



Universiteit
Leiden
The Netherlands

Advancing explanatory and tonal dialectometry

Sung, H.W.M.

Citation

Sung, H. W. M. (2026, February 13). *Advancing explanatory and tonal dialectometry*. LOT dissertation series. LOT, Amsterdam. Retrieved from <https://hdl.handle.net/1887/4291801>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4291801>

Note: To cite this publication please use the final published version (if applicable).

Advancing Explanatory and Tonal
Dialectometry

Published by

LOT
Binnengasthuisstraat 9
1012 ZA Amsterdam
The Netherlands

phone: +31 20 525 2461

e-mail: lot@uva.nl
<https://www.lotschool.nl>

Cover illustration: A photo of Hong Kong, photographed in December 2019 by Matthew Sung.

ISBN: 978-94-6093-494-0

DOI: <https://dx.medra.org/10.48273/LOT0709>

NUR: 616

Copyright © 2026 Ho Wang Matthew Sung. All rights reserved.

Advancing Explanatory and Tonal Dialectometry

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr. S. de Rijcke,
volgens besluit van het college voor promoties
te verdedigen op vrijdag 13 februari 2026
klokke 11:30 uur

door

Ho Wang Matthew Sung

geboren in 1996
in Hong Kong

Promotor: Prof. Dr. Yiya Chen
Copromotor: Dr. Jelena Prokić
Promotiecommissie: Prof. Dr. Carole Tiberius
Prof. Dr. Sjef Barbiers
Prof. Dr. Jack Grieve (University of Birmingham)
Dr. Wilbert Heeringa (Fryske Akademy)

Contents

Acknowledgments	ix
List of Figures	xi
List of Tables	xv
1 Introduction	1
1.1 Definition of ‘Dialect’	1
1.2 Background	2
1.3 Motivations of the dissertation	4
1.4 Research questions	5
2 Traditional Classification of Yue-Pinghua Dialects	9
2.1 Introduction	9
2.2 Traditional classification of Yue and Pinghua dialects	10
2.2.1 Previous classifications of Yue and Pinghua	10
2.2.2 Limitations to the existing classifications	24
3 An Overview of Dialectometry	29
3.1 Dialectometry	29
3.1.1 What is dialectometry?	29
3.1.2 Problems with the isogloss approach	30
3.1.3 An aggregate perspective	31
3.2 A brief history of dialectometry	32
3.2.1 Pre-modern dialectometric studies	33
3.2.2 Modern dialectometric approaches	34

3.2.3	Recent developments	35
3.3	Methodology	37
3.3.1	Distance calculation	37
3.3.2	Multidimensional scaling	40
3.3.3	Cluster analysis	44
3.3.4	Quantitative maps	50
4	Data	51
4.1	Introduction	51
4.2	Data sources	52
4.2.1	Sources	53
4.2.2	Selection of words	54
4.3	Modifications to the original segmental transcriptions	57
4.3.1	Comparison with existing recordings	57
4.3.2	Maintaining contrasts	57
4.3.3	Removal of redundant characters	58
4.3.4	Simplification of overly detailed transcriptions	59
4.3.5	Consistency of onsets	59
4.3.6	Converting Chinese IPA to Standard IPA	59
4.3.7	Phonetic alignment	60
4.3.8	Descriptive statistics	60
4.4	Tonal data	61
4.5	Conclusion	62
5	Segmental Variation of Yue Dialects	63
5.1	Introduction	63
5.2	Dialectometric analysis of Yue segmental variation	64
5.2.1	The Yue-Pinghua dichotomy	65
5.2.2	Comparison with the traditional classification of Yue dialects	68
5.2.3	A dialectometric classification of Yue dialects	72
5.3	Patterns in Yue variation	86
5.3.1	Dialect continua in Yue?	87
5.3.2	Correlates with Yue variation	89
5.4	Conclusion	94
6	Automatic Feature Detection for Yue Dialects	95
6.1	Introduction	95
6.2	Previous approaches	96

6.2.1	Bottom-up approaches	97
6.2.2	Top-down approaches	99
6.2.3	Comparison of feature extraction methods	100
6.3	Methodology	101
6.3.1	Classification of Yue Dialects	102
6.3.2	Feature Extraction	106
6.4	Top features in different Yue sub-dialect groups	108
6.4.1	Guangfu dialects	111
6.4.2	Siyi dialects	112
6.4.3	Coastal dialects	114
6.4.4	Goulou dialects	115
6.4.5	Western dialects	118
6.5	Discussion and Conclusion	119
6.5.1	Features in the traditional classification	119
6.5.2	nPMI threshold	120
6.5.3	Historical interpretation	121
6.5.4	Conclusion	122

7 Comparison of the Existing Tone Distance Calculation

Methods		123
7.1	Introduction	123
7.2	Tonal Languages	125
7.3	Transcription notations and formal representations	126
7.3.1	Transcription Notations	126
7.3.2	Formal representations	130
7.3.3	Optimal representation of tones for measuring tone distances	134
7.4	Previous approaches to measuring tone distances	135
7.4.1	Onset-Contour-Offset	135
7.4.2	Tone-to-string	137
7.4.3	Binary comparison	138
7.4.4	Gandour-Harshman-Tang tone distance measurement	140
7.5	Comparison of the tone distance calculation methods	143
7.5.1	Tone overlaps	145
7.5.2	Comparison with the perceptual dimensions	148
7.5.3	Local incoherence	150
7.5.4	Adjusted Rand Index (ARI)	151
7.6	Summary of the results and discussion	152

7.7	Conclusion	155
8	Tonal Variation of Yue Dialects	157
8.1	Introduction	157
8.2	Dialect Tonometry	158
8.2.1	The Modified OCO (mOCO) representation	158
8.2.2	Calculating aggregate tone distances	160
8.3	Tonal variation under the scope of dialectometry	162
8.3.1	Tonal variation between dialects	162
8.3.2	A geographical continuum?	163
8.4	Tonal vs. segmental variation	168
8.5	Discussion	173
8.6	Conclusion	177
9	Conclusion	179
	References	185
	Appendices	205
	Samenvatting	211
	Summary	213
	粵語摘要	215
	Curriculum Vitae	217

Acknowledgments

I owe my gratitude to everyone who has provided me with direct and indirect support throughout my PhD, the preparation of this thesis, and my development as a researcher.

I would first like to thank my brilliant supervisor, Jelena, for her guidance and her non-stopping support, even before my arrival in Leiden (we were in lock down when I started my PhD). I am very grateful for her backing and visions in the field. She is an inspiration to me; she has influenced me a lot in my work and in finding my direction in academia. Throughout my PhD, she has encouraged me to acquire a variety of skills, including programming, multiple sequence alignment, natural language processing and later, even image processing. In retrospect, all of these skills proved to be essential. I would also like to thank my promotor Yiya for her inputs in the tonal part of the thesis. It was not an easy path to start exploring ways to measure tonal distances and her comments to my early ideas and to the modifications of OCO helped a lot in shaping the current thesis. Last but not least, I would like to thank them both for dedicating a lot of time and energy to reading my thesis and providing me with valuable feedback.

I would like to thank my reading committee, Wilbert Heeringa, Jack Grieve, Sjef Barbiers and Carole Tiberius, for reading and evaluating my thesis. Their comments were very helpful in shaping the final draft of my thesis.

A shout out to my proofreaders who have agreed to proofread my thesis without taking any compensations. These generous friends include: Cliodhna Huges, Nina Markl, Tim Espin and Lily Wood. A special mention to Nef and Raoul for proofreading my summaries in Cantonese and

Dutch respectively.

I would also like to mention a number of people who helped me a lot in getting acquainted with dialectometry. Pavel Iosad and Warren Maguire first introduced the concept of dialectometry to me when I was an undergraduate student in Edinburgh. I was very fortunate to have met Wilbert Heeringa at ICLaVE in Leeuwarden in 2019 and he selflessly spent two hours of his dinner time teaching me the basic methods in dialectometry from scratch. I was also fortunate to have the opportunity to meet John Nerbonne in Freiburg at the end of 2019, where he kindly offered to teach me more about using Gabmap for dialectometry. On a different note, I would also like to express my gratitude to Chris Montgomery for teaching me the principles of creating digital maps on QGIS. Without his initial guidance, I would not be able to develop my cartographic skills which I have obtained today. Remco Knooihuizen in Groningen deserves another mention. On the train back to the Netherlands from the Methods conference in Mainz, we had a discussion on how classic dialectometry lacked explainability. This conversation was stuck in my mind until I got the idea of using nPMI to perform feature extraction.

Leiden has become very special to me and one of the reasons is the people I met here. I want to thank my friends from CM, LUCL and LUCDH. Here are a few mentions: Unnur, Laura, Isabella, Priscilla, Aron, Amos, Jiahui, Qi, Marloes, Hans, other houtblazers, the altviool, Yiya's group and colleagues at the LUCDH. I would also like to mention a few friends outside Leiden who have been there for me, even though they are not physically in Leiden, namely Nef, Raoul, Hedwig, Heize, Veronique.

I also want to thank the friends who have helped me to buy books for me in the past few years. Without their help, I won't have access to these valuable resources otherwise.

Last but not least, I want to thank my parents. Since I was a child, they have been giving me pouring love and support. Without them, I would not be able to fulfill my dream of becoming a dialectologist. I am forever grateful for what they have done to support me.

List of Figures

2.1	Map of Yuan's (1960) classification	11
2.2	Map of Zhan's (2002) classification	12
2.3	Map of Li's (1994) classification	13
2.4	Map of Yang et al. (1985) and Xiong's (1987) classification	14
2.5	LAC Classification Criteria	16
2.6	LAC Classification of Yue dialects (Tree representation) .	17
2.7	Map of Yue-Hashimoto's (2006) classification	18
2.8	Yue-Hashimoto's (2006) Classification (Tree representation)	18
2.9	Map of Xie's (2007) classification	19
2.10	Map of Sung's (2020) classification of historical Yue dialects	21
2.11	Map of Carlyle's (2020) classification	22
2.12	Geographical distribution of Yue, Guinan and Guibei Pinghua dialects	24
2.13	LAC Classification Criteria with a Different Criteria Order	26
2.14	LAC Classification of Yue dialects with a different feature order (Tree representation)	27
3.1	Dutch isogloss map	31
3.2	Workflow of Dialectometry using Levenshtein Distance . .	38
3.3	Illustration of Levenshtein Distance	39
3.4	An example of an MDS plot	42
3.5	An example of an MDS scree plot	43
3.6	Three-dimensional RGB spectrum	44
3.7	Dendrogram of 9 Yue dialects	46

3.8	Comparison of different cluster solutions based on various cluster algorithms	47
3.9	Probabilistic dendrogram of 9 Yue dialects	50
4.1	Sources of Yue dialect transcriptions	53
4.2	Localities and their respective sources	55
4.3	Boxplot of Distances between Raw and Modified Transcriptions.	61
5.1	MDS plots of the segmental distances of 113 Yue and Pinghua dialects	66
5.2	MDS map of the segmental distances of 113 Yue and Pinghua dialects	67
5.3	MDS plot of 104 Yue dialects	68
5.4	LAC classification map	69
5.5	MDS plot of 104 Yue dialects with LAC dialect group annotation (Dimension 1 & 2)	70
5.6	MDS plot of 104 Yue dialects with LAC dialect group annotation (Dimension 2 & 3)	71
5.7	Average Silhouette Scores from different number of clusters obtained with four cluster algorithms	74
5.8	Silhouette Plot for UPGMA, 5-cluster solution	75
5.9	Reference point maps of Heshan, Xindu and Lingui	76
5.10	MDS plots of the 5-cluster solution from four cluster algorithms	77
5.11	Silhouette Plot for Ward's method, 5-cluster solution	78
5.12	Dendrogram of 104 Yue dialects (Segmental distances, Ward's method)	79
5.13	Cluster maps based on the Ward's method	80
5.14	Probabilistic dendrogram of 104 Yue dialects	82
5.15	Cluster map based on the Ward's method with dialect islands	83
5.16	Scree Plot for MDS (Segmental variation)	84
5.17	MDS map of Yue segmental distances	85
5.18	Continua within a continuum in the Yue-speaking area	88
5.19	MDS map of the northern border between Guangdong and Guangxi overlaid with elevation and major river systems	90
5.20	MDS map overlaid with elevation and major river systems	91

5.21	MDS map with Pre-1952 Provincial Borders	93
6.1	Combined Factor Map of the Sprachatlas von Bayerisch-Schwaben (SBS) survey sites	98
6.2	Illustration of distance calculation for the FLD method	100
6.3	Distribution map of the most important feature extracted by various methods	101
6.4	Illustration of Multiple Sequence Alignment	103
6.5	Cluster Map of Yue Dialects in Guangdong and Guangxi (MSA)	105
6.6	Calculation of nPMI scores with toy example	108
6.7	Comparison of associations (nPMI scores) between 3 Dutch dialect variants and the West Flemish area	109
6.8	Distribution maps of the top two Guangfu features	113
6.9	Multivariate distribution map of Goulou dialect features	117
6.10	Ranges of nPMI of top 20 features for each dialect group	120
7.1	Tone symbols for the Middle Chinese tone categories in the Yin register	128
7.2	Gedney's (1989) Tone Box	129
7.3	Typology of tone notations across traditions	130
7.4	Wang's (1967) tone features	131
7.5	Sampson's (1969) modification of Wang's (1967) tone height features	132
7.6	Woo's (1969) proposal of tone features	133
7.7	Illustration of the two most important perceptual dimensions of tones	144
7.8	MDS plot for Tone-to-string	147
7.9	MDS plot for OCO	147
7.10	MDS plot for GH-T	148
7.11	Overall results of evaluation across 4 tone distance calculation methods	152
7.12	Comparison of tones 53 and 12 with 31	154
8.1	2-dimensional MDS plot of tone distances using mOCO representation	160
8.2	Dialect tonometry procedures	161
8.3	MDS plot of aggregate tone distances between Yue dialects	163
8.4	MDS map of tonal variation of Yue dialects	165

8.5	Cluster map of tonal variation of Yue dialects	166
8.6	3-dimensional MDS plot of tone distances using mOCO representation	167
8.7	MDS map of tonal variation of Yue dialects with borders based on cluster analysis	168
8.8	Correlation matrix between each dimension from MDS for tonal and segmental distance matrices	170
8.9	Individual dimension maps for segments (Dimension 3) and tones (Dimension 1)	171
8.10	Individual dimension maps for segments (Dimension 1) and tones (Dimension 3)	172
8.11	Individual dimension maps for segments (Dimension 2) and tones (Dimension 2)	174

List of Tables

3.1	A sample distance matrix	40
4.1	Examples of Raw vs. Modified Transcriptions	60
5.1	Explained Variance of various cluster algorithms	73
5.2	Summary of Yue segmental classifications	85
6.1	Top 20 Features in Guangfu Yue Dialects	112
6.2	Top 20 Features in Siyi Yue Dialects	114
6.3	Top 20 Features in Coastal Yue Dialects	115
6.4	Top 20 Features in Goulou Yue Dialects	116
6.5	Top 20 Features in Western Yue Dialects	118
6.6	Features that are identified by nPMI and the traditional classification	119
7.1	Contours in OCO representation with examples	136
7.2	Calculation of Levenshtein Distance between 221 and 24 in OCO representation	137
7.3	Calculation of Levenshtein Distance between 325 and 15 with tone-to-string method	138
7.4	Examples of lexical distribution differences in tones of two Siyi dialects	140
7.5	Cues and their relative feature values in GH-T	141
7.6	Example of tone distance calculation using GH-T	142
7.7	Differentiation of tones under each tone distance measure	146

7.8	Interpretation of the first two dimensions in the MDS plots of each tone distance measure	149
7.9	Local Incoherence scores for each tone distance measure and segments	150
7.10	ARI Similarity matrix between tone distance measures, traditional classification and segmental classification . . .	151