage 2** | 3 | 3 | 10 | 4 | 20 | **40** |
| **Stage 3** | 2 | 1 | 2 | 1 | 1 | **7** |
| **Stage 4** | 0 | 0 | 0 | 2 | 3 | **5** |
| **Discourse** | 1 | 0 | 0 | 0 | 0 | **1** |
| **Total** | **38** | **39** | **161** | **47** | **118** | **403** |

The normalized frequencies followed much the same pattern, although the gap between conversations was less pronounced than with the raw frequencies. Conversation #4 had the highest normalized frequency at 22.87 uses per thousand words, followed by Conversation #3 at 19.87 and the New NCCU conversation at 16.34. Conversation #1 and #2 had the lowest normalized frequencies at 4.82 and 7.64, respectively. Stage 1 remained the most common use of *shuō*, with an average frequency of 8.59 uses per thousand words. Stage 0 was the next most common use with an average of 2.87, followed by Stage 2 at 1.63. No other stage had an

average normalized frequency above 0.20 per thousand words, though Stage 3 *shuō* had a frequency of 0.25 in two conversations, and in the two conversations in which it occurred, Stage 4 *shuō* had frequencies of 0.31 and 0.58 per thousand words.

Table 21. New Corpus: Normalized frequency of *shuō*

| | **Convo. #1** | **Convo. #2** | **Convo. #3** | **Convo. #4** | **New NCCU** | **Avg.** |
|---|---|---|---|---|---|---|
| **Stage 0** | 0.51 | 2.94 | 3.21 | 5.04 | 2.67 | **2.87** |
| **Stage 1** | 3.55 | 3.92 | 15.18 | 13.18 | 6.99 | **8.56** |
| **Stage 2** | 0.38 | 0.59 | 1.23 | 3.88 | 2.06 | **1.63** |
| **Stage 3** | 0.25 | 0.20 | 0.25 | 0.19 | 0.10 | **0.20** |
| **Stage 4** | 0.00 | 0.00 | 0.00 | 0.58 | 0.31 | **0.18** |
| **Discourse** | 0.13 | 0.00 | 0.00 | 0.00 | 0.00 | **0.03** |
| **Total** | **4.82** | **7.64** | **19.87** | **22.87** | **16.34** | |

As it did not include information about where the participants grew up, the New NCCU conversation was excluded from the analysis of *shuō* use by region. This left 285 *shuō* tokens for analysis. Nearly half of these tokens were Stage 1 as produced by speakers from the North (131). That value was something of an outlier in terms of raw frequency, as the most *shuō* tokens produced in any other category was 32, but its normalized frequency of 10.17 was no further removed from the value for the Central region (7.55) than that value was removed from that of the South (3.45). Whether in terms of raw or normalized frequency, all regions displayed the same usage pattern, with Stage 1 being the most frequent, followed by Stage 0, Stage 2, and Stage 3/4, in that order. The South generally had the lowest rate of usage, with Stage 2 and discourse marker *shuō* being the only exceptions.

Table 22. New Corpus: Normalized frequency of *shuō* by region

| | North | Central | South | Avg. |
|---|---|---|---|---|
| **Stage 0** | 2.49 | 5.60 | 0.76 | **2.95** |
| **Stage 1** | 10.17 | 7.55 | 3.45 | **7.06** |
| **Stage 2** | 0.39 | 2.44 | 0.54 | **1.12** |
| **Stage 3** | 0.23 | 0.24 | 0.22 | **0.23** |
| **Stage 4** | 0.16 | 0.00 | 0.00 | **0.05** |
| **Discourse** | 0.00 | 0.00 | 0.11 | **0.04** |
| **Total** | **13.44** | **15.84** | **5.07** | |

*6.5.4 Comparison of Corpora*

Several trends emerge when looking at the overall use of *shuō* across the three corpora. First, while the normalized frequency of *shuō* increased from 7.4 uses per thousand words in the Old Corpus to 10.3 in the NCCU Corpus and 11.2 in the New Corpus, the proportion of that use that was comprised of Stage 0 tokens decreased from 57.4% to 28.1% to 21.8%. Accompanying this was a steady increase in the relative use of Stage 1 *shuō*, growing from 27.9% in the Old Corpus to 56.7% in the NCCU Corpus and 65.0% in the New Corpus.

Between the NCCU Corpus and the New Corpus, changes were slight and usually not significant. The one difference that did reach the level of statistical significance was the rise in Stage 1 quotative use (LL 5.35, log ratio 0.32). Small, but non-significant increases were also seen in the rate of Stage 3/4 full complementizer use.

Table 23. Comparison of *shuō* use between the NCCU Corpus and New Corpus

| | NCCU Corpus | | New Corpus | | Difference | |
|---|---|---|---|---|---|---|
| | *raw freq.* | *per thousand* | *raw freq.* | *per thousand* | *log likelihood* | *log ratio* |
| Stage 0 | 88 | 2.89 | 88 | 2.45 | 1.25 | -0.24 |
| **Stage 1** | **177** | **5.82** | **262** | **7.28** | **5.35** | **0.32** |
| Stage 2 | 39 | 1.28 | 40 | 1.11 | 0.40 | -0.21 |
| Stage 3 | 5 | 0.16 | 7 | 0.19 | 0.08 | 0.25 |
| Stage 4 | 3 | 0.10 | 5 | 0.14 | 0.23 | 0.49 |
| Discourse | 1 | 0.03 | 1 | 0.03 | 0.01 | -0.24 |
| Total | 313 | 10.30 | 403 | 11.20 | 1.25 | 0.12 |

More pronounced differences were found between regions. There was a significant difference between the North and Central regions in Stage 0 lexical verb (LL 8.26, log ratio 1.17) and Stage 2 semi-complementizer (LL 12.07, log ratio 2.65) use, with the Central speaker using *shuō* in these ways roughly 2-5 times as often as the North speakers. Differences between the Central and South speakers were even more pronounced, with South speakers using *shuō* almost six times less frequently at Stage 0 (LL 26.88, log ratio -2.89), over two times less frequently at Stage 1 (LL 9.38, log ratio -1.13), and over four times less frequently at Stage 2 (LL 8.20, log ratio -2.18). A direct comparison between the North and South speakers also showed the South using Stage 0 (LL 10.19, log ratio -1.53) and Stage 1 (LL 34.89, log ratio -1.56) *shuō* roughly three times less frequently than the North. And though it was not a statistically significant difference, it is also worth noting that the only speaker to use *shuō* as a discourse marker in the New Corpus came from the South.

Table 24. Comparison of *shuō* use in the New Corpus by region

| | North | | Central | | Difference | |
|---|---|---|---|---|---|---|
| | *raw freq.* | *per thousand* | *raw freq.* | *per thousand* | *log likelihood* | *log ratio* |
| **Stage 0** | 32 | 2.49 | 23 | 5.60 | **8.26** | **1.17** |
| Stage 1 | 131 | 10.17 | 31 | 7.55 | 2.36 | -0.43 |
| **Stage 2** | 5 | 0.39 | 10 | 2.44 | **12.07** | **2.65** |
| Stage 3 | 3 | 0.23 | 1 | 0.24 | 0.00 | 0.06 |
| Stage 4 | 2 | 0.16 | 0 | 0.00 | 1.11 | -0.35 |
| Total | 173 | 13.44 | 65 | 15.84 | 1.25 | 0.24 |

| | Central | | South | | Difference | |
|---|---|---|---|---|---|---|
| | *raw freq.* | *per thousand* | *raw freq.* | *per thousand* | *log likelihood* | *log ratio* |
| **Stage 0** | 23 | 5.60 | 7 | 0.76 | **26.88** | **-2.89** |
| **Stage 1** | 31 | 7.55 | 32 | 3.45 | **9.38** | **-1.13** |
| **Stage 2** | 10 | 2.44 | 5 | 0.54 | **8.20** | **-2.18** |
| Stage 3 | 1 | 0.24 | 2 | 0.22 | 0.01 | -0.13 |
| Stage 4 | 0 | 0.00 | 0 | 0.00 | 0.00 | -1.18 |
| Discourse | 0 | 0.00 | 1 | 0.11 | 0.73 | *undefined* |
| **Total** | 65 | 15.84 | 47 | 5.07 | **35.68** | **-1.64** |

| | North | | South | | Difference | |
|---|---|---|---|---|---|---|
| | *raw freq.* | *per thousand* | *raw freq.* | *per thousand* | *log likelihood* | *log ratio* |
| **Stage 0** | 32 | 2.49 | 7 | 0.86 | **10.19** | **-1.53** |
| **Stage 1** | 131 | 10.17 | 33 | 3.45 | **34.89** | **-1.56** |
| Stage 2 | 5 | 0.39 | 5 | 0.54 | 0.27 | 0.47 |
| Stage 3 | 3 | 0.23 | 2 | 0.22 | 0.01 | -0.06 |
| Stage 4 | 2 | 0.16 | 0 | 0.00 | 2.17 | -1.53 |
| Discourse | 0 | 0.00 | 1 | 0.11 | 1.74 | *undefined* |
| **Total** | 173 | 13.44 | 47 | 5.07 | **41.27** | **-1.14** |

### 6.6 Discussion

The predictions for *shuō* use were that there would be a greater degree of individual variation and less use frequent use of *shuō* in the Old Corpus than in the other corpora. In the NCCU Corpus, more consistency of use across conversations expected, as well as increased use in Stages 1 and 2. Between the NCCU Corpus and the New Corpus, a decrease in the use of Stage 0 *shuō* and increase in that of Stages 3 and 4 was anticipated. Finally, when looking at the *shuō* use of speakers from different regions, North speakers were expected to have the bulk of their use in Stages 0-2 with Central speakers following the same pattern but using *shuō* more frequently. South speakers were predicted to have Stage 3 and 4 uses make up a greater portion of their total *shuō* use than the other two regions.

In the Old Corpus, as predicted, there was a high degree of individual variation in the rate of *shuō* use. Some speakers, like the Student, used *shuō* frequently, while others, like the Clerk, used it rarely. The Clerk only used *shuō* once as a Stage 1 quotative, while the Student and the Professor each used Stage 3/4 full complementizer *shuō* more than once. The Professor in particular favored more grammaticalized uses of *shuō*, with a third of her tokens occurring as a Stage 4 full complementizer or as a discourse marker.

(5) Use of Stage 4 complementizer *shuō* by 'the Professor'

| 我 | 又...又 | 怕 | 說 | 一 | 開始 | 教 | 就 | 要 | 考, |
|---|---|---|---|---|---|---|---|---|---|
| wǒ | yòu...yòu | pà | shuō | yī | kāishǐ | jiào | jiù | yào | kǎo |
| 1.SG | again | afraid | $COMP_{SAY}$ | once | start | teach | then | need | test |

| 一 | 考 | 就 | 會 | 亂掉。 |
|---|---|---|---|---|
| yī | kǎo | jiù | huì | luàndiào |
| once | test | then | will | messy |

"I'm afraid that once I start teaching, I will need to take a test, and once I take the test, I'll mess up."

(line 60)

By contrast, in the NCCU Corpus, the rate and type of *shuō* use was much more regular, exactly in line with the predictions of the extended RT framework. Where the average standard deviation in normalized frequency for the Old Corpus was 1.94, the NCCU Corpus had an average of only 0.80. *Shuō* was also used more often in the NCCU Corpus, with a normalized frequency of 10.30 uses per thousand words in comparison to 7.42 in the Old Corpus. The ways in which *shuō* was used in the two corpora also showed the expected differences. Like the Old Corpus, the NCCU Corpus saw the bulk of its *shuō* tokens fall in Stage 0 and Stage 1, however, the balance of that distribution reversed from approximately 60% Stage 0 and 30% Stage 1 in the Old Corpus to roughly 30% Stage 0 and 60% Stage 1 in the NCCU Corpus. Additionally, Stage 2 semi-complementizer *shuō* emerged as a minor pattern in the NCCU Corpus, occurring at least twice in every conversation and with a variety of verbs.

(6) Stage 2 semi-complementizer *shuō* in the NCCU Corpus

| 他 | 在 | 自我 | 介紹 | 說 | 他 | 是 | 衣索比亞 | 人 |
|---|---|---|---|---|---|---|---|---|
| tā | zài | zìwǒ | jièshào | shuō | tā | shì | yīsuǒbǐyǎ | rén |
| 3.SG | at | self | introduce | say$_{\text{SEMI-COMP}}$ | 3.SG | COP | Ethiopia | person |

"He introduced himself saying he is Ethiopian."

(NCCU 1, line 256)

| 你 | 可以 | 就是 | 叫 | 他 | 說 | 你 | 今天 |
|---|---|---|---|---|---|---|---|
| nǐ | kěyǐ | jiùshì | jiào | tā | shuō | nǐ | jīntiān |
| 2.SG | can | just.like | call | 3.SG | say$_{\text{SEMI-COMP}}$ | 2.SG | today |

| 之內... | 應該 | 一 | 個 | 小時 | 嘛 |
|---|---|---|---|---|---|
| zhīnèi | yīnggāi | yī | gè | xiǎoshí | ma |
| within | should | one | CL | hour | SFP |

"You can just tell him you should have one hour today."

(NCCU 6, 136)

| 那 | 體育 | 老師 | 就 | 跟 | 我 | 講 | 說, | 好 | 啊, |
|---|---|---|---|---|---|---|---|---|---|
| nà | tǐyù | lǎoshī | jiù | gēn | wǒ | jiǎng | shuō | hǎo | a |
| that | P.E. | teacher | then | with | 1.SG | say | say$_{\text{SEMI-COMP}}$ | good | SFP |

| 如果 | 他們 | 不 | 乖 | 你 | 就 | 告訴 | 我。 |
|---|---|---|---|---|---|---|---|
| rúguǒ | tāmen | bù | guāi | nǐ | jiù | gàosù | wǒ |
| if | 3.PL | NEG | good | 2.SG | then | tell | 1.SG |

"Then that P.E. teacher told me, "Ok, if they're not good, you can tell me.""

(NCCU 2, line 258)

Finally, the predicted sporadic occurrence of Stage 3/4 *shuō* was also born out by the data, with such tokens appearing at low frequencies in less than half of the conversations.

(7) Stage 3 full complementizer *shuō* in the NCCU Corpus

| 然後 | 就 | 會 | 覺得 | 說 | 整個 | 人 |
|---|---|---|---|---|---|---|
| ránhòu | jiù | huì | juédé | shuō | zhěnggè | rén |
| after | then | will | think | COMP$_{SAY}$ | entire | person |

| 很 | 清爽 |
|---|---|
| hěn | qīngshuǎng |
| very | refreshed |

"Then you will think that the whole person is refreshed."

(NCCU 7, line 407)

(8) Stage 4 full complementizer *shuō* in the NCCU Corpus

| 意思 | 是 | 說 | 我 | 有 | 一 | 個 | 教學 |
|---|---|---|---|---|---|---|---|
| yìsi | shì | shuō | wǒ | yǒu | yī | gè | jiàoxué |
| meaning | is | COMP$_{SAY}$ | 1.SG | have | one | CL | teaching |

| 實習 | 的 | 輔導老師 |
|---|---|---|
| shíxí | de | fǔdǎo lǎoshī |
| practicum | DE | tutor |

"It means that I have a teaching practicum tutor."

(NCCU 6, line 273)

The changes between the NCCU Corpus and the New Corpus were expected to be slight. *Shuō* was anticipated to again increase in frequency, but the overall pattern of its use was not predicted to change: Stage 1 tokens would still be the most frequent, followed by Stage 0 and Stage 2, in that order. What was expected to change was the proportion of tokens occurring in each context. Stage 0 use would decline, while Stage 1 and 2 uses would increase, leading to a wider gap between fully lexical and at least partially grammaticalized uses of *shuō*. This prediction was largely confirmed by the data. *Shuō* did increase its overall rate of use, with a

corpus normalized frequency of 11.20 per thousand words (compared to the NCCU Corpus at 10.30), and there was a decrease in use of Stage 0 *shuō* and increase in use of Stage 1. In fact, the difference in Stage 1 quotative use was the only statistically significant difference between the corpora (LL 5.35, log ratio 0.32). Additional expected differences were an increase in the use of *shuō* as a Stage 3/4 complementizer, with Stage 3 *shuō* occurring at least once in every conversation.

(8) Stage 3 complementizer *shuō* in the New Corpus

| 她 | 就 | 覺得 | 說, | 反正 | 她 | 就 | 是 |
|---|---|---|---|---|---|---|---|
| tā | jiù | juédé | shuō | fǎnzhèng | tā | jiù | shì |
| 3.SG | then | think | $COMP_{SAY}$ | anyway | 3.SG | then | COP |

| 平常 | 上班 |
|---|---|
| píngcháng | shàngbān |
| usually | go to work |

"She thinks that she usually goes to work anyway."

(Conversation #3, line 374)

| 我 | 才 | 反應 | 過來 | 說 | 好像 | 講 | 太 | 快 |
|---|---|---|---|---|---|---|---|---|
| wǒ | cái | fǎnyìng | guòlái | shuō | hǎoxiàng | jiǎng | tài | kuài |
| 1.SG | just.then | realize | EXP | $COMP_{SAY}$ | seem | speak | too | fast |

"I just then realized that I seemed to be speaking too fast."

(Conversation #1, line 54)

Though it was not statistically significant, there was a slight decrease in the normalized frequency of Stage 2 semi-complementizers between the NCCU Corpus and the New Corpus. Given the lack of significance, however, and the continued variety of verbs used in the $V_1$ position, it is extremely unlikely that this decline was due to anything but random chance.

(9) Stage 2 semi-complementizer *shuō* in the New Corpus

| 他 | 有 | | 告訴 | 你 | 說 | | 那 | 個 |
|---|---|---|---|---|---|---|---|---|
| tā | yǒu | | gàosù | nǐ | shuō | | nà | gè |
| 3.SG | AUX$_{\text{COMPLETION}}$ | | tell | 2.SG | say$_{\text{SEMI-COMP}}$ | | that | CL |
| 參考 | 書 | 要 | 推薦 | | 什麼 | | | |
| cānkǎo | shū | yào | tuījiàn | | shénme | | | |
| reference | book | want | recommend | | what | | | |

"Did he tell you what that reference book recommended?"

(Conversation #4, line 17)

| 我 | 等下 | 就 | 問 | 他 | 說 | 嗯.. | 請問... |
|---|---|---|---|---|---|---|---|
| wǒ | děngxià | jiù | wèn | tā | shuō | ń | qǐngwèn |
| 1.SG | wait a bit | then | ask | 3.SG | say$_{\text{SEMI-COMP}}$ | um | excuse me |

"I'll wait a bit and ask him, um, "Excuse me…"

(New NCCU, line 718)

Discourse marker *shuō* also did not experiences significant changes between the two corpora. It was used once in each corpus, both times as a sentence-final intensifier.

(10) *Shuō* used as an intensifier

| 反正 | 他 | 也 | 不 | 要… | 他 | 也 | 不 | 想 |
|---|---|---|---|---|---|---|---|---|
| fǎnzhèng | tā | yě | bù | yào… | tā | yě | bù | xiǎng |
| anyway | 3.SG | also | NEG | want | 3.SG | also | NEG | would.like |
| 睡 | 午覺… | 他 | 就 | 覺得 | 說 | | | |
| shuì | wǔjiào | tā | jiù | juédé | shuō | | | |
| sleep | nap | 3.SG | then | think | SAY | | | |

"Anyway, he also doesn't want…he also doesn't want to take a nap…that's his feeling."

(NCCU #2, line 286)

覺得 他... 他 就 會 很 有 禮貌 啊 說

juédé tā… tā jiù huì hěn yǒu lǐmào a shuō

think 3.SG 3.SG then will very have manners SFP SAY

"I think he…he will be very polite."

(Conversation #2, line 254)

Though the other discourse marker uses noted by Wang et al. (2003) were not present in the corpora examined here, that is likely a consequence of the size of the corpora and infrequency of the feature, rather than of language change.

The regional data also aligned with the predictions of the extended RT approach. As anticipated, the North used *shuō* much more often than the South, with a normalized frequency of 13.44 in comparison to the South's 5.07. This difference was most marked at Stages 0 and 1, where the disparity reached the level of statistical significance. And while the North did use *shuō* as a full complementizer more often than the South in terms of normalized frequency, when looking at the percentage of total use, Stage 3 *shuō* made up 4.3% of all uses for the South speakers, whereas Stage 3 and 4 *shuō* combined only accounted for 2.9% of the North's total *shuō* use. Furthermore, the single instance of discourse marker *shuō* in the New Corpus was by a speaker from the South.

**6.7 Conclusion**

This chapter presented the argument for treating *shuō* in TM as an example of replica grammaticalization. Corpus data was used to demonstrate how *shuō* has 1) increased in its general frequency of use over time, and 2) become more likely to be used in grammaticalized forms over time, both circumstances predicted by RT in general and the extended RT approach

used in the current study. The regional data also matched the predictions of extended RT, both in terms of distribution patterns within regions and frequency of use.

# Chapter 7: Co-verb and pro-verb *yòng* 用

## 7.1 Introduction

Though it has received little scholarly attention, one of the minor use patterns in Taiwan Mandarin that shows evidence of having undergone replica grammaticalization is the use of *yòng* 用 'use' as a co-verb or pro-verb. This chapter will begin with a review of the existing work on co-verb and pro-verb *yòng,* followed by the justification for treating it as an example of replica grammaticalization. Next, the predictions regarding chronological and regional differences will be laid out, followed by the results of the current study. This chapter will then conclude with a discussion of the results as they relate to the extended RT approach.

## 7.2 Literature Review

Early accounts of TM syntax note *yòng* being used in two ways not possible in PTH. First, it could be used to mark instrumental case. Second, when used in conjunction with a nominalized verb (i.e. a verb followed by the particle *de* 的) TM *yòng* could "form a main predicate indicating manner" (Cheng 1985, p. 367).

(1) Manner co-verb *yòng*

| 你 | 要 | 用 | 跑 | 的 | 才 | 來得及 |
|---|---|---|---|---|---|---|
| nǐ | yào | yòng | pǎo | de | cái | láidejí |
| 2.SG | must | use | run | DE | only.then | in.time |

"In order to be on time, you have to run."

Both of these uses are quite common with the TSM equivalent *iēng* (Kubler 1981), which may additionally be used as a pro-verb, as in (2).

(2) Pro-verb *iēng* in TSM

| 伊 | 佮 | 碗 | 用 | 置 | 土腳 |
|---|---|---|---|---|---|
| i | ka | oan | iong | ti | thokha |
| 3.SG | DISPOSAL | bowl | YONG | on | floor |

"He drops the bowl on the floor."

(adapted from Liu & Xu 2013, p. 546)

This pro-verb use does not appear to have been possible in TM until relatively recently. In 2013, Liu and Xu published their study on the emergence of pro-verb *yòng* in TM, identifying it as a construction in which *yòng* serves either as a stand-in for an earlier verb in the discourse or as semantically void verb whose meaning must be inferred from pragmatics.

(3) Pro-verb *yòng*: replaced verb

| **A:** 你 | 可以 | 幫 | 我 | 修 | 車 | 嗎? |
|---|---|---|---|---|---|---|
| nǐ | kěyǐ | bāng | wǒ | xīu | chē | ma |
| 2.SG | can | help | 1.SG | fix | car | Q |

"Can you help me fix the car?"

| **B:** 好， | 我 | 來 | 用 |
|---|---|---|---|
| hǎo | wǒ | lái | yòng |
| ok | 1.SG | come | YONG |

"Ok, I'll come do it (help fix the car)."

(4) Pro-verb *yòng*: inferred verb

| 我 | 昨天 | 整 | 天 | 都 | 在 | 用 | 報告， |
|---|---|---|---|---|---|---|---|
| wǒ | zuótiān | zhěng | tiān | dōu | zài | yòng | bàogào |
| 1.SG | yesterday | entire | day | all | PROG | YONG | paper |

| 超 | 累 | 的 |
|---|---|---|
| chāo | lèi | de |
| super | tired | DE |

"I was writing the paper all day yesterday; I was exhausted."

(adapted from Liu & Xu 2013, p. 541)

As with TSM pro-verb *yòng*, this construction in TM has been restricted to verbs of physical action where the object NP fills the semantic role of Patient (Liu & Xu 2013). While Liu and Xu (2013) attribute the existence of pro-verb *yòng* in TM to the influence of TSM, because their study includes only synchronic data, they ultimately make no claim about whether that influence manifested as polysemy copying or grammaticalization.

### 7.3 Extended RT approach

Viewing *yòng* in TM diachronically, it becomes clear that this feature is a case of replica grammaticalization, not polysemy copying. The use of *yòng* as a co-verb to mark instrumental case is common across Chineses, so it is very likely that MM had this feature at the time of contact with TSM. However, TSM, in addition to its use of *iēng* as a pro-verb, was also unique among Chineses in how it used co-verb *iēng*, employing it to express manner of action, purpose, source, goal, etc. and overtly marking the nominalization of verbs used in the construction (Cheng 1985). There was therefore a gap in degree of grammaticalization between MM and

TSM, a gap that was preserved in early TM, where co-verb *yòng* was used in only some of the more grammaticalized ways found in TSM (i.e. expressing manner of action). This intermediate stage does not occur in cases of polysemy copying, but does in cases of grammatical replication. And as with *yǒu*+VP and complementizer *shuō*, the fact that homophony still existed in the model language (TSM) between the lexical and grammatical forms of the feature means that grammatical replication manifested as replica grammaticalization. Furthermore, even though contemporary TM and TSM may appear to be identical in their use of *yòng*/*iēng*, closer examination will reveal that there are slight differences between the co-verb forms. There is therefore ample reason to not treat this feature as a polysemy copy, but as a replica grammaticalization.

### 7.4 Predictions

Taking an extended RT approach, then, we can make several predictions about the use of co-verb and pro-verb *yòng* in TM. Initially, we would expect some speakers to use co-verb *yòng* occasionally to communicate manner of action with overtly nominalized verbs. Over time, the overall frequency of this co-verb *yòng* pattern would increase, but the explicit marking of nominalization, as a rare element in the feature pool, would fall out of use. This is the predicted state of TM in the NCCU Corpus. It is not until the New Corpus that any uses of pro-verb *yòng* are expected. When examining the New Corpus data by region, the North and Central regions are expected to pattern similarly as their feature pools would have been relatively alike: Hakka, like MM, only allows main verb and instrumental co-verb uses of *yòng*/*yung*$^{4}$ (Chappell & Lamarre 2005). The South is predicted to have a distinct distribution pattern, with a greater

portion of *yòng* tokens occurring as co-verbs and pro-verbs.

## 7.5 Results

Adjectival uses of *yòng* (i.e. *yǒuyòng* 有用 'useful') were excluded from analysis, as were compound verbs that contained *yòng* (e.g. *shǐyòng* 使用 'to make use of').

### *7.5.1 Old Corpus*

In the Old Corpus, *yòng* was used a total of 12 times, with 8 of those uses employing *yòng* as a main verb.

Table 25. Old Corpus: Raw frequency of *yòng* by speaker

| | **Professor** | **Student** | **Clerk** | **Housewife** | **Interviewer** | **Total** |
|---|---|---|---|---|---|---|
| **Main verb** | 7 | 0 | 0 | 0 | 1 | **8** |
| **Co-verb** | 0 | 0 | 0 | 2 | 2 | **4** |
| **Pro-verb** | 0 | 0 | 0 | 0 | 0 | **0** |
| **Total** | **7** | **0** | **0** | **2** | **3** | **12** |

The 4 uses of *yòng* as a co-verb all occurred in the interview with the Housewife, with each speaker using co-verb *yòng* twice to mark instrumental case.

(5) Instrumental co-verb *yòng*

用 誰 的 房子 去 抵?
yòng shéi de fángzi qù dǐ
use who DE house go offset
"Whose house was used to offset it?"

('the Housewife', Interviewer, line 25)

The only other interview in which *yòng* appeared was 'the Professor', where the Professor

repeatedly used *yòng* as a main verb. The normalized frequency of *yòng* used as a main verb or co-verb in this corpus was 1.31 uses per thousand words.

### *7.5.2 NCCU Corpus*

Verbal *yòng* was used consistently, though infrequently, in the NCCU Corpus, 75% of conversations using it at least once, but no conversation using it more than 6 times.

Table 26. NCCU Corpus: Raw frequency of *yòng*

| | NCCU1 | NCCU2 | NCCU3 | NCCU4 | NCCU5 | NCCU6 | NCCU7 | NCCU8 | Total |
|---|---|---|---|---|---|---|---|---|---|
| **Main verb** | 0 | 1 | 1 | 4 | 0 | 4 | 0 | 1 | **11** |
| **Co-verb** | 2 | 1 | 0 | 1 | 0 | 2 | 0 | 0 | **6** |
| **Pro-verb** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** |
| **Total** | **2** | **2** | **1** | **5** | **0** | **6** | **0** | **1** | **17** |

Of the 6 co-verb uses of *yòng*, 5 mark instrumental case. The single token of *yòng* used to indicate manner of action did not mark the verb with *de*.

(6) Manner co-verb *yòng*

**A:** 可是 他 有 這麼 容易 收服 嗎?
kěshì tā yǒu zhème róngyì wàifú ma
but 3.SG YOU$_{\text{ASSERTION}}$ so easy conquer Q
"But is he indeed so easy to conquer?"

**B:** 就 用 看看 吶
jiù yòng kànkàn ne
just use look-look P
"Just take a look."

(NCCU 6, lines 151-152)

The normalized frequency of *yòng* in the NCCU Corpus was 0.56 uses per thousand words.

*7.5.3 New Corpus*

Use of *yòng* varied greatly between conversations in the New Corpus, with Conversation #2 using it not at all and the New NCCU conversation using it 32 times. The New NCCU conversation is something of an outlier, however, as most conversations used verbal *yòng* fewer than 10 times.

Table 27. New Corpus: Raw frequency of *yòng*

| | **Convo. #1** | **Convo. #2** | **Convo. #3** | **Convo. #4** | **New NCCU** | **Total** |
|---|---|---|---|---|---|---|
| **Main verb** | 0 | 0 | 6 | 7 | 24 | **37** |
| **Co-verb** | 3 | 0 | 1 | 1 | 7 | **12** |
| **Pro-verb** | 0 | 0 | 0 | 0 | 1 | **1** |
| **Total** | **3** | **0** | **7** | **8** | **32** | **50** |

This gap is less extreme when considering normalized frequencies, although even then the New NCCU conversation used *yòng* more than twice as often as any other conversation.

Table 28. New Corpus: Normalized frequency of *yòng*

| | **Convo. #1** | **Convo. #2** | **Convo. #3** | **Convo. #4** | **New NCCU** | **Avg.** |
|---|---|---|---|---|---|---|
| **Main verb** | 0.00 | 0.00 | 0.74 | 1.36 | 2.47 | **4.56** |
| **Co-verb** | 0.38 | 0.00 | 0.12 | 0.19 | 0.72 | **1.42** |
| **Pro-verb** | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | **0.10** |
| **Total** | **0.38** | **0.00** | **0.86** | **1.55** | **3.29** | |

Co-verb uses of *yòng* made up about 25% of uses in the New Corpus, all of them used to mark instrumental case. There was one use of *yòng* as a pro-verb whose meaning had to be inferred from context.

(6) Pro-verb *yòng* (inferred)

**A:** 能 幾 句 話 然後 就 加 一 次

néng jǐ jù huà ránhòu jiù jiā yī cì

can a.few CL word after then add one instance

那 個 符號 我 就 很 煩

nà gè fúhào wǒ jiù hěn fán

that CL symbol 1.SG then very annoyed

"(He) can say a few words and then add that symbol once, which makes me very annoyed."

**B:** 那 以後 就 用

nà yǐhòu jiù yòng

then in.the.future just YONG

"Then in the future, just say so."

**A:** 什麼 意思?

shénme yìsi

what meaning

"What do you mean?"

**B:** 以後 就 跟 他 說

yǐhòu jiù gēn tā shuō

in.the.future just with 3.SG speak

"In the future, just tell him."

(New NCCU, lines 676-679)

The combined normalized frequency of main verb, co-verb, and pro-verb *yòng* in the New Corpus was 1.39 uses per thousand word.

Of the three geographical regions, the North used verbal *yòng* the most. The Central region was the only one not to use *yòng* as a co-verb, and was also the region that used *yòng* the

least overall. Only the South used co-verb *yòng* more than main verb *yòng*.

Table 29. New Corpus: Raw frequency of *yòng* by region

| | **North** | **Central** | **South** | **Total** |
|---|---|---|---|---|
| **Main verb** | 9 | 2 | 2 | **13** |
| **Co-verb** | 2 | 0 | 3 | **5** |
| **Pro-verb** | 0 | 0 | 0 | **0** |
| **Total** | **11** | **2** | **5** | **18** |

For both the North and South, co-verb *yòng* was used only to mark instrumental case.

### *7.5.4 Comparison of Corpora*

The Old Corpus and NCCU Corpus saw similar distributions of verbal *yòng*, with roughly two-thirds of *yòng* tokens used as main verbs and the remaining serving as co-verbs. While the NCCU Corpus had a lower normalized frequency for *yòng* than the Old Corpus, it did use *yòng* in more varied ways, using co-verb *yòng* both to mark instrumental case and to form a main predicate of manner. These patterns changed between the NCCU Corpus and the New Corpus. Firstly, there was a statistically significant increase in normalized frequency of *yòng* (LL 11.90, log ratio 1.31), with the bulk of that difference coming from a significant increase in use of *yòng* as a main verb (LL 10.83, log ratio 1.51).

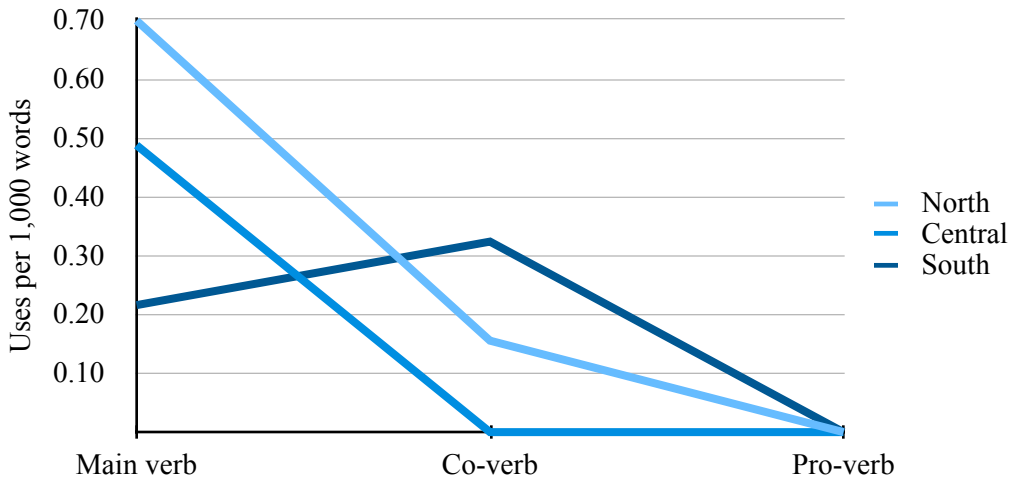
Table 30. Comparison of *yòng* use between the NCCU Corpus and New Corpus

| | NCCU Corpus | | New Corpus | | Difference | |
|---|---|---|---|---|---|---|
| | *raw freq.* | *per thousand* | *raw freq.* | *per thousand* | *log likelihood* | *log ratio* |
| **Main verb** | 11 | 0.36 | 37 | 1.03 | **10.83** | **1.51** |
| Co-verb | 6 | 0.20 | 12 | 0.33 | 1.16 | 0.76 |
| Pro-verb | 0 | 0.00 | 1 | 0.03 | 1.22 | *undefined* |
| **Total** | 17 | 0.56 | 50 | 1.39 | **11.90** | **1.31** |

Secondly, co-verb *yòng* was used more consistently in the New Corpus, with 80% of conversations containing at least one such token, as opposed to 50% in the NCCU Corpus. The types of usages, however, decreased between the two corpora, with co-verb *yòng* used only to mark instrumental case in the New Corpus. Pro-verb *yòng* appeared for the first time in the New Corpus, though not in one of the conversations that recorded the participants' region of origin.

When comparing regions, although the North used *yòng* the most overall, speakers from the South were more likely to use co-verb *yòng*, with it making up 60% of their *yòng* tokens as opposed to 18.2% in the North. The Central region data patterned similarly to that of the North, with rates of use falling as degree of grammaticalization increased. The South, by contrast, saw an increase in normalized frequency from main verb to co-verb use.

Figure 7. Comparison of *yòng* use between regions




**7.6 Discussion**

These results are in line with the predictions of the extended RT approach. More grammaticalized uses of *yòng*, namely, co-verb and pro-verb uses, do increase in frequency over time, as does overall use of *yòng*. Also expected was the lack of overt marking on the manner

co-verb used in the NCCU Corpus, as the presence of this feature only in TSM and the low frequency of its occurrence therein made this feature one that was unlikely to be selected for TM. What was somewhat surprising was that manner co-verb *yòng* was not attested in the New Corpus, though as it was very rare in the NCCU Corpus, accounting for only about 0.5% of all *yòng* tokens, it is entirely possible that this was merely a matter of chance. Very much in line with the extended RT predictions, though, was the appearance of pro-verb *yòng* for the first time in the New Corpus. Interestingly, this particular use of pro-verb *yòng* did not occur in place of a verb of physical action, but rather one of communication, indicating that this construction has continued to undergo extension in the decade or so since Liu and Xu's (2013) study. Regional differences in the New Corpus also followed the predictions of extended RT, with the North and Central regions using *yòng* primarily as a main verb while the South used it most of the time as a co-verb.

### 7.7 Conclusion

Co-verb and pro-verb *yòng* in Taiwan Mandarin have been presented here as an example of replica grammaticalization. While current usage in TM may give the impression that these use patterns for *yòng* are simply the result of polysemy copying from TSM, a more detailed examination of both the synchronic and diachronic data shows subtle differences in how *yòng* is used in TM and TSM, indicating that it is indeed grammatical replication that has been at work. Additional evidence for this conclusion is found in the regional data, where differences in how *yòng* is used in different parts of Taiwan fit the predictions of the extended RT approach used in

this study, and in the fact that pro-verb *yòng* in TM now more grammaticalized than its counterpart in TSM.

# Chapter 8: A-not-A questions

## 8.1 Introduction

Among the many ways to form a question in Chinese languages is the A-not-A construction, in which a verb is followed by its negative counterpart. The specifics of how this construction is used can vary greatly between Chineses, making it an interesting subject for diachronic and contact studies. TM is no exception in this regard. What is particularly interesting about this use pattern in comparison to those discussed earlier in this work, however, is that this seems to be a case where MM, rather than TSM, has served as the model language. Additionally, while, as a syntactic pattern, A-not-A questions cannot be said to have grammaticalized in TM in the classical sense, in the RT sense of extending to new contexts to shift from minor use pattern to major use pattern status, the framework is still applicable. To support this analysis, this chapter will review the relevant literature on A-not-A questions in TM, present the results of the present study, and then demonstrate how those results argue for an extended RT approach to this feature that treats MM as the model language.

## 8.2 Literature Review

In TSM, several constraints exist on the use of the A-not-A construction. First, only auxiliaries (or verbs that can act as auxiliaries, such as ‘to have’) and the copula may occur in an A-not-A question with only three exceptions: the verb ‘to know’ and the adjectives ‘good’ and ‘right’. Additionally, A-not-A questions in TSM experience separation, with the initial ‘A’

occurring before the main verb and the ‘not-A’ portion coming at the end of the sentence (Lin 1974).

(1) Separation A-not-A in TSM

| 汝 | 有 | 菸 | 還是 | 無 | 菸? |
|---|---|---|---|---|---|
| lí | ū | hûn | a-sǐ | bóu | hûn |
| 2.SG | have | cigarette | or | NEG.have | cigarette |

“Do you have cigarettes?”

(adapted from Cheng 1985, p. 369)

There are two exceptions to this pattern, the first being the copula, which occurs in its entirety between the subject and predicate, and the second being tag questions, where the entire A-not-A structure occurs sentence-finally.

(2) Copula A-not-A in TSM

| 汝 | 是 | 毋 | 是 | 學生? |
|---|---|---|---|---|
| lí | sī | m̄ | sī | hàk-seng |
| 2.SG | be | NEG | be | student |

“Are you a student?”

(adapted from Li 1974, p. 39)

Another unusual feature of TSM A-not-A questions is the obligatory deletion of the second syllable of a bisyllabic verb in the initial ‘A’ position.

(3) Deletion A-not-A in TSM

| 知 | 毋 | 知道? |
|---|---|---|
| cāi | m̄ | cāi-ià |
| know | NEG | know |

“Do (you) know?”

(adapted from Kubler 1981, p. 102)

A-not-A questions in the PTH of the mid-20th century were substantially different. Main verbs, auxiliaries, and adjectives could all freely occur in A-not-A questions, and bisyllabic verbs did not obligatorily undergo deletion (though it was optional). However, there was some overlap with TSM when it came to separation, as when auxiliary verbs were used in an A-not-A question, the 'not-A' portion occurred sentence-finally.

(4) Separation A-not-A in PTH

| 你 | 會 | 打 | 牌 | 不 | 會? |
|---|---|---|---|---|---|
| nǐ | huì | dǎ | pái | bú | huì |
| 2.SG | can | play | cards | NEG | can |

"Can you play cards?"

(Adapted from Cheng 1985, p. 369)

Undoubtedly, separation is a complex structure, as the main verb phrases may be long and intricate, even to the point of containing embedded clauses. This complexity is likely why TSM, PTH, and Hakka are the only varieties of Chinese to use separated A-not-A questions when the 'A' in question is an auxiliary (Cheng 1985). Given this rarity among Chinese languages, there is no reason to expect that this pattern would have been selected for MM from the feature pool formed by the Mainlanders on Taiwan. Instead, MM would have used the adjoined pattern for all A-not-A questions, and would have allowed all classes of verbs (including adjectives as stative verbs) to be used as 'A'. As more grammaticalized forms are less contextually constrained in their use than less grammaticalized forms, the use of A-not-A questions with all verbs in MM but functionally one class of verbs in TSM makes the MM pattern the more grammaticalized one. This is therefore a case where MM, rather than TSM, should be treated as the model language with respect to TM (replica language).

### 8.3 Predictions

If MM is treated as the model language for A-not-A questions in TM, then we can expect these questions to be most TSM-like in the Old Corpus. In other words, they should occur with fewer classes of verbs, appear more often in the separation pattern, and apply the deletion pattern more universally when compared to the more recent corpora. The deletion pattern in particular is expected to feature prominently, as it was noted by Kubler (1981) as being applied to all A-not-A constructions, even those that have no TSM equivalent, such as ones involving the potential complement construction.

(5) "Can you see it?" in PTH and TM

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **PTH:** | 你 | 看 | 得 | 見 | 看 | 不 | 見? |
| | nǐ | kàn | de | jiàn | kàn | bu | jiàn |
| | 2.SG | look | DE | see | look | NEG | see |
| **TM:** | 你 | 看 | 得 | 看 | 不 | 見? | |
| | nǐ | kàn | de | kàn | bu | jiàn | |
| | 2.SG | look | DE | look | NEG | see | |

(adapted from Kubler 1981, p. 103)

By the time of the NCCU Corpus, use of A-not-A questions should be more similar across speakers, and are expected to have extended to occurring sometimes with non-auxiliary verbs. The separation pattern is predicted to be less frequent, as its absence in MM and innate complexity make it unlikely to be selected from the feature pool, and the deletion pattern should no longer be applied to verb constructions that are not found in TSM. In the New Corpus, the separation pattern is expected to be rare, if not completely absent, and non-auxiliary verbs are predicted to make up a greater portion of A-not-A usages. Deletion is not expected to apply to

potential complement constructions, though the tendency of both frequent and grammaticalizing patterns to experience phonetic erosion makes it likely that deletion will still be applied to bisyllabic verbs more often than not. As for regional differences in the New Corpus, the North is, in this case, expected to demonstrate a higher degree of grammaticalization, using more main verbs and adjectives in A-not-A questions than the other two regions. The South is predicted to more heavily favor the use of auxiliary verbs in this construction. Finally, because of the greater proportion of Hakka speakers, the Central region is considered most likely under extended RT to still use the separation pattern, as separation is not only present in Hakka, but frequent (Chappell & Lamarre 2005). However, as Hakka A-not-A questions are, like MM, not restricted in the class of verbs they can occur with, the Central region data is expected to resemble that of the North with respect to the types of verbs used.

## 8.4 Results

Tag questions that utilize the A-not-A pattern were excluded from analysis, as they are not subject to separation or the verb choice restrictions that apply to other A-not-A questions.

### *8.4.1 Old Corpus*

The A-not-A construction was used 32 times in the Old Corpus, giving it a normalized frequency of 3.48 uses per thousand words. The vast majority of tokens came from the Interviewer, with the remaining few produced by the Clerk. Half of all A-not-A questions occurred with auxiliary verbs, and in all the tokens where the main verb participated in the A-not-A structure, that main verb was either *yǒu* 有, which can also function as an auxiliary in TM,

or the copula *shì* 是. Only 3 tokens of the separation pattern occurred in the corpus, all of them with *yǒu* 有. Each of the 5 bisyllabic verbs used — none of which were *zhīdào* 知道 'to know' — underwent deletion.

Table 31. Old Corpus: Raw frequency of A-not-A

| | **Professor** | **Student** | **Clerk** | **Housewife** | **Interviewer** | **Total** |
|---|---|---|---|---|---|---|
| **Auxiliary** | 0 | 0 | 3 | 0 | 13 | **16** |
| **Main** | 0 | 0 | 0 | 0 | 8 | **8** |
| **Deletion** | 0 | 0 | 1 | 0 | 4 | **5** |
| **Separation** | 0 | 0 | 2 | 0 | 1 | **3** |
| **Total** | **0** | **0** | **6** | **0** | **26** | **32** |

Figure 8. Old Corpus: A-not-A verbs

| **Auxiliary** | **Main** | **Deletion** | **Separation** |
|---|---|---|---|
| 有 'have' | 有 'have' | 反對 'oppose' | 有 'have' |
| 是 'be' | 是 'be' | 適合 'fit' | |
| 會 'will' | | 合適 'suit' | |
| 能 'can' | | 可以 'can' | |
| | | 喜歡 'like' | |

*8.4.2 NCCU Corpus*

Of the 67 total A-not-A questions, 50 occurred with a monosyllabic auxiliary. No instances of separation were present, and though deletion always occurred with a bisyllabic verb, only two such tokens were present in the corpus. Both instances of deletion occurred with the same auxiliary verb: *kěyǐ* 可以 'can'. One main verb token used the adjective *duō* 多 'many'; the rest of the main verb usages involved verbs that can also act as auxiliaries.

Table 32. NCCU Corpus: Raw frequency of A-not-A

| | NCCU1 | NCCU2 | NCCU3 | NCCU4 | NCCU5 | NCCU6 | NCCU7 | NCCU8 | Total |
|---|---|---|---|---|---|---|---|---|---|
| **Auxiliary** | 3 | 10 | 4 | 8 | 5 | 5 | 6 | 9 | **50** |
| **Main** | 0 | 6 | 0 | 3 | 2 | 0 | 3 | 1 | **15** |
| **Deletion** | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | **2** |
| **Separation** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0** |
| **Total** | **3** | **16** | **4** | **11** | **8** | **5** | **10** | **10** | **67** |

Figure 9. NCCU Corpus: A-not-A verbs

| Auxiliary | Main | Deletion | Separation |
|---|---|---|---|
| 有 ‘have’ | 有 ‘have’ | 可以 ‘can’ | |
| 是 ‘be’ | 是 ‘be’ | | |
| 會 ‘will’ | 多 ‘many’ | | |
| 要 ‘want’ | | | |

Though all conversations used A-not-A questions, there was a disparity in the rate of usage, from 0.89 in NCCU1 to 4.49 in NCCU2. The normalized frequency for the corpus as a whole was 2.20 uses per thousand words.

Table 33. NCCU Corpus: Normalized frequency of A-not-A

| | NCCU1 | NCCU2 | NCCU3 | NCCU4 | NCCU5 | NCCU6 | NCCU7 | NCCU8 | Total |
|---|---|---|---|---|---|---|---|---|---|
| **Auxiliary** | 0.89 | 2.81 | 1.08 | 1.52 | 1.64 | 1.10 | 1.48 | 3.15 | **13.67** |
| **Main** | 0.00 | 1.68 | 0.00 | 0.57 | 0.65 | 0.00 | 0.74 | 0.35 | **4.00** |
| **Deletion** | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 0.00 | 0.25 | 0.00 | **0.57** |
| **Separation** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | **0.00** |
| **Total** | **0.89** | **4.49** | **1.08** | **2.09** | **2.62** | **1.10** | **2.46** | **3.50** | |

*8.4.3 New Corpus*

In the New Corpus, monosyllabic auxiliary verbs were again the most common verb class

used in A-not-A questions, accounting for 42 tokens out of a total 59. Main verb uses were the next most common type at 11 tokens, with only a single example of the separation pattern occurring with the copula *shì* 是. Non-auxiliary *bào* 報 'to report' was also used once in the A-not-A construction. The remaining 6 A-not-A tokens occurred with one of two bisyllabic verbs — auxiliary *kěyǐ* 可以 'can' or main verb *xǐhuān* 喜歡 'to like' — and universally underwent deletion.

Table 34. New Corpus: Raw frequency of A-not-A

| | **Convo. #1** | **Convo. #2** | **Convo. #3** | **Convo. #4** | **New NCCU** | **Total** |
|---|---|---|---|---|---|---|
| **Auxiliary** | 8 | 13 | 7 | 0 | 14 | **42** |
| **Main** | 3 | 3 | 4 | 0 | 1 | **11** |
| **Deletion** | 0 | 1 | 2 | 0 | 2 | **5** |
| **Disjunction** | 0 | 0 | 1 | 0 | 0 | **1** |
| **Total** | **11** | **17** | **14** | **0** | **17** | **59** |

Figure 10. New Corpus: A-not-A verbs

| **Auxiliary** | **Main** | **Deletion** | **Disjunction** |
|---|---|---|---|
| 有 'have' | 有 'have' | 可以 'can' | 是 'be' |
| 是 'be' | 是 'be' | 喜歡 'like' | |
| 會 'will' | 報 'report' | | |
| 要 'want' | | | |
| 能 'can' | | | |

A-not-A questions were infrequent in the New Corpus, with a normalized frequency of 1.64 uses per thousand words. Their use, however, was quite uniform, as, if Conversation #4 is excluded as an outlier, the standard deviation in normalized frequencies across conversations was only 0.87.

Table 35. New Corpus: Normalized frequency of A-not-A

| | Convo. #1 | Convo. #2 | Convo. #3 | Convo. #4 | New NCCU | Total |
|---|---|---|---|---|---|---|
| **Auxiliary** | 1.01 | 2.55 | 0.86 | 0.00 | 1.44 | **5.86** |
| **Main** | 0.38 | 0.59 | 0.49 | 0.00 | 0.10 | **1.56** |
| **Deletion** | 0.00 | 0.20 | 0.25 | 0.00 | 0.21 | **0.65** |
| **Separation** | 0.00 | 0.00 | 0.12 | 0.00 | 0.00 | **0.12** |
| **Total** | **1.40** | **3.33** | **1.73** | **0.00** | **1.75** | |

Considered by region, the distribution pattern is much the same, with auxiliary uses making up at least half of all tokens in each region, followed by a lower number number of main verb usages. Deletion and separation tokens both remained either fairly rare or absent. The Central speaker was the only one to use the separation pattern, and was also responsible for 1 of the 3 deletion tokens in the regional data. While the South data contained no examples of the deletion pattern, it also had no use of a bisyllabic verb in the A-not-A construction, and thus no contexts in which the pattern could be applied. In terms of raw numbers, the North used A-not-A questions the most out of any region and the Central region the least. However, when looking at normalized frequencies, the Central region had the highest rate at 1.95 uses per thousand words, followed by the North at 1.79 and South at 1.19.

Table 36. New Corpus: Raw frequency of A-not-A by region

| | North | Central | South | Total |
|---|---|---|---|---|
| **Auxiliary** | 16 | 4 | 8 | **28** |
| **Main** | 5 | 2 | 3 | **10** |
| **Deletion** | 2 | 1 | 0 | **3** |
| **Separation** | 0 | 1 | 0 | **1** |
| **Total** | **23** | **8** | **11** | **42** |

Table 37. New Corpus: Normalized frequency of A-not-A by region

| | North | Central | South | Total |
|---|---|---|---|---|
| **Auxiliary** | 1.24 | 0.97 | 0.86 | **3.08** |
| **Main** | 0.39 | 0.49 | 0.32 | **1.20** |
| **Deletion** | 0.16 | 0.24 | 0.00 | **0.40** |
| **Separation** | 0.00 | 0.24 | 0.00 | **0.24** |
| **Total** | **1.79** | **1.95** | **1.19** | |

Of the regions, the South used the fewest individual verbs in A-not-A questions, all of which were auxiliaries or main verbs that can also be used as auxiliaries. Both the North and Central regions used one non-auxiliary verb in the A-not-A construction. The single separation token, produced by the Central speaker, used the copula *shì* 是.

Figure 11. New Corpus: A-not-A verbs by region

| Auxiliary | Main | Deletion | Separation |
|---|---|---|---|
| (N/C/S) 有 'have' | (N/S) 有 'have' | (N/C) 可以 'can' | (C) 是 'be' |
| (N/C/S) 是 'be' | (N/C) 是 'be' | (N) 喜歡 'like' | |
| (N/S) 會 'will' | (C) 報 'report' | | |
| (N/C) 要 'want' | | | |
| (N) 能 'can' | | | |

*8.4.4 Comparison of Corpora*

Between the Old Corpus and the NCCU Corpus, there was a decline in the frequency of A-not-A questions, from 3.48 uses per thousand words to 2.20. The distribution of tokens also changed, with an increase in the share of A-not-A questions that occurred with monosyllabic auxiliary verbs, but a decrease in the proportion that occurred with main- and bisyllabic verbs, with the separation pattern vanishing entirely. There were also fewer main verbs used in the

NCCU Corpus that could not also be auxiliary verbs, with a single use of the adjective *duō* 多 'many' being the only example. However, the NCCU Corpus did see comparatively more regular use, with every conversation containing at least one A-not-A question, even if the rate of use within each conversation varied.

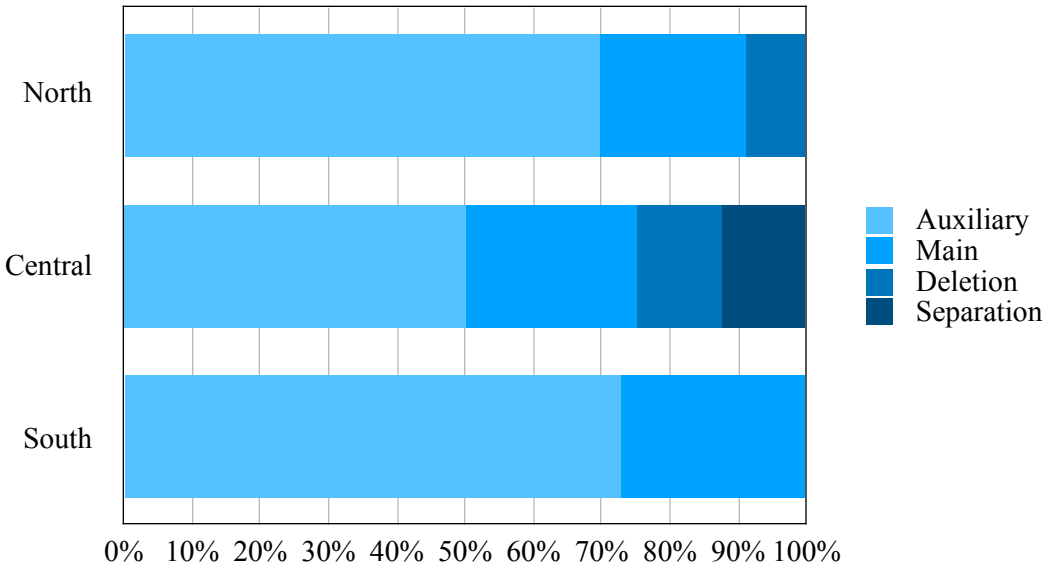
Between the NCCU Corpus and the New Corpus, there was again a decline in overall frequency but an increase in general regularity of use. The distribution pattern changed only slightly, with a small decrease in auxiliary and main verb use (approximately 4% in each case), moderate increase in deletion, and the addition of a single separation token. None of these differences were statistically significant.

Table 38. Comparison of A-not-A questions between the NCCU Corpus and New Corpus

| | NCCU Corpus | | New Corpus | | Difference | |
|---|---|---|---|---|---|---|
| | *raw freq.* | *per thousand* | *raw freq.* | *per thousand* | *log likelihood* | *log ratio* |
| Auxiliary | 50 | 1.64 | 42 | 1.17 | 2.70 | -0.49 |
| Main | 15 | 0.49 | 11 | 0.31 | 1.48 | -0.69 |
| Deletion | 2 | 0.07 | 5 | 0.14 | 0.87 | 1.08 |
| Separation | 0 | 0.00 | 1 | 0.03 | 1.22 | *undefined* |
| Total | 67 | 2.20 | 59 | 1.64 | 2.75 | -0.43 |

When comparing the regions, there were again no statistically significant differences. The distribution of A-not-A tokens was also fairly similar between regions. Monosyllabic auxiliaries made up the majority of uses in each region, followed by monosyllabic main verbs, and then bisyllabic verbs with deletion (in the North and South). The Central data also saw the use of a single separation token, making it the region with the most varied use of A-not-A questions.

Figure 12. Comparison of A-not-A question use between regions




**8.5 Discussion**

As predicted by the extended RT approach, there was a great deal of individual variation in the Old Corpus, with only two of the five speakers using A-not-A questions. However, a good deal of this can almost certainly be attributed to genre effects, as each conversation was structured as an informal interview. With this in mind, it is unsurprising that the vast majority of A-not-A tokens came from the Interviewer and that three of the four interviewees asked no A-not-A questions at all, since it is normal in interviews for only one speaker to ask questions. It cannot be concluded with certainty, therefore, that the inter-speaker variation was truly caused by the lack of an established norm for TM.

Even without the benefit of direct comparison to the Old Corpus, however, the NCCU Corpus fits very well with the extended RT predictions. A-not-A questions in that corpus were used more regularly and tended towards being used with auxiliary verbs. The deletion pattern

was applied to every bisyllabic verb used in the construction, and the class of verbs that could be used expanded to include adjectives, as in (6).

(6) Adjective A-not-A

| 印 | 錯 | 的 | 訂單 | 多 | 不 | 多? |
|---|---|---|---|---|---|---|
| yìn | cuò | de | dìngdān | duō | bù | duō |
| print | wrong | DE | order form | many | NEG | many |

"Are there many wrongly-printed order forms?"

(NCCU 4, line 512)

And while the lack of separation tokens was somewhat surprising, the fact that one such token was used in the New Corpus suggests that this absence was due to chance, rather than to a change in the characteristics of TM.

As for the New Corpus, it also largely fit the predictions of extended RT. Though there was a decline in overall A-not-A use, the pattern was used more often with bisyllabic verbs and with just as many unique non-auxiliary main verbs. Of most interest in the New Corpus, though, was the regional data, specifically, the separation A-not-A produced by the Central speaker and one of the auxiliary A-not-A questions used by a speaker from the North. The Central speaker token is unusual from the perspective of TM, seemingly double-marking the question facet of the utterance by using both the A-not-A construction and the question marker *ma* 嗎.

(7) Central speaker A-not-A

| 是 | 因為 | 那 | 個 | 老師 | 也 | 沒 | 有 |
|---|---|---|---|---|---|---|---|
| shì | yīnwèi | nà | gè | lǎoshī | yě | méi | yǒu |
| COPULA | because | that | CL | teacher | also | NEG | YOU$_{\text{ASSERTION}}$ |

| 一定 | 會 | 來 | 不 | 是 | 嗎? |
|---|---|---|---|---|---|
| yīdìng | huì | lái | bù | shì | ma |
| definitely | will | come | NEG | COPULA | Q |

"Is it because that teacher also will not definitely come?"

(Conversation #3, line 481)

However, this structure is almost identical to one sometimes used in Hakka with A-not-A questions.

(8) Hakka A-not-A

| 你 | 愛 | 唔 | 愛 | 同 | 我 | 跳舞 | 冇? |
|---|---|---|---|---|---|---|---|
| ngi$^{2}$ | oi$^{5}$ | m$^{1}$ | oi$^{5}$ | tung$^{2}$ | ngai$^{2}$ | tiau$^{5}$vu$^{3}$ | mo |
| 2.SG | want | NEG | want | with | 1.SG | dance | NEG |

"Do you want to dance with me?"

(adapted from Ager 2024)

Therefore, if the greater influence of Hakka in the feature pool of the Central region is taken into account (as it is in the extended RT approach due to incorporation of the founder principle), this superficially strange pattern becomes entirely expected.

What remains somewhat puzzling, however, is the negator used in (9) by one of the speakers from the North.

(9) Unusual negation in A-not-A

| 你 | 覺得 | 到底 | 要 | 沒 | 要 | 回 | [city B] | 啊 |
|---|---|---|---|---|---|---|---|---|
| nǐ | juédé | dàodǐ | yào | méi | yào | huí | [city B] | a |
| 2.SG | think | in.the.end | will | NEG | will | return | [city B] | P |

"Do you think you will return to [city B] in the end?"

(Conversation #2, line 38)

Outside of perfective aspect contexts, *méi* 沒 is only used as a negator for the verb *yǒu* 有; all other verbs take the negator *bù* 不. Most probably, this use of *méi* 沒 with the auxiliary *yào* 要 was just a slip of the tongue on the part of the speaker, but there is also the slight possibility that it is a sign of incipient decategorialization and the blurring of the distinction between *bù* 不 and *méi* 沒 verbs.

### 8.6 Conclusion

This chapter examined the A-not-A question construction in Taiwan Mandarin. While this feature has been given very little attention in the literature, what is discussed in early studies suggests that it is a rare case of MM serving as the model language for TM. Data from the NCCU Corpus and New Corpus indicate that A-not-A questions are slowly grammaticalizing in TM along the pathway predicted by the extended RT framework. And though it is still in the early stages of grammaticalization, the results of this study suggest that the A-not-A structure in TM may in time prove to be a valuable source of data for scholars taking an RT approach to language contact.

# Chapter 9: Conclusion

## 9.1 Summary of Findings

This dissertation sought to demonstrate that grammatical replication, rather than borrowing or polysemy copying, is responsible for the non-standard syntactic features and constructions that characterize Taiwan Mandarin. Under an extended RT approach, it was predicted that, after an initial period of significant individual variation, features that were shared by TSM and MM and/or that were of very frequent use in one variety would be selected for use in TM. These features would at first exist in TM at a degree of grammaticalization intermediate between MM and TSM, then gradually progress through the process of grammaticalization, becoming more frequent and being used in more contexts over time (extension) before losing their semantic content (desemanticization) and morphosyntactic properties (decategorialization), such as the ability of verbs to take aspectual affixes (e.g. *le* 了, *guo* 過, etc.). Loss of phonetic substance (erosion), would be the next stage in this process, though it was not predicted that any of the features examined here would have yet progressed that far. Regional differences were also predicted, with the North generally expected to exhibit a lower degree of grammaticalization than the South, and the Central region converging on either the North or South pattern depending on whether Hakka most resembles MM or TSM with respect to the specific feature under consideration. A-not-A questions were predicted to be the single exception to this general pattern, as they are a feature for which MM, rather than TSM, served as the model language, and thus the North was expected to have the highest degree of grammaticalization and the South the lowest.

The results of the corpus analysis performed in this study align with these predictions and support the validity of the extended RT approach. In the Old Corpus, which was collected approximately 30 years after the influx of Mainlanders to Taiwan, speakers vary considerably both in how frequently they use the features considered here and in how grammaticalized their uses of those features are. By the time the NCCU Corpus was collected just over 20 years later, a distinct norm had emerged, with speakers behaving much more similarly in their use of those features. While those features were used in more grammaticalized ways than in MM, they were still less grammaticalized than the TSM constructions they were modeled after, precluding a borrowing/polysemy copying analysis. In the New Corpus, which was collected approximately 15 years after the NCCU Corpus, all of those features showed signs of grammaticalization. The predicted regional differences were also present in the data. When TSM served as the model language with respect to a certain feature, the South, with its higher proportion of non-Mainlander residents, used more grammaticalized forms at a proportionally higher rate than the North. When the model language was MM, that pattern reversed and it was the North that used more grammaticalized forms. The Central region, meanwhile, exhibited the expected signs of Hakka influence, patterning with either the North and South based on whether the MM or TSM pattern was more like the Hakka one. These results support the claim of this study that using the World Englishes concepts of the founder principle and feature pool to more accurately determine the structure of language varieties in a contact situation can improve both the predictive and explanatory power of the RT framework.

## 9.2 Directions for Future Research

As with all research, this study had limitations that can and should be addressed in future research on TM syntax. Firstly, it would be beneficial to use larger corpora, as the relatively small size of the corpora used here meant that low-frequency items, such as discourse marker *shuō* and pro-verb *yòng*, were unlikely to be used more than once in each corpus. Because of this, the full range of contexts in which such features occur was likely unobserved in the current study. Secondly, though gender differences in language use were not addressed here, other research has indicated that this can be a salient consideration when examining how speakers from Taiwan structure their utterances (see Farris 1991, Kuo 2003, Su & Chang 2019). Whether these differences in preferred use patterns lead to differences in degree of grammaticalization is certainly an interesting question to consider. Another demographic factor that could not be fully taken into consideration in this study was education level, as such information was not recorded in the NCCU Corpus and the New Corpus contained only data from speakers who had completed at least some college/university. However, given the role of Mandarin in the education system on Taiwan, it is likely that the differences noted by Kubler (1981) between speakers of different education levels have persisted. The extended RT framework would expect so, since the difference in social networks would affect the inputs to the feature pools for these speakers, but of course this hypothesis needs to be tested. Finally, there are many contact-influenced features of TM that were not addressed in this study, such as directional complements, sentence-final particles, and reduplication/triplication. Any of these features would be a worthwhile focus of future language contact studies.

It is not only the TM features and demographic characteristics that this study did not cover that merit further research, however. In particular, the regional differences noted in this work would be an interesting avenue of study from both a dialectology and language contact perspective. As only the New Corpus included information about the speakers' region of origin, diachronic developments in the regional varieties had to be inferred from synchronic data, making future studies necessary in order for any truly diachronic perspective to be taken. Indeed, as an RT approach is inherently historical in nature, more studies will be needed to track the ongoing grammaticalization of the features considered here whether at the national or regional level. More broadly, incorporating World Englishes elements into RT to form the extended RT framework used here opens up the interesting possibility of approaching global varieties of Chinese in the same way the World Englishes field has approached global varieties of English.

# Works Cited


Ager, S. (2024). *Useful Hakka phrases*. Omniglot. https://www.omniglot.com/language/phrases/hakka.php

Brubaker, B. L. (2012). *The normative standard of Mandarin in Taiwan: An analysis of variation in metapragmatic discourse*. (Publication No. 3532778) [Doctoral dissertation, University of Pittsburgh]. ProQuest Dissertations Publishing.

Central Intelligence Agency. (2022, May 11). *Taiwan*. The World Factbook. https://www.cia.gov/the-world-factbook/countries/taiwan/

Chappell, H. (2008). Variation in the grammaticalization of complementizers from *verba dicendi* in Sinitic languages. *Linguistic Typology, 12*, 45-98.

Chappell, H. (2015). Say-complementizers shuō 說, waa3 話, kong1 講. In Rint Sybesma (Ed.), *Encyclopedia of Chinese language and linguistics*. Brill. http://dx.doi.org/10.1163/2210-7363_ecll_COM_00000098

Chappell, H., Ming, L., & Peyraube, A. (2007). Chinese linguistics and typology: The state of the art. *Linguistic Typology, 11*(1), 187-211.

Chappell, H. & Lamarre, C. (2005). *Grammar and lexicon of Hakka: Historical materials from the Basel Mission Library*. École des Hautes Études en Sciences Sociales, Centre de Recherches Linguistiques sur l'Asie Orientale.

Chen, P. (1999). *Modern Chinese: History and sociolinguistics*. Cambridge University Press.

Chen, P. H. (2001). Policy on the selection and implementation of a standard language as a source of conflict in Taiwan. In N. Gottlieband & P. Chen (Eds.), *Language planning and language policy: East Asian perspectives* (pp. 95-110). Curzon Press.

Cheng, R. L. (1985). A comparison of Taiwanese, Taiwan Mandarin, and Peking Mandarin. *Language, 61*(2), 352-377.

Chiung, W. V. (2001). *Language and ethnic identity in Taiwan*. Paper presented at the North American Taiwan Studies Conference, University of Washington, Seattle.

Collart, A. & Su, H.-K. (2022). Expressing the existence of an event with ‘you (to have) + VP’ in Taiwan Mandarin: A corpus-based investigation. *Concentric, 48*(2), 249-284.

Corne, C. (1995). A contact-induced and vernacularized language: How Melanesian is Tayo?. In P. Baker (Ed.), *From contact to creole and beyond* (pp. 121-148). University of Westminster Press.

Country Reports. (2022). *Taiwan Demographics*. https://www.countryreports.org/country/Taiwan/population.htm

D'Arcy, A. (2020). The relevance of World Englishes for variationist sociolinguistics. In D. Schreier, M. Hundt & E.W. Schneider (Eds.), *The Cambridge handbook of World Englishes* (pp. 436-458). Cambridge University Press.

Department of Census Directorate-General of Budget Accounting and Statistics (DGBAS). (2010). *2010 Population and housing census*. https://eng.stat.gov.tw/News.aspx?n=2400&sms=11716

Encyclopædia Britannica. (2021). *Taiwan*. https://www.britannica.com/place/Taiwan

Farris, C. S. (1991). The gender of child discourse: Same-sex peer socialization through language use in a Taiwanese preschool. *Journal of Linguistic Anthropology, 1*(2), 198-224.

Gates, H. (1981). Ethnicity and Social Class. In E. M. Ahern & H. Gates (Eds.), *The anthropology of Taiwanese society* (pp. 241-281). Stanford University Press.

Han, C. (2017). *The emergence of the YOU construction in Mandarin Chinese: The perspective of contact-induced grammaticalization*. (Publication No. 10754424) [Doctoral dissertation, National University of Singapore]. ProQuest Dissertations Publishing.

Hashimoto, M. J. (1973). *The Hakka dialect: A linguistic study of its phonology, syntax, and lexicon*. Cambridge University Press.

Heine, B. & Kuteva, T. (2003). On contact-induced grammaticalization. *Studies in Language, 27*(3), 529-572.

Heine, B. & Kuteva, T. (2005). *Language contact and grammatical change*. Cambridge University Press.

Heine, B. & Kuteva, T. (2010). Contact and grammaticalization. In R. Hickey (Ed.), *The handbook of language contact* (pp. 86-105). Oxford: Wiley-Blackwell.

Her, O.-S., Hammarström, H. & Allassonnière-Tang, M. (2022). Defining numeral classifiers and identifying classifier languages of the world. *Linguistics Vanguard, 8*(1), 151-164.

Huang, J. (2021). The semi-complementizer shuō and non-referential CPs in Mandarin Chinese. *Proceeds of the Linguistic Society of America, 6*(1), 882-895.

Huang, S. (2003). Doubts about complementation: A functionalist analysis. *Language and Linguistics, 4*(2), 429-455.

Huang, S. F. (1993). *Yuyan, shehui yu zuqun yishi* [Language, Society and Ethnicity]. The Crane Publishing Company.

Hsiau, A. C. (1997). Language Ideology in Taiwan: The KMT's language policy, the Tai-yu language movement, and ethnic politics. *Journal of Multilingual and Multicultural Development, 18*(4), 302-315.

Hsiau, A. C. (1998). *Crafting a nation: Contemporary Taiwanese cultural nationalism*. (Publication No. 9824653) [Doctoral dissertation, University of California, San Diego]. ProQuest Dissertations Publishing.

Huang, C. M. (1997). *Language education policies and practices in Taiwan: From nationism to nationalism*. (Publication No. 9819250) [Doctoral dissertation, University of Washington, Seattle]. ProQuest Dissertations Publishing.

Klöter, H. (2017). Taiwan: Language situation. In R. Sybesma, W. Behr, Y. Gu, Z. Handel, C. T. Huang & J. Myers (Eds.), *Encyclopedia of Chinese language and linguistics* (vol. 4, pp. 263-267). Koninklijke Brill NV.

Kubler, C. C. (1981). *The development of Mandarin in Taiwan: A case study of language contact*. (Publication No. 8119529) [Doctoral dissertation, Cornell University]. ProQuest Dissertations Publishing.

Kubler, C. C. & Ho, G. (1984). *Varieties of spoken standard Chinese: A speaker from Taipei*. (Vol. 2). De Gruyter.

Kuo, S.-H. (2003). Involvement vs detachment: Gender differences in the use of personal pronouns in televised sports in Taiwan. *Discourse Studies, 5*(4), 479-494.

Kuo, Y. H. (2005). *New dialect formation: The case of Taiwanese Mandarin*. (Publication No. U486873) [Doctoral dissertation, University of Essex]. ProQuest Dissertations Publishing.

Li, C. N. & Thompson, S. (1981). *Mandarin Chinese: A functional reference grammar*. University of California Press.

Li, X. (2019). *On "You + VP" construction in Mandarin Chinese*. (Publication No. 22617366) [Doctoral dissertation, University of Wisconsin-Madison]. ProQuest Dissertations Publishing.

Lim, L. (2020). The contribution of language contact to the emergence of World Englishes. In D. Schreier, M. Hundt & E.W. Schneider (Eds.), *The Cambridge handbook of World Englishes* (pp. 72-98). Cambridge University Press.

Lin, S.-F. (1974). Reduction in Taiwanese A-not-A questions. *Journal of Chinese Linguistics, 2*(1), 37-78.

Liu, C.-T. (2011). Motivations for grammaticalization: A case study of the *realis* marker YOU (有) in Taiwan Mandarin. *The 12th Chinese Lexical Semantics Workshop (CLSW-12)*, 42-49.

Liu, M. & Xu, R. (2013). *YONG* 用 as a pro-verb in Taiwan Mandarin. In D. Ji and G. Xiao (Eds.), *13th Chinese Lexical Semantics Workshop (CLSW 2012) Revised Selected Papers* (pp. 540-550). Springer.

Mo, R. P. J. (2000). *Taiwan on the brink of reversing language shift: Its current development and contributory factors.* (Publication No. 3018248) [Doctoral dissertation, Purdue University]. ProQuest Dissertations Publishing.

Mufwene, S. S. (1996). The founder principle in creole genesis. *Diachronica, 13*(1), 83-134.

Mufwene, S. S. (2001). *The ecology of language evolution*. Cambridge University Press.

Mufwene, S. S. (2002). Competition and selection in language. *Selection 3*(1), 45-56.

Mufwene, S. S. (2020). Population structure and the emergence of World Englishes. In D. Schreier, M. Hundt & E.W. Schneider (Eds.), *The Cambridge handbook of World Englishes* (pp. 99-119). Cambridge University Press.

Schreier, D. (2020). World Englishes and their dialect roots. In D. Schreier, M. Hundt & E.W. Schneider (Eds.), *The Cambridge handbook of World Englishes* (pp. 384-407). Cambridge University Press.

Scott, M. & Tiuⁿ H. K. (2007). Mandarin-only to Mandarin-plus: Taiwan. *Language Policy, 6*, 53-72.

Su, Y. & Chang, Y. (2019). Intra-lingual pragmatic variation in Mandarin Chinese apologies: Influence of region and gender. *East Asian Pragmatics, 4*(1), 59-86.

Sullivan, J. P. (1980). The validity of literary dialect: Evidence from the theatrical portrayal of Hiberno-English forms. *Language in Society, 9*(2), 195-219.

Szeto, P. Y. (2019). *Typological variation across Sinitic languages: Contact and convergence*. [Doctoral dissertation, The University of Hong Kong]. HKU Theses Online. http://hdl.handle.net/10722/279263

Thomason, S. & Kaufmann, T. (1988). *Language contact, creolization, and genetic linguistics*. University of California Press.

Trudgill, P. (1986). *Dialects in contact*. Basil Blackwell Publishing.

Tsao, F. F. (2008). The language planning situation in Taiwan: An update. In R. B. Kaplan & R. B. Baldauf Jr. (Eds.), *Language planning and policy in Asia: Japan, Nepal, and Taiwan and Chinese characters* (vol. 1, pp. 285-300). Multilingual Matters.

Tse, J. K. P. (1986). Standardization of Chinese in Taiwan. *International Journal of the Sociology of Language, 59*, 25-32.

Wang, Y.-F., Katz, A. & Chen, H. (2003). Thinking as saying: shuo ('say') in Taiwan Mandarin conversation and BBS talk. *Language Sciences, 25*, 457-488.

Wang, Y.-F., Tsai, P.-H. & Ling, M.-Y. (2007). From informational to emotive use: meiyou ('no') as a discourse marker in Taiwan Mandarin Conversation. *Discourse Studies, 9*(5), 677-701.

Wei, J. M. (2006). Language choice and ideology in multicultural Taiwan. *Language and Linguistics, 7*(1), 87-107.

Wei, J. M. (2008). *Language choice and identity politics in Taiwan*. Lexington Books.

Weinreich, U. (1953). *Languages in contact: Findings and problems.* Linguistic Circle of New York.