Probing into the Historical and Geographical Variants of Mandarin: A Computational Approach

Annie Chen

Follow this and additional works at: https://scholarworks.umass.edu/masters_theses_2

Part of the Chinese Studies Commons

Recommended Citation
https://scholarworks.umass.edu/masters_theses_2/624

This Open Access Thesis is brought to you for free and open access by the Dissertations and Theses at ScholarWorks@UMass Amherst. It has been accepted for inclusion in Masters Theses by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
Probing into the Historical and Geographical Variants of Mandarin:
A Computational Approach

A Thesis Presented

by

ANNIE HONGJIE CHEN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of
Master of Arts

May 2018

Chinese
Probing into the Historical and Geographical Variants of Mandarin:  
A Computational Approach  

A Thesis Presented  

by  

ANNIE HONGJIE CHEN  

Approved as to style and content by:  

Zhongwei Shen, Chair  

Amanda C. Seaman, Member  

David K. Schneider, Member  

David K. Schneider, Director  
East Asian Languages and Cultures  

William Moebius, Chair  
Department of Languages, Literatures, and Cultures
ACKNOWLEDGEMENTS

For the past two years studying in the Chinese program housed in the East Asian Languages and Cultures department at the University of Massachusetts Amherst, I owe a debt of gratitude to the faculty members and staff for a memorable graduate experience. The professional and supportive environment facilitates Master's studies and encourages academic pursuits. The Chinese program has a well-rounded curriculum and a group of well-established scholars with versatile specialties to offer a solid fundamental training in Sinology. Although my research project was on Chinese linguistics, I personally enjoyed courses on modern Chinese literature, classical Chinese literature, and ancient Chinese philosophy to deepen my appreciation of the literary body of Chinese on a full-spectrum. The classes I took with Professor Elena Suet-Ying Chiu, Professor David K. Schneider, and Professor Enhua Zhang were full of valuable insights of the Chinese culture and civilization. I also gained a good understanding of second-language acquisition and Chinese grammars thank to Professor Zhijun Wang.

On my master thesis project probing into the development of the Chinese language, I cannot have enough good words to express my respect to Professor Zhongwei Shen for his bottomless knowledge and tremendous patience in guiding me all the way through the finish line. As the Graduate Director who responded to all graduate students' needs and questions diligently, Professor Shen still made the effort and managed to advise me on this project closely and energetically on top of his own research writing contract and the class teaching obligation. With their warm-hearted encouragements and brilliant suggestions, Professor Shen and my other two committee members, Professor David K. Schneider and Professor Amanda C. Seaman, constituted the best work triangle of committee that I could ever ask for. Their support functioned as the cornerstone for me to carry on, especially, during the time when I was preparing for the defense while my father just passed away.

The Chinese program at UMass Amherst will always be the home to my journey of Sinology. My experience here will keep nourishing and inspiring me on the road of academic exploration for the rest of my life. On my faith in academics, I will always remember where it all begins ---- the molded quality of discipline and high-achieving back in the years of honors college, and the genuine love for and endless curiosity of my own heritage flourished over the past two years of the Master in Chinese. I would like to end my acknowledgments with the following quote from Qian Mu's preclude, being my long-time inspiration, to his milestone work outlining the history of China.

“凡读本书请先具下列诸信念：一、当信任何一国之国民，尤其是自称知识水平线以上之国民，对其本国以往历史，应该略有所知。二、所谓对其本国以往历史略有所知者，尤必附随有一种对其本国以往历史之温情与敬意。三、所谓对其本国以往历史有一种温情与敬意者，至少不会对其本国以往历史抱有一种偏激的虚无主义，二将我们当身种种罪恶与弱点，一切诿卸与古人。四、当信每一国家必待其国民具备以上诸条件者比较渐多，其国家乃再有向前发展之希望。”

---- 钱穆《国史大纲》
ABSTRACT

PROBING INTO THE HISTORICAL AND GEOGRAPHICAL VARIANTS OF MANDARIN:
A COMPUTATIONAL APPROACH

MAY 2018

ANNIE HONGJIE CHEN, B.A., BOSTON COLLEGE
M.A., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Zhongwei Shen

This computational study reveals the primacy of language contact in the variation of language (Sarah Grey Thomason 2003). The visualization and further analysis confirm the reconceptualization of Chinese linguistic history with the theory of Horizontal Transmission (Shen 2016). Horizontal Transmission situates the development of Mandarin and other Chinese dialects in a sociopolitical landscape as a cultural complex and introduces imperfect learning to the time-capsulated process of Language Shift as an inevitable social phenomenon.

The nature of language largely determines how it can change (Janda and Joseph 2003). We have to ruminate on the fact that the grammar of language is a symbolic system of representation while living language is a complex adaptive system generated and regenerated by individuals (Shen 2015).
The descriptive capacity of Shen’s theory is compatible with the nature of language being dynamic idiolects alongside a real linguistic history embodied by individual speakers in time and space. Only by understanding the change mechanism of Chinese from the perspective of language contact and through the lens of language shift, the variation of Mandarin and emergence of Chinese dialects find their explanations in a salient chain of logic to create a holistic account of Chinese evolution where the intertwined influence of languages finds its manifestation.
# Table of Contents

ACKNOWLEDGEMENTS......................................................................................................... iv
ABSTRACT.................................................................................................................................. v
LIST OF TABLES...................................................................................................................... viii
LIST OF FIGURES................................................................................................................... ix

## CHAPTER

I. INTRODUCTION ................................................................................................................ 1
   1.1 Overview: Content and Purpose.............................................................................. 1
   1.2 Existing Problems..................................................................................................... 3
      1.2.1 Object of Study.............................................................................................. 3
      1.2.2 Mechanism of Language Change................................................................. 6
      1.2.3 Formation of Chinese Dialects..................................................................... 11
   1.3 Theoretical Framework: Language Shift and Horizontal Transmission.............. 13
   1.4 The Origin of Mandarin......................................................................................... 18

II. RESEARCH DESIGN.......................................................................................................... 24
   2.1 Source Materials and Research Method............................................................. 24
   2.2 Twenty Phonological Features............................................................................. 30

III. NEIGHBOR-NET AND TABLE OF COHESIVENESS.................................................. 39
   3.1 The Network of Mandarin Closeness................................................................... 39
   3.2 Cohesiveness of Chinese Variants....................................................................... 42

IV. CONCLUSION.................................................................................................................... 50
   4.1 Research Findings................................................................................................. 50
   4.2 Horizontal Transmission....................................................................................... 51
      4.2.1 The Primacy of Language Contact.............................................................. 51
      4.2.2 A Mesohistory of Chinese............................................................................ 54

## APPENDICES

I. TWO BASIC CHINESE DIALECTAL MAPS....................................................................... 58

II. MATRIX OF INTERESTING ISOGLOSSSES.................................................................... 60

III. NEIGHBOR-NET OF TWENTY-EIGHT CHINESE DIALECTAL VARIETIES................. 61

IV. NUMBER OF COHESIVENESS..................................................................................... 62

BIBLIOGRAPHY..................................................................................................................... 65
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1: Cohesiveness of Original Mandarin and Its Varieties</td>
<td>44</td>
</tr>
<tr>
<td>2.2: Mandarin Sub-dialect Chengdu and Other Chinese Dialectal Varieties</td>
<td>47</td>
</tr>
<tr>
<td>2.3: Original Mandarin Variety and Non-Mandarin Varieties</td>
<td>48</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------</td>
</tr>
<tr>
<td>1.1</td>
<td>Language Shift Model of Horizontal Transmission</td>
</tr>
<tr>
<td>2.1</td>
<td>The Closeness Web of Mandarin Varieties</td>
</tr>
</tbody>
</table>
CHAPTER I
INTRODUCTION

1.1 Overview: Content and Purpose

This study attempts to understand the linguistic history of Chinese from a perspective that situates language in time and space with a computational demonstration. Instead of counting on the similarities to reconstruct genealogy, the closeness of dialectal varieties, diachronically and synchronically, will come under the focus of this study to bring an unconventional light onto the development of Chinese. Sampling Chinese dialects, twenty-five language varieties of twenty-five major cities across the territory of China will be investigated. Another three ancient Mandarin varieties roughly in the same geographical area, centered around modern city of Beijing, for three succeeding dynasties about a thousand years ago make the total investigated language varieties twenty-eight. Twenty diagnostic features will be compared for these twenty-eight Chinese language varieties with a computational application.

The computational grouping and numerical data give rise to our later analysis that confirms two hypotheses: a) If we compare the closeness between diachronic Mandarin varieties and synchronic Mandarin varieties, we observe that the former is closer than the latter; then, we say that space is a more critical factor than time when it comes to Mandarin variation; b) If we can generate numerical evidence demonstrating that Mandarin, in the same region over a considerable amount of time, varied less than it did across China, then we can conclude confidently that the development of Mandarin is due to language contact. It is worthy to note that the representative dialectal varieties selected for this study actually include non-Mandarin varieties as well, which gives us the full
picture of Chinese variation on a macro-scale in the diagram.¹

To demonstrate the closeness of the investigated dialectal varieties, we use the computational method employed by linguists such as Alexandre Francois (2015) to visualize with NeighborNet. The net diagram as a whole gives us the visualization of the map of Chinese dialects in terms of closeness, which matches our overall perception of Chinese dialectal classification.² Besides, numbers of Cohesiveness between each pair of dialectal varieties facilitate our comparative analysis of their closeness. Interesting comparisons reveal that dialectal classification is categorical after all and that the reality of Chinese variation is dynamically complex. Demonstrating our linguistic evidence with crystal clarity, the computational method and mathematical embodiment confirm our two hypotheses of this study. At the same time, they support valuable insights for the mechanism of Chinese variation. Readings of the diagram and numerical tables construct explanations that rigorously refine a Chinese variation mechanism and naturally lead to Language Shift Model of Horizontal Transmission (Shen 2016).

In reconceptualizing the development of Chinese against the real linguistic history of China, Zhongwei Shen proposed a ground-breaking theory of Horizontal Transmission. Inspired by analysis of areal linguistic features, Horizontal Transmission acknowledges the primacy of language contact in language change and situates the

---

¹ Please see Appendix III for a full diagram of Chinese dialectal NeighborNet illustrating the closeness of the twenty-eight Chinese language varieties in terms of the twenty diagnostic features in this study.
² The classification of Chinese dialect groups has its own historical development of standards. As Bangxin Ding proposed in his milestone paper in 1982 on how to classify Chinese dialects, it is widely accepted by today's Chinese linguists that early historical innovations distinguish major dialect groups, late historical innovations distinguish sub-dialects within a dialect group, and modern differences rather than historical innovations distinguish newly evolved dialects.

"以汉语语音史为根据，用早期历史性的条件区别大方言；用晚期历史性条件区别次方言，用现在平面性的条件区别小方言。”（《汉语分区的条件》丁邦新， 1982）Mandarin, Jinyu, Wuyu, Huiyu, Ganyu, Xiangyu, Minyu, Kejia-hua (Hakka), Yueyu, and Pinghua are the ten conventional Chinese dialect groups. For a visualized Chinese classification on the geographical map of China, please refer to the first map in Appendix I.
formation of Chinese dialects in the sociopolitical landscape of Chinese history. Shen spotted influence of non-Han Chinese on Han-Chinese and correlated development of Chinese dialects. Alongside contemplation over the historical cultural complex of China and recognition of the time-capsulated, thus gradual, process of linguistic variation, Shen introduced imperfect learning at different levels of language standardization into the mechanism of Language Shift in Chinese variation. Moreover, Horizontal Transmission responds to the true nature of living language as a complex adaptive system where idiolects carried by individuals organically evolve (Shen 2015).

The ultimate purpose of this study is for a deep understanding of the history of Chinese from a synchronized viewpoint of culture and language. Along the way, we will touch on issues related to the essence of language and the perception of language change in historical linguistics in order to reflect what constitutes a meaningful study of language in change. Of course, we focus on the development of Chinese particularly. We want to both find the necessity of a theoretical revolution in the current field of Chinese dialectology and reinforce the resolution of Horizontal Transmission in advancing the field's conceptualization. In such an endeavor, may this study invite future research on Chinese dialects from the perspective of Horizontal Transmission to bring about exciting discoveries of Chinese linguistic history and findings to settle controversial debates of Chinese linguistics.

1.2 Existing Problems

1.2.1 Object of Study

Historical progress is linear only if we connect the dots of events on a timeline; however, real history evolves against a complex web of events taking place in different locales, and the development of history has far more empty room to be filled with
imagination and rationale beyond a timeline. What we call history is embodied by human activities in time and space. Other than being pinned down on a timeline for simplified comprehension, history does not progress in a linear fashion. As William S.Y. Wang puts it, almost all human systems in history advance in a non-linear fashion, like the Brazilian butterfly causing a Tornado in Texas. He points out that “Human culture is a complex adaptive system, and so is human language.”3 Sharing this same observation of language, Zhongwei Shen spent the past decade in expanding the horizon of research in the academic field of Chinese dialectology. He first re-examined language change by taking language from the perspective of a complex adaptive system (Shen 2015). This perspective highlights language as an ensemble of idiolects constantly interacting with each other and adapting to the norm. Only by situating language change in its real vehicle, idiolect, can the study of language change be compatible with real linguistic history. Through such a dynamic lens, Shen provided the field valuable and inspiring insights into linguistic evidence that had long been there for analysis but had been taken for granted. He asked and answered over-looked questions, which eventually re-conceptualized the mechanism of Chinese variation (Shen 2016).

Since the year of 2015, Shen has worked on reforming the theoretical framework for the historical study of Chinese and the data analysis of modern Chinese dialectal varieties. From his major theoretical publications, we can see a pattern of emphasizing the importance of acquiring both an in-depth understanding of language and language in change alongside the cultural history of China before designing and conducting any Chinese linguistic research at all. It calls for a fundamental reshaping of the object of

---

3 Wang made comment on language being a complex adaptive system on a discussion panel led by him in the International Conference of Language and Human Complex Adaptive System at Peking University in 2014. Notes organized from this panel became published literature in 2017.

(《语言与人类复杂系统第一讲：复杂系统与音节语言的形成》王士元主讲: 19)
study in the field and a revolution in methodologies of analyzing and collecting linguistic evidence. This is particularly essential for diachronic studies of language whose object of analysis is living language rather than the symbolic structure of language as in synchronic studies.

The common sense of understanding the subject of research before even starting to research has not received enough attention in the field of Chinese Linguistics. In the academic field of Linguistics in general, the methodologies of synchronic study and diachronic study of language have had little reflection upon a justification in accordance with their respectively different subjects of study. The subject of study for synchronic linguistics is the symbolic entity of any language in its relatively stable stage of variation. Thus, the ultimate quest is the mapping from sound to meaning. The linguistic universal of synchronic studies is the linguistic system of representation. Diachronic linguistics, on the other hand, is all about language in variation and in history. Thus, the object of diachronic studies of language should assume a different point of view acknowledging the dynamics of living language. The dynamics of living language lie in human communication.

4 Such an understanding of the object of synchronic linguistic studies was first systematically elaborated by Ferdinand de Saussure, the father of structural linguistics. He made famous metaphors comparing language, as a symbolic entity, to a virtual dictionary shared by community speakers or a piece of symphony played by different musicians. Noted by Saussure, the symbolic reality of language can be played out differently by the user of the dictionary or the musician performing the notes. He made a distinction between langue (language) and parole (speaking), where speech is external to the internal symbolic entity.

5 The definition of language has many versions. However, from a survey of language definitions, we find that common ground is in communication. As S.I. Hayakawa and Alan R. Hayakawa (1939) put it, language is used to communicate, facilitating human cooperation for survival purposes. Wenguo Pan (2001) concludes, after almost exhausting definitions of language in history, that language is essentially a human process.

6 Natural language is human spoken language. Although sign languages do arise spontaneously, they are not defined as natural language in linguistics. Constructed languages like Esperanto, having estimated two million speakers worldwide, do not fall under the definition of natural language either.
structural linguistics, the object of diachronic linguistic studies has a default problem. How can we examine living language properly if human speech, the very loci of living language, is considered external to language? Edward Sapir, in his pioneering monograph *Language: An Introduction to the Study of Speech*, points out that “language is a purely human and non-instinctive method of communicating ideas, emotions, and desires by means of a system of voluntarily produced symbols” (Sapir 1912: 8). Sapir makes a profound and clear point in his book title; human speech is not peripheral to the study of language, but central. Going back to the point made by Shen when taking language as a complex adaptive system, language evolves as an organic ensemble of idiolects that intelligently respond to each other and gradually coordinate a direction of change. Shen correctly asserts that “language always exists in idiolects” (Shen 2015: 169). Individual speeches and the interaction among them are essential to the study of living language and its evolution in history. However, with the preoccupation of speech being external to language, language contact is considered an external linguistic phenomenon, and thus it becomes the external factor for language to change.7

1.2.2 Mechanism of Language Change

In exploring language in variation, diachronic linguistic studies mainly have put the focus on comparing similarities among today's languages to align with the working principle of the Comparative Method. Through the Method, a Stammbaum (Schleicher 1860) or a family tree graphically represents the relatedness among languages. Invented and rooted in studying Indo-European languages, the Tree Model and the Comparative

---

7 Summarized by Canadian historical linguist Mark Hale, there are five major conventional factors contributing to language change: 1). Directionality limitations on change events, 2). Triviality indices of change events, 3). Probability of borrowing, 4). Chance, 5). Iconic similarities. The first two are of the internal factor, and the rest are of the external. External factors are credited to language contact.
Method were traditionally practiced by historical linguists. When directionality limitations apply to the change events of language or other types of evidence showing favors of an internal change, historical linguists will then apply the Method to examine similarities among daughter languages. (Hale 2014: Pg) These similarities are thought to be contributed by internal factors of language rather than external. Since only external factors are considered to have their trigger in language contact, the Method by design excludes language contact from forming language family trees. Taken as the mechanism of linguistic change, the Tree Model thus credits the divergence of linguistic branches to the internal change of language. In other words, internal diversification of language becomes the cause of language change. Consequently, the linguistic history is simplified into a procedure of vertical diversification. With this mechanism, reconstructing historical phonologies and assessing the validity of language classification refer to the simplified linguistic history in a vertical and linear fashion, which becomes mainstream discourse in diachronic linguistic studies. With the Method, daughter languages offer clues in tracing intermediate parent languages and eventually lead to the ultimate parent language, a proto-language. Throughout this process of reconstruction, linguistic features shared by daughter languages construct the grammar of their historical varieties, and “the internal structure of a language family” reveals itself in a tree. In turn, the Tree claims to mimic the development of language and to visualize language classification.

Although the Tree does inform language history, it lacks the descriptive capacity of real linguistic history. Neither can the Tree represent the historical development of language. In fact, a close examination of the Method and the Tree Model reveals that a reconceptualization of linguistic variation mechanism is in demand. First of all, it is

8 “Such a tree captures a set of claims about the internal structure of a language family.” (Francois 2014: 163)
problematic that the Method only takes on linguistic similarities to construct language
history. Similarities among daughter languages already are selective characteristics
favorable to the paradigm of single proto-language and its vertical diversification. In
fact, the difference among daughter languages very likely will link them to multiple
ancestral languages and thus establish a complete profile of parent language. The
Method has a fallacy of slanting by calculating similarities alone and claiming to
reconstruct the complete genealogical history of daughter languages. The Method's
paradigm of a single proto-language and its default change mechanism of vertical
diversification are to be blamed for the model's theoretical fallacy. Secondly, since the
Method presumes that the cause of language change is internal, along with the slanting
fallacy of selecting linguistic similarities, it already has presupposed a tree. In other
words, the Tree Model by nature is demonstrative, not in the sense of proving but
displaying a subjective projection of language history based on the graphic representation
of a tree. This is to say it is arguing in a circle by using the Tree Model to reason
language change having a variation mechanism of internal vertical diversification.

The linguistic evidence used in the Method mainly is phonological, and natural
sound change occupies the center of the stage where internal vertical diversification of
language takes place. Physiological capacity of human acoustics in producing and
directing natural sound change generates energetic debates in current Chinese linguistic
discourse. In his milestone three volumes on linguistic change, William Labov (1994)
has a volume contributed to the constraints problem of internal language change
(Weinreich, Labov, and Herzog 1968). Labov gives a decent summary and discussion on
various theoretical developments of internal language change in the volume, including
the regularity hypothesis, mergers and splits, chain shifts, and functionalism. As a
possible fundamental reason behind all of these, natural sound change ultimately serves as the cause of internal language change. The argumentation of vertical diversification as the linguistic variation mechanism builds on the assumption that natural sound change is the internal factor of language change in contrast to that language contact is the external factor of language change. In other words, the mechanism of vertical diversification excludes any change caused by language contact. It is critical to point out that the tendency of easy utterance behind natural sound change, if there is any, is not in language as a symbolic system, but in human speech. No language can process a natural sound change by itself. By definition, natural sound change is a linguistic dynamic actualized by individuals, the carrier of idiolects. In this sense, language has no internal facet to change per se. As a matter of fact, it is not legitimate to bi-sect the factor of language change into internal and external at all. Even if biological evolution does push humans into lazy ways of making sounds, there is nothing internal of easy utterance in causing a linguistic variation. The easy utterance needs to compete to substitute the difficult one in contact so that the “natural” outcome of such competition has a chance to register in the linguistic variation. As we can see, the legitimacy of the Method again is called into question for crediting internal language change and negating language contact in shaping daughter languages at the root.

Reconstructing a one and only proto-language for every language family presents another logical fallacy. Conventionally, daughter languages A and B share similarities of parent language C so that they reconstruct language C to be the parent language, and historical linguists take the Tree to say parent language C diverts into different languages on her own. The logical form of this reasoning is fallacious by taking an inverted induction as a deduction. As a result, it leaves out the vital possibility of other languages
contributing to the formation of daughter languages, and the number of parent languages is largely reduced. Not only does the quantity of parent languages fall short of reality, but also does the linguistic universal of proto-languages become far off the reality of ancestral linguistic varieties in history. As in a current linguistic definition, a language has systematic varieties called dialects. However, with the Tree Model, the reconstructed proto-language cannot have dialects, thus is not consistent with this fundamental descriptive characteristic of a language. Mark Hale (2014) tried to resolve this theoretical deficiency of the Tree Model by accommodating the proto-language to be an individual language rather than areal. He argues that taking the proto-language as a reconstructed idiolect gives us an individual grammar. An individual grammar is not a language, thus it does not have varied forms. Hale gave a good try to defend the Tree Model by re-conceptualizing proto-language. However, the reconceptualization justifies proto-language in accordance with general linguistic universal by sacrificing a salient chain of reasoning. Areal daughter languages are grouped idiolects; they have no legitimate reason to reconstruct an individual grammar. After all, who's idiolect is the proto-language? The reconceptualization of proto-language being an idiolect is like winning a battle by losing the war.

In general, the proto-language reconstruction is to recover the ancestor language and to recreate the genealogy of language family. The Tree Model is a categorical representation of language relatedness. Language relatedness is not language evolution because relatedness does not speak for the causality of change. No causal development is embodied in the Tree; the Tree is incapable of reflecting the mechanism of linguistic variation or describing the real linguistic history. The categorical is helpful for us to understand the object of study, language, in abstract and simplified terms; but the
categorical is not the reality of language, not to mention the model's logical fallacies in representing linguistic reality. As a conceptual representation, the Tree Model is not universally inclusive to open up a theoretical space for the complex adaptive dynamics of individual speeches. The Tree is just a profile of language, which means it is informative of linguistic history, but is not in full power of describing linguistic history.

1.2.3 Formation of Chinese Dialects

Dialects are “mutually intelligible forms of a language” (Fromkin and Rodman 1974: 276). However, Chinese dialects (方言) literally means local speeches in Chinese, and Fangyan is a political umbrella term for language varieties in China. With “one common system of orthography” (Y.R. Chao 1976: 24), the Chinese writing system is supra-dialectal, but the concept of a unified Chinese writing system should not overshadow the difference in Chinese language. In the first scholastic and survey literature of Chinese dialects, Xiong Yang (53 B.C. - 18 A.D.) listed synonymous lexical items followed by their geographical occurrences. Such an awareness of regional dialect essentially is an illustration of language contact. Different pronunciations of the same lexical term co-exist in a region, and local speeches are making contact.

Xiaofan Li and Mengbing Xiang make the observation of language difference in their widely-used textbook of Chinese dialectology to be the result of language regenerating through language use (Li and Xiang 2009: 10). Created and recreated in

---

9 Sub-dialects of a dialect in China may fit in the definition of mutual intelligibility, but Chinese dialectal groups are by large mutually unintelligible. Sub-dialects of a dialect in China can be mutually unintelligible as well. In other words, we think of distinctive grammars when we talk about Chinese dialects.

10 The political means of “书同文”, making the writing system equivalent to language, in China helps the centralization of power by standardizing and facilitating the communication of bureaucracy.

11 “Generally speaking, the Fangyan gives the set of synonymous lexical items followed by the occurrences of the listed words in various geographical areas. On the basis of the regions mentioned one can put the dialects in groups.” (“Chinese Dialects in Time and Space” Chin-Chuan Cheng, 2005)
human application, language lives with humans and dies with humans in their view. With such an emphasis on language usage, they endorse the Wave Model (1872) developed by Johannes Schmidt to complement his teacher August Schleicher's Tree Model in explaining dialectal diffusion of a language. Based on the linguistic evidence of Indo-European languages, the Wave Model acknowledges language contact in language variation, but in complementing the Tree Model, it does not challenge internal language change or respond to the cause of splits on the Tree. Without involving causality, the Wave is another model in describing linguistic phenomena, not in revealing the mechanism of language change. Taking Indo-European linguistic theoretical framework without giving it a second thought, Chinese dialectologists explain the formation of Chinese dialects from either the internal language change perspective or the external. Both internal vertical diversification and external language contact are credited for the formation of Chinese dialects. As Rulong Li put it, “the formation of Chinese dialects is the result of divergence or migration (of Chinese) in history” (Li 2003: 5).

No matter how obvious the integrity of language and humanity is in these scholars' own words, language still comes to their minds as a conceptual notion rather than a historical reality. In the diachronic study of Chinese, humans making speech contact and idiolects carrying linguistic variations never have received the mainstream attention they deserve for situating the formation of Chinese dialects in the real history of Chinese variation. The variation of Chinese has been confined by the Tree into the development from a single proto-Chinese, and thus the formation of Chinese dialects is the diversification of Chinese (You 1992:1). Although scholars have voiced on the influence of non-Han Chinese on forming Chinese dialects (Pan 2004; Chen B. 2005; Shen 2007), in response to the Chinese language's own diversification, historical
stratification of Chinese and classification of Chinese dialects dominate the research field of Chinese dialectology. With today's vast digital data of Chinese dialects and for future extensive fieldwork of Chinese dialects, Chinese dialectology is in urgent need of an indigenous linguistic model appropriating to the formation of Chinese dialects.

1.3 Theoretical Framework: Language Shift and Horizontal Transmission

As previously discussed, the Tree takes the cause of language change for granted and gives us a subjective language history that credits the vertical diversification of language from a single proto-language. This framework of language change is insufficient for our research on Chinese dialects in real history. We are trying to design a study where language is the ensemble of idiolects constantly in the interaction carried out by individual speakers. We believe that the nature of language largely determines how it can change (Janda and Joseph 2003). From this point of view, we then have to ruminate on the fact that language grammar is a symbolic system of representation while living language is a complex adaptive system generated and regenerated by individual speakers (Shen 2015). Thus, we want to probe into linguistic variations and treat language change as they are the dynamics of a complex adaptive system.

In Shen's work (2016) on proposing his theory of Horizontal Transmission to be the main mechanism of Chinese variation, he examined areal linguistic features with valuable insights. Shen led us to focus on the reason why non-Han Chinese phonological features present in regional modern Chinese dialects, and Middle Chinese (MC) phonological features present in non-Han Chinese. For instance, the Yue dialect group preserves all tonal categories of Middle Chinese with the phonological condition of having a contrast in vowel length. Contrast in vowel length is an iconic feature of the ethnic Zhuang language spoken in the geographical region of Yue historically (Shen
In other words, the phonological condition that helps preserving MC tonal categories in the Yue dialect group which may have connection to the phonology of the Zhuang language.

Yue is a southern dialect group, and Wu is a south-eastern dialect group. Wu also displays a similar phonological pattern with a regional ethnic language. Without the three-way contrast of stop codas, [-p], [-t], and [-k] in MC phonology, Wu parallels with the ethnic language of Miao around the area in terms of having a weak contrast for the ending\(^\text{12}\) in the structure of a Chinese syllable (Shen 2016: 26). Shen considered the preservation and loss of MC phonological features in modern Chinese dialects in the light of non-Han languages spoken by ethnic Chinese historically habituated or still habituated around the area. His panoramic way of reasoning the evidence of Chinese variation brought about reflection over the role of language contact, particularly in the context of language transferring through speakers, playing in the linguistic development of Chinese. The historical migration of Chinese people into non-Chinese territories is not just a signature event in Chinese history, but also critical in shaping the linguistic history of Chinese.

When talking about history, factors of time, space, and people compose an analytical trinity of consideration. The same consideration should apply to language history as well. Constant contact is common and natural of language in the history of Chinese. As grouped idiolects, languages first make contact to each other through individuals, who are the vehicles of idiolects. In other words, language change is a

\(^{12}\) The first layer in the structure of a Chinese syllable is tone and syllable without the tone. The second layer divide the syllable into initial and final. The final can divide into medial and rhyme. The rhyme is composed of vowel and ending. Initial, medial, vowel, and ending are abbreviated to be IMVE. IMVE is the basic structure of a Chinese syllable because these four structural sloths in a Chinese syllable will either have one phoneme or zero.
linguistic phenomenon caused by idiolects transmitting through people across a certain territory over a time span.¹³ Heavily influenced by linguistic theories and models abstracted from and conceptualized for Indo-European languages, Chinese linguistic specialists have long been lacking a conceptual guidance and a theoretical framework to serve as the pivotal point of reference for conducting Chinese linguistic research in the context of Chinese linguistic history. Shen proposed a reconceptualization of Chinese variation, introducing language shift, rather than vertical diversification into the formation of Chinese dialects.

¹³ “Language is human speech, thus language change has to do with human activities in time and space.”
“语言是人的语言，语言变化和人在时间和空间中的活动有关。” (Shen 2016: 21)
In his theory of Horizontal Transmission, L1 makes contact with L2 and produces a modified L1. In the time-capsulated process of language shift, the acquisition is never perfect, and the culturally robust L1 becomes a varied version that is a result of L2 speakers imperfectly learning L1. Consequently, linguistic features of L2 will present in L1'; and, L1' will either preserve or lose features of L1, depending on the linguistic compatibility of L2 to L1. Shen's model of Chinese variation completely steps aside from the variation mechanism of vertical diversification. The Language Shift Model of Horizontal Transmission is especially conceptualized for the variation of Chinese. In the case of Chinese, grouped idiolects carried by Han people, migrating along a sociopolitical
landscape making contact with non-Han languages, gradually scattered into different geographical areas of China and formed varieties of modified Chinese in replacement of indigenous local speeches. Another aspect of Horizontal Transmission in conceptualizing the development of Chinese is the possibility where various varieties of Chinese in turn engaged in language shift. For instance, the Mandarin sub-dialects are not just products of vertical diversification of Mandarin itself, and our study computationally generates numbers for comparison analysis revealing astonishing linguistic closeness of Mandarin sub-dialects to non-Mandarin dialects. It is worthy to note that the dynamic of language shift is complexly adaptive, but the final outcome of language shift is directional (Labov 1994). This is because, on the macro-scale, language that is more socially prominent tends to replace other languages in contact. As a result, the socially prominent language becomes modified differently and repeatedly because of imperfect learning\textsuperscript{14} at different stages of language diffusion.

As a robust culture in the history of China, Han Chinese promoted their language, and their language was modified by ethnic languages in the process of Sinicization (Shen 2016: 33). “A New Account of the Tales of the World” by Yiqing Liu (403 - 444 A.D), during the time of the Southern and Northern Dynasties, contains a historical anecdote showing Sinicization around the Han and Wei-Jin periods. In the anecdote, a local official spoke a Non-Han language in a region where a form of Chinese dialects prevails today. In the time-encapsulated process of Sinicization, Non-Han people learned the language of Han people as a necessity of standardization. However, their learning would never be perfect; features in their first languages were carried into their learned Chinese.

\textsuperscript{14} Imperfect learning gives us “Interlangauge” which is the novel language variety imprinted with the influence of languages in contact. Weinreich introduced the linguistic phenomenon of interference as “those instances of deviation from the norms of either language which occur in the speech of bilinguals as a result of their familiarity with more than one language, i.e. as a result of language contact” (1968:1).
Phonological features of their first language either helped to preserve the MC phonological system or to substitute and cause loss of certain phonological contrasts. Consequently, varied forms of Chinese, displaying impact of ethnic Chinese, became varieties of 方言 Fangyan, Chinese dialects.

Horizontal Transmission is both indigenous to Chinese variation and realistic to Chinese history. It best captures the gradual process of language change in time and space through people making contact. To quote Zhongwei Shen, “all Chinese dialects can be seen as simplified forms of Middle Chinese for MC phonological features show up in modern Chinese dialects more or less.” Although modern Chinese is a variety of Middle Chinese, the vertical diversification of Chinese is an illusion. This is because vertical diversification of a language is a manifestation of the timeline of horizontal transmission in reality. The mechanism of change is still horizontal. Vertical diversification is informative but not essential in describing language change and variation. In the case of Chinese linguistic history, change events are under the intertwined influence of languages spoken by Han and Non-Han people. Features of original local languages have an impact on the imperfectly learned Chinese, which varied the local speech of Chinese from the standard and formed Chinese dialects.

1.4 The Origin of Mandarin

In order to establish a solid foundation upon which we rationalize our research design, we give a thorough account of reflection over the essence of language and language change alongside the real linguistic history of Chinese in the above sections. We are probing into Mandarin varieties and its variation. Thus, before we get into the research design, we'll take one more detour to where everything about Mandarin begins ---- the origin of Mandarin.
Mandarin has long been speculated to originate in 幽燕 You-yan area, centered in modern Beijing, back in the Liao dynasty. Altaic speakers from the north ruled the three succeeding dynasties of Liao, Jin, and Yuan in the You-yan area. To seek the origin of Mandarin, non-Chinese transcription materials, which are almost undecipherable, of Khitan Lesser, Jurchen, and hP’ags-pa Scripts are ground-breaking contributors to confirm the real origin of Mandarin in the sense of actualizing its date and geographical locale. They help us to expand the horizon of understanding Chinese variation and reconsidering Chinese linguistic history in time and space.

Traditionally compiled by Deqing Zhou in 1324, the rhyme book Zhongyuan Yinyun (ZYYY) is taken as the earliest evidence of Mandarin's phonological system. As a standard textbook for composing poetry and rhymed passages to compete in the Imperial Civil Examination, the phonological system presented in ZYYY is undoubtedly a literary standard. The meaning of being a literary standard is two-fold. First, literary language offers the base upon which ZYYY categorized its phonological system. Second, test standard language is one of many Chinese varieties back in the day. This is to say the best ZYYY can provide us is a relatively stable linguistic entity which existed in the time of 1314. ZYYY cannot provide us information on the spoken language heard in the street back then. Neither does it cover languages of the time across the map of China. In fact, Chinese linguistic history has long been compressed into a linear development connecting the dots of phonological systems in rhyme books, which is problematic for bringing out the linguistic universal of Chinese in history that is compatible with the real development of Chinese variation.

Since Yintang Zhao (1936), scholars have worked on reconstructions of phonetic

15 Chinese rhyme books are historical records of Chinese phonologies in traditional Chinese phonological terms.
values of the phonological system in ZYYY. Such reconstruction effort is either based on the comparison of Middle Chinese and modern Mandarin, or based on the more phonetically transparent source materials, such as non-Chinese transcriptions. For example, Naisi Yang (1981) used hP’ags-pa spellings to reconstruct phonetic values of the phonological categories of ZYYY. With such valuable transcription scripts that reflect phonetic values instead of phonological categories, scholars definitely made advances in expanding the horizon of reconstruction. However, rhyme books give us phonological categories, not phonetic values, and reconstructions are intellectual guesses after all. The key issue overlooked and left out in the research field of Chinese linguistics is the history of Chinese variation in reality.

Specializing in non-Chinese transcriptions and dedicated to the scientific study of language change, Zhongwei Shen (2011) backdated the origin of Mandarin about three hundred years earlier than the time of ZYYY. More importantly, he has gradually worked toward the reconceptualization of Chinese variation with insights drawn from materials and methodologies of his studies. When he deciphered the Khitan Lesser Script of the Liao and discovered in them major phonological features of modern Northern Mandarin, he proceeded to prove his hypothesis that the phonology of Mandarin has already established by the eleventh century (Shen 2011: 25). Shen compared seven phonological properties for the five diachronic Mandarin varieties spoken roughly in the same area centered around modern Beijing. Phonologies represented in Khitan Lesser script of the Liao, Jurchen script of the Jin, and hP’ags-pa script of the Yuan were the three historical varieties of Mandarin, which preceded the Chinese variety represented in ZYYY. Besides these four varieties, the Beijing sub-dialect of Mandarin also participated in the comparison. The seven phonological properties were historically innovative. The
features compared were as followed:

- Diphthongization of MC syllables with coda \([-k]\)
- Devoicing of MC voiced obstruents
- Labiodentalization of MC bilabial initials
- Merger of MC zhi and zhao initials
- Loss of MC stop codas \([-p\), \([-t\), \([-k]\]
- Loss of MC velar nasal initial
- Loss of MC bilabial nasal coda \([-m]\)

The comparison gave us a matrix with reflexes of “+/yes” and “-/no” for the intersection of the seven phonological properties and the five varieties. Among these seven phonological properties, modern Mandarin showed all seven reflexes of yes. The phonological system of ZYYY occupied the first six properties. hP’ags-pa script took the first six with minor variation on “Devoicing of MC voiced obstruent” and “Loss of MC bilabial nasal coda \([-m]\)”. For these two “-” reflexes, Shen found out that the first “no” could be erroneous transcription of the script while the second was actually indicative of the feature being exact in the process of varying. The first five properties took “yes” reflexes for both Jurchen and Khitan Lesser scripts. The one property distinguishing Mandarin from all other Chinese modern dialects, astonishingly and informatively, presented in all three ancient non-Chinese scripts.

Unlike other analyses compressing Chinese variation in a linear fashion on a historical timeline, Shen was fully aware that the materials he examined had not only chronological but also areal significance. Such awareness allowed him to take his research result and further his findings to another level of comprehension. First, the most
distinguishable feature of Modern Mandarin presented in the three non-Chinese scripts of transcriptions. This is to say Mandarin formed its origin about a thousand years ago with the confirmation of three succeeding historical varieties in its modern geographical area. Second, among the seven phonological properties examined for the five varieties, the reflex of “yes” had an increase in correlation with time. Especially, the fact a certain feature was found at the stage of variation highlights that Mandarin should not be seen as a linguistic entity of static status. Its change was political and time-capsulated. To make this point clear, the three non-Chinese scripts were writing systems of three ethnic rulers, the Khitans, Jurchens, and Mongols. These three ethnic rulers were all Altaic language speakers, but with their own linguistic varieties. As they make contacts with the Chinese speakers in the north, Mandarin had correlated variation in terms of phonological features. In other words, alongside the confirmation of Mandarin's origin, the variation of Mandarin demonstrated an ever-evolving essence of linguistic change, and the motivation behind its change is politically involved with dynasty alteration. With ethnic language speakers in the picture of Mandarin variation, language contact and language transferring took the spotlight. Shen even proposed in his paper that with the same logic applied to Mandarin phonological features found in southern Chinese dialects, the possibility of Mandarin transmitted southward across China as a politically superior linguistic force to participate in forming Chinese dialects cannot be ruled out. Propelled by Shen's milestone finding of Mandarin's origin, we came to realize that the linguistic universal carried by Chinese speakers, including ethnic Chinese, is complicated and dynamic in contrast to the literary, standard, and static phonological systems represented by rhyme books from different times in history. Therefore, we asked questions against the complexity and dynamics of Chinese variation. We want to set up a study that fully
captures Chinese variation chronologically and geographically in respect to our above understanding of language, language change, and language history.
CHAPTER 2
RESEARCH DESIGN

2.1 Source Materials and Research Method

We have three cornerstone source materials for this study: *Hanyu Fangyan Gaiyao 2nd Edition*, *Hanyu Fangyan Diaocha Zhibiao*, and *Hanyu Fangyin Zihui 2nd Edition*. We use Jiahua Yuan’s *Hanyu Fangyan Gaiyao 2nd Edition* and other scholastic works for selecting diagnostic features for comparison. *Hanyu Fangyan Diaocha Zhibiao*, edited by Chinese Academy of Social Science in 1981, provides us with Chinese characters falling into the phonological description, in traditional Chinese terms, of each diagnostic feature. To investigate their phonetic values, we use *Hanyu Fangyin Zihui 2nd Edition*. It offers both literary and colloquial phonetic values for a good amount of characters among the three thousand characters. We do not use the literary phonetic values for this study, but only the colloquial ones. This is because we try to investigate Chinese as a language in change from the perspective of idiolects ever-evolving as they adapt to the speech environment in a complex fashion; colloquial Chinese is the most vital participant in Chinese variation on a constantly evolving base. As a matter of fact, the colloquial phonetic value is of dialects themselves, and the literary value reflects the influence of standard Chinese at various historical stage. Of course, we cannot exclude the influence of literary Chinese on the variation of Chinese in history (Xu 2018). However, it is only appropriate to use the colloquial phonetic values in this study for the sake of our research design.

The colloquial phonetic values of characters then confirm the reflexes of their correspondent diagnostic feature for each dialectal variety. The confirmation of some reflexes have to rely on other sources such as an online database to complement
characters' phonetic values for certain representative varieties.\textsuperscript{16} Twenty diagnostic features will be compared for twenty-eight Chinese varieties. The dialectal features selected for this study are phonological, descriptive of initial, final, and tone in the case of Chinese. The chosen phonological features are not just any phonological feature, but diagnostic ones. A diagnostic phonological feature should be a historical innovation, or specifically a sound change, that maximally differentiates the Chinese varieties under investigation. It is worthy to note that the criteria for selecting innovative features for this study are as followed: a). they are historical innovations thus are not just descriptive of a symbolic linguistic entity for any given time period; b). at least one of our language varieties\textsuperscript{17} can be distinguished by this feature from the rest; c). the innovation can be an incomplete sound change that does not apply to all characters in the phonological category correspondent to the feature description. A historical innovation does have a time of occurrence. The reason why we do not take its time of occurrence into consideration is complicated. The two most important factors are as followed. First, to find out the mechanism of variation, the occurrence of a sound change, regardless of its time frame, is indicative of historical linguistic development in the framework of variation being constantly on-going. Second, an in-depth exploration of the development of an individual feature will give us interesting facts peeping into language contact on a micro-scale; however, we do not study language contact of various circumstances, but of its directional influence on Chinese variation on a macro-scale. Throughout the

\textsuperscript{16} Wuyun Pan leads the effort in digitalizing Chinese linguistic evidence, which provides open access with vast resources on \url{www.eastling.org}.

\textsuperscript{17} To avoid stereotyped understanding of dialects as subordinates diverged from a language and the possible confusion between Chinese dialect groups and our representative Chinese varieties, we decide to use the neutral term of “language variety” for our twenty five synchronic and diachronic Chinese varieties to level them for a fresh look on their relations. After all, we are focusing on regional Chinese varieties instead of classified Chinese dialects.
discussion of this chapter, we will go into details on the rationale behind these criteria of feature selection.

As introduced in the overview section, twenty-five synchronic and three diachronic varieties make a total of twenty-eight representative Chinese dialectal varieties for this study. Shen's findings from phonological comparative work on Chinese transcriptions in Khitan, Jurchen, and hP’agas-pa Scripts provide us with references to confirm the reflexes of innovative Chinese phonological features in the three succeeding dynasties of Liao, Jin, and Yuan (Shen 2007, 2011, 2012). We label these three ancient Mandarin varieties, sharing almost the same locale with modern Mandarin sub-dialect of Beijing, as “Khitan Lesser Script”, “Jurchen Script”, and “hP’agas-pa Script”. We label the other twenty-five representative Chinese varieties with their city names. We list the labels of our twenty-eight representative varieties in the following section. The first one stays true to our spirit of studying these language varieties at the same linguistic level and with an awareness of their geographical relationship. The variety labels are listed in an order generally from the north to the south across China. The other two lists separate them into groups of Mandarin and non-Mandarin varieties to help sort out representative varieties in our later introduction of diagnostic features mainly by dialect groups.

**Compared Varieties From North to South (28):**

Shenyang (沈阳), Beijing (北京)/Khitan Lesser Script (契丹小字)/Jurchen Script (女真文)/hP’agas-pa Script (八思巴字), Lanzhou (兰州), Taiyuan (太原), Jinan (济南), Zhengzhou (郑州), Xi’an (西安), Nanjing (南京), Yangzhou (扬州), Hefei (合肥), Suzhou (苏州), Chengdu (成都), Wuhan (武汉), Nanchang (南昌), Wenzhou (温州), Changsha (长沙), Shuangfeng (双峰), Fuzhou (福州), Jian'ou (建瓯), Xiamen (厦门), Kunming (昆明), Meixian
Since the three ancient Mandarin varieties, Khitan Lesser Script, Jurchen Script, and hP’agas-pa Script, almost share the same geographical region centered around modern Beijing, they will be considered to have identical geographical locale with the representative variety of Beijing.

**Mandarin Dialectal Varieties (16):**

Khitan Lesser Script (契丹小字), Jurchen Script (女真文), hP’agas-pa Script (八思巴字), Shenyang (沈阳), Beijing (北京), Lanzhou (兰州), Taiyuan (太原), Jinan (济南), Zhengzhou (郑州), Xi'an (西安), Nanjing (南京), Yangzhou (扬州), Hefei (合肥), Chengdu (成都), Wuhan (武汉), Kunming (昆明)

**Other Chinese Dialectal Varieties (12):**

Suzhou (苏州), Nanchang (南昌), Wenzhou (温州), Changsha (长沙), Shuangfeng (双峰), Fuzhou (福州), Jian'ou (建瓯), Xiamen (厦门), Meixian (梅县), Chaozhou (潮州), Guangzhou (广州), Yangjiang (阳江)

The reflexes of twenty phonological innovations for twenty-eight Chinese varieties constitute the matrix of primitive data for computational calculation. The matrix (table) of reflexes compares diagnostic features for the language varieties with representations of “yes” and “no”. These yes/no reflexes will take a numerical transcription of “1” and “0” to computationally calculate the relative strength between each pair of language varieties under investigation. The numerical transcription can fluctuate between “0” and “1” to evaluate the weight of features. For our study, we do not proceed to designate different weights for our selected features. Since we try to study living language as a complex adaptive system, we find that variations, even minor ones, reflect “impure language”\(^\text{18}\). In this sense, all the selected features are equally weighted.

---

18 Shen made the observation that the object of historical linguistic studies is “impure language” (不纯洁的语言). (2015)
With the weight calculated, the actual numbers of numerical representation will be
different. However, the question is whether polished numbers matter to the content and
purpose of our study. Moreover, the weight of different features is evaluated against an
on-going debate on the classification of Chinese dialects. The classification of Chinese
dialects is not the focus of our study. Going into the controversies of classification will
steal the spotlight of the research from variation mechanism, and may even be complicit
in the stereotypical perception of dialectal distance based on classification. In fact, we
attempt to reveal the mechanism of Chinese variation on a macro-scale from a fresh
perspective, breaking the boundaries of classification. By large, dialectal classification
can obscure contact between differently classified varieties in a region. As long as the
general mapping is sufficient for a meaningful analysis, and the numbers generated are
indicative of the primary factor in Chinese variation, polishing up the numerical
representation is not obligatory in our study. Of course, brief explanations on how
different weights, other than “0” and “1”, play in the outcome of our study will make
their way into later discussions of the chapter.

To demonstrate the closeness of investigated dialectal varieties, we use the
computational method employed by linguists such as Alexandre Francois (2015) to
visualize with NeighborNet. The diagram as a whole gives us the visualization of the map
of Chinese dialects in terms of closeness. This net diagram is by nature multi-dimensional
but is displayed on a two-dimensional paper. However, such will not affect our analysis
of it; for the legend of the diagram will give us a solid reading of the distance between
every pair of dialectal varieties. However, the significance of reading the diagram is not
the actual distance between our dialectal varieties, but an interesting perspective of
understanding the dynamic contact among our dialectal varieties. This mathematical
embodiment will further facilitate an analysis of distance comparison with a numerical table generated. With the numbers in the table, we are able to make comparisons of distances between dialectal varieties. The analysis of comparison not only confirms our hypotheses, but also offers insights into the reconceptualization of Chinese variation.

The software we will use to compute the intersecting isoglosses acquired with the matrix is the program SplitsTree4 that was originally developed to process unrooted phylogenetic networks from molecular sequence data. With this application, we are able to use the method of neighbor-net to capture the distance between language varieties based on difference rather than similarity in a network. Demonstrating the relations among Chinese dialectal varieties from the aspect of difference is particularly significant to the design of our study in contrast to the traditional Comparative Method. The Method has largely influenced research minds to focus on the similarities in linguistics. However, with language contact being the core of this study's grand design, we need an approach acknowledging the difference of varieties to echo the real dynamics of Chinese variation. More details on this point will have their display in later discussion on the computational result reading. In short, difference among language varieties gives us the relative strength or distance without the presumption that these language varieties are sprung from any familial origin. Thus, the network of closeness is not the development of one protolanguage reconstructed from the similarities of investigated language varieties. It is a diagram showcasing the distance among language varieties by implementing a true inductive logic in diachronically analyzing linguistic data of “impure language”. (Shen 2015) In other words, the diagram leads to an analysis of the objective rather than the symbolic linguistic universal of Chinese in historical variation. To reveal

19 “Isogloss” is a linguistic term referring to the concept of regional dialectal features. Consequently, “intersecting isoglosses” means shared features among dialects.
the real variation of Chinese in the history, we calculate the cohesiveness between each pair of the twenty-eight varieties in the scope of the twenty diagnostic features. The comparison of these numbers of cohesiveness provides us different facets of Chinese transmitting across space and making contacts.

2.2 Twenty Phonological Features

History is dynamic, and so is linguistic history. The linguistic dynamics are so complex that the branching of linguistic history has to take the complexity of variation into consideration. The complexity of linguistic variation is a result of idiolects constantly adapting to each other and is perfectly embodied by on-going linguistic changes. Thus, to make our selected diagnostic features maximally distinguishable in revealing dialectal closeness, we go for innovative features regardless of their time of establishment. Rather, we cherish the impure phenomenon of language; phonological features in variation will be considered equally distinguishable in dialects with well-established innovative features that are conventionally recognized. Our study is designed to find the variation mechanism of Chinese situated under the perspective of a complex adaptive system. Therefore, innovative features in variation are, in fact, insightful and critical linguistic changes that have long been over-looked and underestimated in forming Chinese dialects. The above rationale discounts the necessity of distributing weights to the diagnostic features when computing. Besides, the evaluation scale of weighing the feature can be problematic: we should go with expert's subjective opinion on the importance of a feature or with considerations of the entire evolving history of an innovative feature. Either one is an open standard, which means the actual number we assign to the feature invites debates essentially belonging to issues of classification by dialectal feature. The classification issue involves a whole new set of questions waiting
for scholastic examination other than our research questions. We focus on the closeness
of dialectal varieties in revealing the variation mechanism of Chinese particularly in an
effort to avoid stereotyped perception caused by dialectal classification.

Before reading the computational results, we give a brief introductory account to
every diagnostic feature selected for the comparison.

1. Devoicing of MC voiced obstruent

Y.R. Chao first proposed this feature to distinguish Mandarin from other Chinese
dialects. The feature is innovative to all seven Mandarin sub-dialects. Thus, it
groups Chinese dialects into Mandarin and non-Mandarin. As previously stated
in the section on the origin of Mandarin, the devoicing of MC voiced obstruent is
a critical piece of linguistic evidence. This evidence contributes to the assertion
that Mandarin was formed before the Yuan dynasty because it is also found in the
phonology of Chinese transcriptions in non-Han scripts from the reign of the
Khitan and the Jurchen. This innovative feature is unique to Mandarin varieties,
and it is also a conventional major feature of Mandarin. This is the kind of
language change that is well-established in contrast to on-going variation.

However, if we extend the linguistic description of devoicing to voiceless
aspirated and voiceless unaspirated to give a complete phonation profile, we will
further divide Mandarin sub-dialects. The significance of bringing out this point
is two-fold. First, even the most well-established feature can experience on-going
change, which is a strong persuasion for us to study language from a dynamic
perspective. Second, feature weight comes in vain considering how extensive
linguistic description can vary the feature's degree of distinguishability in terms of
offering us different ways of grouping linguistic varieties. In other words, the
description of dialectal features is not standardized and gives us infinite ways to
group dialectal varieties, which produces overlapped group identities for a
dialectal variety. Therefore, different ways of grouping dialects are just different
ways to categorize them, which goes exactly against what inspires this project.
We want to level stratified Chinese dialects so that we can study Chinese
variation from a holistic viewpoint that is objective without any preoccupation.
Giving diagnostic features weights simply is not essential for examining living
language that is a trans-lingual phenomenon of impure interlanguage.(Gillian
2001)

2. Diphthongization of MC syllables with coda [-k]

Those MC Chinese characters with a [-k] ending sound had a VE composed of a
vowel and a consonant. Now both syllable structural slots are occupied with
vowels. This feature distinguishes Northern Mandarin, a sub-class of Mandarin,
from the rest of Mandarin varieties. A very important principal of cladistics is to
classify by difference. Thus, it is used to identify Northern Mandarin. What is
profoundly interesting in revealing the current open standard of classification in
Chinese dialectology is that the Northern Mandarin variety Zhengzhou does not
have this feature. In other words, this phonological variation is taken as a
Northern Mandarin representative feature not for exclusive reason but for
majority principle. On the other hand, other classification features, such as our
first feature, is exclusive for Mandarin. Sample characters for this feature are: 百
\textit{hundred}, 藥 \textit{medicine}, 腳 \textit{foot}, etc.

3. Labiodentals derived from MC bilabial initials

When the phonological condition of “[front-] palatal medial and rounded vowel”
is met by MC characters with bilabial initials b-, p', p-, these characters' bilabial initials change to labiodental fricative f-. This feature presents in Northern Mandarin, central Chinese dialects, and Hakka dialect in the south. Meixian is a representative Hakka dialectal variety. Sample characters for this feature are: 分 divide, 焚 burn, 芬 fragrance, and etc.

4. Loss of MC stop codas [-p], [-t], [-k]

Except for southern Chinese dialects, all other Chinese dialects had completed this sound change. Since we try to keep the consistency of selected features being innovative or indicating change, we describe this feature as a “loss.” However, it is worthy to note that the preservation of MC stop codas is distinguishing southern Chinese dialects in relation to non-Han language spoken by local Zhuang ethnics. “The more distinguishing the feature shared by two languages, the less probable the feature is developed separately in the two languages” (Shen 2016). When other Chinese dialects all experienced the loss of this feature, the preservation of this feature in southern dialects becomes extremely distinguishing. Such a highly distinguishing feature presents in local non-Han Chinese alludes to language contact of Han Chinese and non-Han Chinese in the past.

5. Loss of MC bilabial nasal coda [-m]

Dialects in northern and central China have this linguistic phenomenon with different expressions. The Xiang dialect and the Wu dialect completely lose the phonological contrast of the [-m] ending, which means these two central Chinese dialects don't have any characters pronounced in an -m ending sound. Differently, the loss of MC bilabial nasal coda expresses in Mandarin through alveolar nasal -n replacing bilabial nasal [-m]. On the other hand, southern
Chinese dialects preserve the bilabial nasal coda.

6. Loss of MC velar nasal initial

Again by majority principle, this feature distinguish Southwestern Mandarin from other Mandarin sub-dialects. Kunming, a Southwestern Mandarin dialectal variety, does not have this sound change. Among the representative dialectal varieties that have this innovative feature, not all of them complete the sound change. Taking Beijing as an example, when the phonological condition “[front-]palatal medial” exists, the MC velar nasal initial has a replacement of alveolar nasal initial instead of zero initial. Beijing pronunciation of 牛 cattle and 凝 condense substantiates this incomplete sound change.

7. Loss of MC ru tone

This feature can reflect the relation between Mandarin and non-Mandarin; and at the same time, it is conventionally used to distinguish the Jin dialect, represented by Taiyuan, from Mandarin. Rong Li first proposed using this feature to distinguish the Jin dialect. He uses this feature to distinguish the Jin dialect because this feature stands out in the Mandarin region. However, Jianghuai Mandarin also has this feature, but not geographically next to the area of the Jin dialect. In other words, the Jin dialect gets to be distinguished out from Mandarin not because it is different from Mandarin, but from the Mandarin dialects surrounding it. Obviously, this again reveals what an open standard Chinese dialectal classification can be. For those preserving MC ru tone, the Yue dialect has exactly the same corresponding MC phonological contrast of ending, but the Wu and Jin dialects both have glottal plosive in replacement of MC codas for MC
ru tone characters.

8. Merger of MC ni and lai initials

This is an iconic, essentially conventional, feature of Southwestern Mandarin. Jianghuai Mandarin representatives Nanjing and Yangzhou also have this feature. Southwestern Mandarin representative Kunming does not have this feature though.

9. Merger of MC alveolar affricates or fricatives and velar stops or fricatives in front of high-front vowels

This phonological innovation presents in most representative Mandarin varieties in our study. Northern Mandarin variety Zhengzhou and Jianghuai Mandarin variety Nanjing don't have it. Its development in Northern Mandarin representative Beijing is a contributor to substantiate Horizontal Transmission, which will be discussed toward the end of this paper.

10. Palatalization of initials for MC Division II syllables under xie Rhyme with jian and xia initials

This feature presents in Northern Mandarin. Mandarin representatives Wuhan, Chengdu, and Kunming in the south do not have it. For example, the initial of 街 street is “k-”, and the initial of 鞋 shoes is “x-” in Southwestern Mandarin Chengdu, which is still of MC phonology.

11. Nasal reflex of MC ri initial

This feature is not innovative in the sense of being new; in fact, it is a MC feature. Our criterion of innovation in feature selecting is its sense of change. In later discussion, we will see how Language Shift mechanism of Horizontal Transmission highlights preservation as a change.
12. Centralization of high-front vowel [i] for MC Division III syllables of zhi Rhyme with ri initial

Mandarin variety Wuhan does not have this feature. However, this feature is shared by other Mandarin varieties and Gan dialect representative Nanchang in central China. Other central Chinese dialects and southern Chinese dialects in southern dialects do not have this feature, but all have the tendency of this centralization.

13. Contrast of phonetic values of fu (夫) and hu (虎)

Except for Chengdu, Mandarin representative varieties all share this feature with Min dialect and Hakka dialect. The Gan dialect, Xiang dialect, and Yue dialect don't make distinction in these two pronunciations.

14. Labial Medial for MC ge Rhyme syllables with duan group initials

This feature differentiates Jin dialect and Northern Mandarin. Southwestern Mandarin and Jianghuai Mandarin do not share this feature with other Mandarin sub-dialects. Sample characters for this feature are: 多 many, 拖 drag, 搬 move, 搓 rub, 左 left, and etc.

15. Palatal Medial for MC Division II unrounded syllables with xi initial

Back in the time of confirmed Mandarin origin, it had already completed this innovation. Thus, it is a well-established feature of Mandarin. Although we appreciate incomplete innovation as being in the stage of variation, we don't take a single account of sound change as in the stage of variation. For instance, there is one character, among all characters associated with the description of this feature, 夾 mix pronounced with a palatal medial in southern Chinese variety Meixian. Other ru-tone characters falling into the feature description don't have
the change in Meixian pronunciation. Sample characters for this feature are: 家 (home), 觉 (sleep), 鸦 (crow), 甲 (armor), and etc.

16. Labial Medial for dang and jiang Rhymes with MC zhuang initial
This is a feature of Mandarin varieties. On the contrary, it does not present in southern Chinese dialects. Sample characters for this feature are: 双 (twin), 床 (bed), 壮 (vigor), and etc.

17. Puckered Mouth for MC Division III rounded syllables of yu Rhyme with initial groups of jing and ni
This feature shows up in all Mandarin representative varieties except Kunming. It also presents in Gan dialect representative Nanchang and Xiang dialect representative Shuangfeng. This feature can well differentiate Northern Min and Southern Min. Sample characters for this feature are: 徐 (slow), 取 (take), 聚 (gather), and etc.

18. Merger of zhi and zhao initials
All Min dialect representative varieties do not have this innovation, while all Mandarin varieties show positive on this feature. This initial merger has already taken place in original Mandarin back in the Liao dynasty. Sample characters of merged initial pronunciation are: “张” “章”, “桌” “捉”, “哲” “折”, and etc.

19. Vowel raising for MC Division III syllables of ma Rhyme
This is a feature differentiating northern and southern Chinese dialects. In northern Chinese dialects, MC Division III syllables of ma Rhyme have a vowel raising from low-back [a] to mid-front [e], -jā > -je. In southern Chinese dialects, the vowel of these characters keep at mid-low on the vowel chart. Sample
characters for this feature are: 也 also, 葅 people, 邪 evil, 且 and, 夜 night, and etc.

20. Merger of phonetic values of san (ㄋ) and shan (ㄕ)

This feature is known as the linguistic phenomenon carried by individual speakers who are incapable of making a distinction of characters with a retroflexed initial. It is conventionally coined to be a Southwestern Mandarin feature. As a matter of fact, dialects in central and southern China do not make such a distinction.
3.1 The Network of Mandarin Closeness

The computation of the matrix of intersecting isoglosses produces a diagram NeighborNet of our representative language varieties. As introduced in the section of method design, the net diagram is essentially multi-dimensional. The full diagram can be found in the Appendix. To read the diagram, every set of parallel lines in the web divide the twenty-eight dialectal varieties into two groups. There are twenty sets of parallel lines drawn by our twenty diagnostic features. The more evenly the diagnostic feature can divide up the twenty-eight varieties, the longer the set of parallel lines will be. A representative language variety can fall into different groups because of different intersected isoglosses it has with other varieties. For instance, the net reveals the general classification of Chinese dialect groups by a set of parallel lines, the longest among them all, in the middle of the web to distinguish Southern dialect groups on the left and Northern dialect groups on the right. The set of parallel lines in the middle is computationally drawn by the feature “Devoicing of voiced obstruents”, the critical linguistic variation in confirming the origin of Mandarin. This innovative feature divides up our twenty-eight representative varieties into two groups with numbers of the smallest difference, and this is how the computation visualizes the general Chinese dialectal map separating Northern and Southern dialect groups with the set of longest parallel lines in the middle of the web. However, this does not mean that Northern and Southern dialect groups don't share certain degree of closeness on other features. Actually, representative

---

20 “Southern” and “Northern” here are not in a strictly geographical sense, but in the sense of linguistic classification. For example, Mandarin variety Chengdu and Kunming belong to the Northern dialect, but geographically they are in the south.
language varieties across China are examined in this study for their mutual influence regardless of their classification of dialect groups. NeighborNet is particularly useful to alternate conventional focus on classification of dialect groups and to reveal the closeness of language varieties despite of the preoccupation of their categorical definition. Its multi-dimensional capacity of demonstration is going to capture relatedness of all investigated varieties along every twenty diagnostic dimension. In concrete terms, other sets of parallel lines besides the longest one in the middle of the web are going to divide our twenty-eight varieties into groups that have varieties from Northern and Southern dialect groups.

Such demonstrative capacity gives rise to interesting readings of the diagram. The readings align with language essence being dynamically evolving idiolects and with linguistic history being a reality of language transmitting. Of course, the closeness of varieties in this study is in terms of our twenty diagnostic features. With feature number expanded or feature description updated, we receive a different web weaving varieties in different ways. Future research utilizing the demonstrative tool of NeighborNet can display webs with varied details; but the general mapping of Northern and Southern dialects will stay the same as along as major classification features are compared. In our study, the diagram mainly serves as a visual aid in displaying the relationship among Chinese dialectal varieties while comparison of cohesiveness numbers in the next section will speak for details of language contact in Chinese variation.

Since the method of neighbor-net is originally designed for computationally sorting out unrooted genetic data to create a web of relative strength, it is able to demonstrate the closeness of the twenty-eight Chinese language varieties for our twenty diagnostic features. If taking Khitan Lesser as the reference point of reading as in Figure
2.1, the variation of Mandarin from its oldest variety Khitan Lesser is basically in correlation with geographical distance. The further the locale of a modern representative Mandarin variety is away from the locale, centered around modern Beijing, of the three ancient Mandarin varieties, the further the variety is away from Khitan Lesser in the diagram. Together, Khitan Lesser, Jurchen Script, hP'agas-pa Script, and modern Mandarin variety Beijing are stretching to the furtherest right corner of the web although they are about a thousand years apart in time.

Figure 2.1: The Closeness Web of Mandarin Varieties
As mentioned earlier, the two hypotheses of this study are: a). If we compare the closeness between diachronic Mandarin varieties and synchronic Mandarin varieties, we observe that the former is closer than the latter; then, we say that space is a more critical factor than time when it comes to Mandarin variation; b). If we can generate numerical evidence demonstrating that Mandarin, in the same region over a considerable amount of time, varied less than it did across China, then confidently we can conclude that the variation of Mandarin is due to language contact. The cropped part of NeighborNet in Figure 2.1 clearly visualizes our first hypothesis with Khitan Lesser, Jurchen Script, hP'agas-pa Script, and Beijing sub-dialect stretching to the right corner of the web. Mandarin varieties stay close to each other over a thousand years when it does not travel far, but modern Mandarin varieties spreading across China into all directions display an expansion of the web of closeness. Thus, space does play a critical role in Mandarin variation. In the next section, with comparison of cohesiveness numbers, not only we can confirm our second hypothesis, we can also reveal possible dialectal contacts in the development of Chinese variation alongside Mandarin variation.

3.2 Cohesiveness of Chinese Variants

The NeighborNet of Chinese representative varieties visualizes a relational web highlighting closeness from the perspective of diagnostic features differentiating them. With the number generated behind the NeighborNet, we are able to do an analysis of cohesiveness between each pair of varieties for the sake of our two hypotheses. Along the full display of NeighborNet, the complete list of cohesiveness numbers for the twenty-eight Chinese varieties and the matrix of intersecting isoglosses are included in the Appendix. In this section, we will first introduce the numerical formula that gives us the numbers of cohesiveness, and then we will take a close look at these numbers for a
linguistic analysis.

The formula of calculation is as followed:

\[
\text{Cohesiveness} = \frac{\text{Number of intersected features}}{\text{Total number of occupied features}}
\]

Taking the pair of varieties, Khitan Lesser Script and Jurchen Script, as an example, fourteen out of the twenty diagnostic features are occupied by both of the varieties. In the case of this pair of varieties, fourteen out of the fourteen occupied features are intersected features. Thus, the fraction of fourteen over fourteen gives us the cohesiveness number “1.000” for the oldest Mandarin variety by record and its succeeding variety roughly in the same geographical area in history. The cohesiveness of “1.000” of Khitan Lesser and Jurchen Script indicates that Mandarin spoken in the region of Liao rule, later taken by the Jin, stayed relatively staple without much variation in terms of our investigated phonological features. The closer the cohesiveness number is to “1”, the closer the cohesiveness of compared varieties is. With the examination of our selected twenty diagnostic features, the cohesiveness of Khitan Lesser Script and Jurchen Script reveals that the historical varieties of Mandarin in Liao and Jin are linguistically close to each other.

From the table below with cohesiveness numbers of Mandarin varieties compared to their original variety Khitan Lesser Script, we can see a general pattern of decreased closeness in correlation to the geographical distance away from the region of original Mandarin variety. Of course, such correlation is not strictly north-south directional; we
have to keep in mind that language contact in Mandarin variation takes place when Mandarin spread into all directions of the territory, and that varied Mandarin can keep influencing each other depending on their geographical closeness and population migration. However, the general north-south directional tendency of decreased closeness or increased variation undoubtedly presents in our numerical evidence. In brief, the number proves geographical transmission intensifies variation. This claim can further find its justification by dividing the numbers into two groups. One group is composed of the cohesiveness of three ancient Mandarin varieties; the other group is composed of the modern varieties. Although some of the modern varieties have high cohesiveness with original Mandarin, it does not affect the observation that modern varieties as a group in general tend to have less of a close relationship to original Mandarin comparing historical varieties sharing the same geographical area. The number of cohesiveness again confirms our first hypothesis: Mandarin variation does show that variation on geography is more drastic than variation over a time span.

Table 2.1: Cohesiveness of Original Mandarin and Its Varieties

<table>
<thead>
<tr>
<th>Variety</th>
<th>Khitan Lesser Script</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jurchen Script</td>
<td>1.000</td>
</tr>
<tr>
<td>hP'ags-pa Script</td>
<td>0.875</td>
</tr>
<tr>
<td>Shenyang</td>
<td>0.824</td>
</tr>
<tr>
<td>Beijing</td>
<td>0.824</td>
</tr>
<tr>
<td>Lanzhou</td>
<td>0.722</td>
</tr>
<tr>
<td>Taiyuan</td>
<td>0.611</td>
</tr>
<tr>
<td>Jinan</td>
<td>0.875</td>
</tr>
<tr>
<td>Zhengzhou</td>
<td>0.706</td>
</tr>
<tr>
<td>Xi'an</td>
<td>0.875</td>
</tr>
<tr>
<td>Nanjing</td>
<td>0.560</td>
</tr>
<tr>
<td>City</td>
<td>Cohesiveness</td>
</tr>
<tr>
<td>----------</td>
<td>--------------</td>
</tr>
<tr>
<td>Yangzhou</td>
<td>0.421</td>
</tr>
<tr>
<td>Hefei</td>
<td>0.588</td>
</tr>
<tr>
<td>Chengdu</td>
<td>0.556</td>
</tr>
<tr>
<td>Wuhan</td>
<td>0.526</td>
</tr>
<tr>
<td>Kunming</td>
<td>0.650</td>
</tr>
</tbody>
</table>

Also, grouping analysis of Mandarin cohesiveness numbers can confirm our second hypothesis directly and indirectly. The second hypothesis goes as: If we can generate numerical evidence demonstrating that Mandarin, in the same region over a considerable amount of time, varied less than it did across China, then confidently we can conclude that the variation of Mandarin is due to language contact. In Table 2.1, all the modern Mandarin varieties having high cohesiveness with original Mandarin representative Khitan Lesser Script are either sharing roughly the same geographical region with it or really close to its geographical area. They are mainly of Northern Mandarin sub-dialect groups which are developed in the region of the three ancient Mandarin varieties. This is to say Mandarin did not vary much over a thousand years in the region. On the contrary, sub-dialects of Mandarin further down into the south, such as Wuhan and Chengdu, only have about fifty percent cohesiveness with original Mandarin.

Other ways of grouping the data and conducting comparison analysis indirectly confirm the second hypothesis as well. In Table 2.2, we have Mandarin sub-dialect Chengdu compared with other twenty-seven representative varieties, including both Mandarin and Non-Mandarin varieties. The cohesiveness between Chengdu and other modern Mandarin varieties is closer than that between Chengdu and the ancient Mandarin varieties. The close cohesiveness among modern Mandarin varieties is reasonably
anticipated; mutual influence among Mandarin variants during the process of original Mandarin transmitting down to the south is inevitable. With rapidly advanced means of transmitting languages nowadays, the standard Mandarin prevails its influence on all Chinese variants in digital ways. Since Mandarin sub-dialects already belong to the group of Mandarin dialects and undoubtedly are linguistically adaptable to the standard Mandarin, the tie among Mandarin sub-dialects become even closer due to the prevailing influence from the standard Mandarin. What is more essential to our hypothesis from examining the cohesiveness numbers in Table 2.2 is the relationship between Mandarin variety Chengdu and non-Mandarin varieties. For example, Changsha is a non-Mandarin variety, but shares a cohesiveness of 0.600 with Chengdu. Comparing to the cohesiveness between Chengdu and other Mandarin varieties, a cohesiveness of 0.600 indicates that Changsha and Chengdu may not share linguistic features the way Mandarin varieties share but do share linguistic features more than anticipated. In other words, the contact between Mandarin varieties and non-Mandarin varieties is an undeniable possibility in Mandarin variation. What is even more telling is the general low cohesiveness of coastal non-Mandarin varieties with Chengdu in the south. Chengdu is the capital city of southwestern and in-land province Sichuan. Geographically, Hunan and then Jiangxi provinces separate Sichuan from coastal provinces in the south. The representative language variety of Jiangxi province is Nanchang that has a cohesiveness of 0.467 with Chengdu. The cohesiveness of Nanchang to Chengdu is less than that of Hunan to Chengdu, but greater than that of all the coastal representative dialects, such as Fuzhou and Guangdong. These numbers first demonstrate that Mandarin varieties can share certain closeness to non-Mandarin varieties. Thus, language contact between Mandarin and non-Mandarin exists. Also, the closeness between Mandarin variety
Chengdu and non-Mandarin varieties again is in correlation to the geographical distance.

Table 2.2: Mandarin Sub-dialect Chengdu and Other Chinese Dialectal Varieties

<table>
<thead>
<tr>
<th>City</th>
<th>Chengdu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Khitan Lesser Script</td>
<td>0.556</td>
</tr>
<tr>
<td>Jurchen Script</td>
<td>0.556</td>
</tr>
<tr>
<td>hP'ags-pa Script</td>
<td>0.579</td>
</tr>
<tr>
<td>Shenyang</td>
<td>0.632</td>
</tr>
<tr>
<td>Beijing</td>
<td>0.632</td>
</tr>
<tr>
<td>Lanzhou</td>
<td>0.632</td>
</tr>
<tr>
<td>Taiyuan</td>
<td>0.706</td>
</tr>
<tr>
<td>Jinan</td>
<td>0.667</td>
</tr>
<tr>
<td>Zhengzhou</td>
<td>0.611</td>
</tr>
<tr>
<td>Xi'an</td>
<td>0.667</td>
</tr>
<tr>
<td>Nanjing</td>
<td>0.647</td>
</tr>
<tr>
<td>Yangzhou</td>
<td>0.688</td>
</tr>
<tr>
<td>Hefei</td>
<td>0.588</td>
</tr>
<tr>
<td>Suzhou</td>
<td>0.375</td>
</tr>
<tr>
<td>Wuhan</td>
<td>0.813</td>
</tr>
<tr>
<td>Nanchang</td>
<td>0.467</td>
</tr>
<tr>
<td>Wenzhou</td>
<td>0.467</td>
</tr>
<tr>
<td>Changsha</td>
<td>0.600</td>
</tr>
<tr>
<td>Shuangfeng</td>
<td>0.533</td>
</tr>
<tr>
<td>Fuzhou</td>
<td>0.250</td>
</tr>
<tr>
<td>Jianou</td>
<td>0.235</td>
</tr>
<tr>
<td>Xiamen</td>
<td>0.063</td>
</tr>
<tr>
<td>Kunming</td>
<td>0.647</td>
</tr>
<tr>
<td>Meixian</td>
<td>0.200</td>
</tr>
<tr>
<td>Chaozhou</td>
<td>0.125</td>
</tr>
<tr>
<td>Guangzhou</td>
<td>0.286</td>
</tr>
<tr>
<td>Yangjiang</td>
<td>0.214</td>
</tr>
</tbody>
</table>
As the hypotheses confirmed, Chinese variation is a matter of space rather than time. The drastic variation of Chinese across China is the result of idiolects transmitting when making contact. The cohesiveness among Chinese varieties is not determined by the internal diversification of Chinese but largely by its history of language contact. Of course, our numbers cannot make the micro-scale situations of language contact transparent, but the general tendency of variation on a macro-scale is captured and presented. In Table 2.3, we can see that all non-Mandarin dialects are low in number on relative strength to original Mandarin. Considering our analysis of Table 2.2, non-Mandarin is almost linguistically nowhere near original Mandarin, but can have a cohesiveness convincing of language contact with neighboring Mandarin varieties. This observation takes our data to the next level of an inferential reading. The reading again indirectly proves our second hypothesis with an even persuasive argument as followed. The formation of Mandarin variation has contributions from non-Mandarin varieties that, in terms of language classification, are linguistically external to Mandarin. Thus, language variation in China should embrace a mechanism that is not confined within a single linguistic system but roots in languages making contact. Under the framework of Horizontal Transmission, the influence of non-Han Chinese and Chinese dialects themselves on the formation of Chinese dialects finds its proper loci in the mechanism of Chinese variation.

Table 2.3: Original Mandarin Variety and Non-Mandarin Varieties

<table>
<thead>
<tr>
<th>Variety</th>
<th>Khitan Lesser Script</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suzhou</td>
<td>0.294</td>
</tr>
<tr>
<td>Nanchang</td>
<td>0.294</td>
</tr>
<tr>
<td>Wenzhou</td>
<td>0.294</td>
</tr>
<tr>
<td>Changsha</td>
<td>0.412</td>
</tr>
<tr>
<td>City</td>
<td>Value</td>
</tr>
<tr>
<td>------------</td>
<td>-------</td>
</tr>
<tr>
<td>Shuangfeng</td>
<td>0.353</td>
</tr>
<tr>
<td>Fuzhou</td>
<td>0.176</td>
</tr>
<tr>
<td>Jianou</td>
<td>0.235</td>
</tr>
<tr>
<td>Xiamen</td>
<td>0.063</td>
</tr>
<tr>
<td>Meixian</td>
<td>0.125</td>
</tr>
<tr>
<td>Chaozhou</td>
<td>0.125</td>
</tr>
<tr>
<td>Guangzhou</td>
<td>0.200</td>
</tr>
<tr>
<td>Yangjiang</td>
<td>0.214</td>
</tr>
</tbody>
</table>
4.1 Research Findings

Our study is set to investigate twenty-eight representative Chinese dialectal varieties for their cohesiveness comparison in the scope of twenty diagnostic phonological features. Under the framework acknowledging the place of language contact in language variation, we analyze diagram produced by and data collected from the computation of intersecting isoglosses. The reading results pierce the veil of Chinese dialectal classification and give the real closeness between dialectal varieties. By “real closeness”, we mean the closeness acquired from a full-spectrum comparison of Chinese dialects, which gives due respect to the reality of language evolving through transmission wherever and whenever. The closeness of Mandarin, non-Mandarin, and Chinese dialects in general is in correlation with geographical distance, and grouped comparisons of cohesiveness indicate that language contact must intervene with language variation.

Both the NeighborNet of Chinese dialects and cohesiveness data of them confirm the two hypotheses of this study: a). If we compare the closeness between diachronic Mandarin varieties and synchronic Mandarin varieties, we observe that the former is closer than the latter; then, we say that space is a more critical factor than time when it comes to Mandarin variation, If we can generate numerical evidence demonstrating that Mandarin, in the same region over a considerable amount of time, varied less than it did across China, then confidently we can conclude that the variation of Mandarin is due to language contact. First, the further Mandarin modern varieties down into the south away from the region of original Mandarin, the less cohesive they are with ancient Mandarin varieties. Ancient Mandarin varieties, roughly developed in the same region, are very
close to each other, which means Mandarin from its origin did not vary much through three dynasties. Moreover, modern Northern Mandarin, geographically close to old Mandarin and represented by Beijing and Shenyang, is quite close to old Mandarin. Over a thousand years, Mandarin stayed linguistically close to its origin around the same geographical region. Such observation speaks for the inadequacy of internal diversification as the mechanism of Chinese variation. With persuasive cohesiveness between Mandarin sub-dialects and non-Mandarin dialects, particularly the neighboring ones, we finally get to demonstrate the relationship of Chinese dialects from a real historical perspective responding to the complexity of historical events.

The data of cohesiveness reveals language contact among neighboring Mandarin sub-dialects and between Mandarin sub-dialects and non-Mandarin dialects. Second, Mandarin sub-dialects' cohesiveness to original Mandarin is in correlation with geographical distance, and at the same time, Mandarin sub-dialects' cohesiveness to non-Mandarin dialects also show such a pattern of variation. Being the cohesiveness of a single linguistic entity or the cohesiveness of dialects from different linguistic classification, the mechanism of variation emerges as horizontal transmission across space rather than vertical diversification over time.

4.2 Horizontal Transmission

4.2.1 The Primacy of Language Contact

The Tree Model inversely induces vertical diversification of language as the mechanism of language change, and linguists, in turn, take the Tree to describe language history. The Method's theoretical framework has major logical flaws of false assumption and a disrupted chain of reasoning. From the aspect of linguistic descriptive capacity, the
Tree falls short to capture the picture of linguistic history in full due to its exclusion of language contact. As a linguistic phenomenon, language contact is constant if we take language as an ever-evolving ensemble of dynamic idiolects carried by individuals. Only by situating a method against the variation of dynamic idiolects can research set off to explore language history from an objective point of view and with a true inductive logic. If we see language and language change through the perspective of dynamic idiolects, we see language history as a process of ensembles of idiolects adapting to complex cultural environments. Taking examples from our own linguistic experience, every time we move to a new place, make a new friend, or watch a new TV program we catch on some new habitual phrase or accent. These new linguistic influences are usually not complete or perfect for it depends on their compatibility with our native tongue. Moreover, psychological impulses in selecting new features, and population prevailing in reserving new features can both add layers of complexity in the eventual outcome of social linguistic directional change. (Wang 2017) To make the picture of language history on a micro-scale more complicated, any settled new features can compete out features newly risen with the power of being the standard. However, among all the forces altering language history, political ones or authoritative standards (Francois 2015) are the most powerful forces on a macro-scale to language variation.

The Tree does not represent the history of language, because it does not have a descriptive capacity matching up with the dimensions of language. The history of language develops along dimensions of Time, Space, and People. (Shen 2016) Time indicates that linguistic history is the developmental process of language. Space states the fact that language change takes place at a locale where contact occurs and languages meet. People pinpoint that the language speaker is the vehicle through which real
language evolves.
The two essential questions of “What is language and language history?” and “Where
does the history of language take place?” start to shine through the above contemplation.
Shen made the answers explicitly in his 2015 work that language is the combination of
sound and meaning and that language history is an ensemble of idiolects transmitting in
time and space through language carriers. William Croft also found the necessity to
consider language change by situating languages and every piece of linguistic evidence
under particular refrained time and space. (Croft 2000) Time and space are the two very
basic dimensions of the entire universe; as “universe” is translated as 宇宙 in Chinese,
the significance of time and space to the laws of universe speaks for itself. Linguists
have taken language, as a subject of scientific study, on various attempts of scientific
analysis. Psycholinguistics and Neurolinguistics associate language with behavior and
neurobiology. They are scientific indeed; they provide us valuable scientific facts about
natural language. Nevertheless, scientific facts are not enough for a revolutionary
scientific study of language. A scientific study of language should locate its subject of
study in the universe, tracing its root of evolution in the real world. Where does language
evolve in the real world? Language evolves with language carriers as mutually interacted
idiolects, not as writing systems or dictionaries. Written language does change, but it's the
result of change brought by individual speakers first into oral language. This is not to say
oral language is never altered by the written; as one type of standard language, written
language can be immensely powerful in wielding the direction of language change for a
political reason. However, despite the direction of change, alteration takes place in the
speech contact of individual language carriers. Consequently, idiolect or individual
speech is the smallest unit of language change at its root of evolution.
4.2.2 A Mesohistory of Chinese

As a linguistic phenomenon of language variation, the formation of mutually unintelligible Chinese dialects has recently become more and more a fascination for researchers in the current discourse of Chinese linguistics. Studies of Chinese dialects draw theoretical reference from historical linguistics by default; however, this is like playing baseball on a soccer field. The conventional theory and methodology of historical linguistics does not hold an adequate descriptive capacity in explaining linguistic history in general, and thus will raise discrepancy against the real linguistic history of Chinese in studies of Chinese dialects. Thus, the variation of Chinese needs an indigenous and realistic explanation taking Chinese history into account. In the case of Chinese variation, we find not just finger prints of language contact, but a causative role played by language contact in forming the closeness of Chinese diachronic and synchronic dialectal varieties. Without language contact assuming the causality of Chinese variation, the significance of geographical correlation of dialectal distance in our research findings cannot be addressed properly. As a matter of fact, the geographical correlation of dialectal distance is the key to unlock the real linguistic history of Chinese. Feature sharing among Chinese dialects is common knowledge in Chinese dialectology, and this linguistic phenomenon has a conventional explanation of external influences. A shared feature automatically becomes an add-on property to the dialect that is not classified by that feature. As we discussed when considering assigning weights to different features, the classification by feature is open-standard and the classification alone still has unsettled debates. Thus, the preoccupation of classification induces countless problems especially in explaining agent and agency of the linguistic alternation. If we take a step further to examine the geographical correlation of feature sharing rather
than just taking feature sharing at its face value, we start to fit feature sharing in and for
the whole picture of Chinese variation.

From our research findings on the geographical correlation of feature sharing, we
can see the efficacy of Shen's theory of Horizontal Transmission in honestly reflecting
and highly abstracting the reality of Chinese dialectal relation. Horizontal Transmission
provides a framework where the Chinese dialectal relation finds the place it deserves in
Chinese variation through a mechanism acknowledging language contact. The Language
Shift Mode of Horizontal Transmission is able to make theoretical room for the dialectal
relation that is indispensable when rationalizing the formation of Chinese dialects. This
innovative theory of the Chinese variation mechanism reveals the intricate human setting
of language evolution. It best addresses the cultural complexity against which language
is organically integrated in. As inferred by the geographical correlation of feature
sharing, neither the formation of Mandarin sub-dialects nor Chinese dialects is a self-
evolving process. The process of Chinese evolution includes intertwined influences of
Han Chinese, local non-Han Chinese, and varied Han Chinese, which may well be
repeated in history for various social reasons. Standardization triumphs among the social
forces in deciding the direction of language shift. Since standardization can be on
different levels (Shen 2016), the Language Shift Model is inclusive of Chinese variation
in all situations. In other words, either Sinicization on a macro-scale or social preference
on a micro-scale is explained with the model for causing Chinese to vary. By situating
the development of Chinese in a sociopolitical landscape, Horizontal Transmission
highlights the dialectal relation in Chinese variation, which is to highlight the causality of
Chinese variation. The cause of Chinese variation ultimately determines the outcome of
dialectal relation. All in all, the descriptive capacity of Horizontal Transmission is
compatible with the true nature of living language and the complex adaptation of
dynamic idiolects.

With the mechanism of Chinese variation revealed with our research findings,
Sound Change and its cultural impetus are synchronized in the linguistic history of
Chinese. As a result, Chinese linguists start to find culturally synchronized explanations
for linguistic phenomena of Sound Change in the case of Chinese. For example, Weimin
Wang found that the course of “Merger of MC alveolar affricates or fricatives and velar
stops or fricatives in front of high-front vowels” feature development in northern China
can prove that this variation in modern Beijing dialect originally traces to Manchu
speakers' imperfect learning of Chinese. (Wang 2018) Horizontal Transmission
embodies a mesohistory of Chinese in the spirit of reasserting language carriers in time
and space; in this way, it rebuilds the language reality of Chinese as a complex adaptive
system. (Shen 2016) The linguistic phenomenon of variation or language change in the
case of Chinese finds the most appropriate theoretical framework in Horizontal
Transmission for its panoramic historical view. The dialectal relation visualization and
comparison data analysis of this study both confirm the efficacy of Language Shift as the
mechanism of Chinese variation and the necessity of Horizontal Transmission as the
reconceptualization of Chinese linguistic history.

In summary, this study echoes the primacy of language contact in the variation of
language (Sarah Grey Thomason 2003). The visualization and further analysis confirm
the reconceptualization of Chinese linguistic history with the theory of Horizontal
Transmission (Shen 2016). Horizontal Transmission situates the development of
Mandarin and other Chinese dialects in a sociopolitical landscape as a cultural complex
and introduces imperfect learning to the time-encapsulated process of standardization as
an inevitable social phenomenon (Shen 2016). The descriptive capacity of Shen’s theory is compatible with the nature of language being dynamic idiolects alongside a real linguistic history embodied by individual speakers in time and space. Only by understanding the change mechanism of Chinese from the perspective of language contact and through the lens of language shift, the variation of Mandarin and emergence of Chinese dialects find their explanations in a salient chain of logic to create a holistic account of Chinese evolution where the intertwined influence of languages finds its manifestation.
Map 1. The Atlas of Chinese Dialects by Ten Dialectal Groups

Copyright by:

Map 2. The Atlas of Chinese Dialects with Mandarin Sub-dialectal Groups

Copyright by:
https://www.dartmouth.edu/~chinese/maps/map4.html
APPENDIX II

MATRIX OF INTERESTING ISOGLOSSES

*Twenty diagnostic features are in their numerical orders from 1 to 20.
APPENDIX III

NEIGHBOR-NET OF TWENTY-EIGHT CHINESE DIALECTAL VARIETIES
### APPENDIX IV

**NUMBER OF COHESIVENESS**

<table>
<thead>
<tr>
<th>City</th>
<th>Number of Cohesiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beijing (北京)</td>
<td>0.024 0.024 0.024</td>
</tr>
<tr>
<td>Lanzhou (兰州)</td>
<td>0.024 0.024 0.024</td>
</tr>
<tr>
<td>Taiyuan (太原)</td>
<td>0.024 0.024 0.024</td>
</tr>
<tr>
<td>Jinan (济南)</td>
<td>0.024 0.024 0.024</td>
</tr>
<tr>
<td>Zhengzhou (郑州)</td>
<td>0.024 0.024 0.024</td>
</tr>
<tr>
<td>Xi'an (西安)</td>
<td>0.024 0.024 0.024</td>
</tr>
<tr>
<td>Nanjing (南京)</td>
<td>0.024 0.024 0.024</td>
</tr>
<tr>
<td>Yangzhou (扬州)</td>
<td>0.024 0.024 0.024</td>
</tr>
<tr>
<td>Hefei (合肥)</td>
<td>0.024 0.024 0.024</td>
</tr>
<tr>
<td>Suzhou (苏州)</td>
<td>0.024 0.024 0.024</td>
</tr>
<tr>
<td>Khitan Lesser Script (契丹小字)</td>
<td>1.000 1.000 1.000</td>
</tr>
<tr>
<td>Jurchen Script (女真文)</td>
<td>1.000 1.000 1.000</td>
</tr>
<tr>
<td>hP'ags-pa Script (巴思八字)</td>
<td>1.000 1.000 1.000</td>
</tr>
<tr>
<td>Shenyang (沈阳)</td>
<td>1.000 1.000 1.000</td>
</tr>
</tbody>
</table>
Chengdu (成都),
Wuhan (武汉),
Shanghai (上海),
Nanjing (南京),
Jinan (济南),
Zhengzhou (郑州),
Xian (西安),
Chongqing (重庆),
Chaozhou (潮州),
Jianlou (景德镇),
Xi'an (西安),
Kunming (昆明),
Chaozhou (潮州),
Xiamen (厦门),
Changzhou (常州),
Nanchang (南昌),
Wenzhou (温州),
Hangzhou (杭州),
Shangfeng (上峰),
Shandong (山东),
Beijing (北京),
Lanzhou (兰州),
Taiyuan (太原),
Jinan (济南),
Zhengzhou (郑州),
Xi'an (西安),
Nanjing (南京),
Yangzhou (扬州),
Hefei (合肥),
Suzhou (苏州),
Khitan Lesser Script (契丹小字),
Jurchen Script (女真文),
hPags-pa Script (巴思八字),
Shenyang (沈阳),
Beijing (北京),
Lanzhou (兰州),
Taiyuan (太原),
Jinan (济南),
Zhengzhou (郑州),
Xi'an (西安),
Nanjing (南京),
Yangzhou (扬州),
Hefei (合肥),
Suzhou (苏州).
BIBLIOGRAPHY

English Sources:


Chinese Sources:
周振鹤和游汝杰（1984）《方言与中国文化》，上海人民出版社，上海。

李如龙（2001）《汉语方言的比较研究》，商务印书馆，北京。

游汝杰（1992）《汉语方言学导论》，上海教育出版社，上海。

李小凡和项梦冰编著（2009）《汉语方言学基础教程》，北京大学出版社，北京。

汪峰和林幼菁主编（2017）《语言与人类复杂系统》，云南大学出版社，昆明。

沈钟伟（2006）辽代北方汉语方言的语音特征，《中国语文》6：483-498页。