Computational Linguistics in the Netherlands Journal 14 (2025) 135-165 Submitted 12/2024; Published 07/2025

On Dialect Typicality and Transitional Dialects

Ho Wang Matthew Sung[*] H.W.M.SUNG@HUM.LEIDENUNIV.NL

[*] *Leiden University, Leiden, The Netherlands*

Abstract

Since the late 19th century, dialectologists have already noticed that dialect areas do not have abrupt borders, but gradual transitions from one zone to another. One could speak of 'focal' or 'core' vs. 'transitional' areas for relatively homogenous areas vs. regions with linguistic variants from different neighbours. Dialectometry, the computational branch of dialectology, offers tools to capture and visualise dialect transitions. These techniques can show the global transition patterns, but not for which features, and in what way are they contributing to the transition. On the other hand, detailed studies on individual (often singleton) features in variationist sociolinguistics offer us insights to the details in the way dialects transition from one area to another. But the limit of close examination is the possible over-generalisation of the pattern of one feature over many other features in the dialects. In addition, it is not possible for humans to analyse every single variable in the data manually.

In this paper, a new approach is proposed to explore dialect transitions on the feature level, namely the *dialect typicality decay analysis*. This novel approach builds on previous approaches in dialectometry (automatic dialect classification and feature extraction), and it explores how transitional (and prototypical) dialects possess characteristic features of a particular dialect group. Two main issues are explored using dialect typicality analysis: 1) what does dialect transition look like based on their top characteristic features? and 2) do transitional dialects favour adopting the most typical features from a certain dialect group?

1. Introduction

Since the 19th century, dialects have been realised to form a geographical continuum (Meyer 1877, Paris 1888). A dialect continuum implies that there are no abrupt boundaries between different groups of dialects, but a gradual transition from one type of dialect to another (Chambers and Trudgill 1998, Heeringa and Nerbonne 2001). Dialectologists have identified several ways in which dialects can transition from one zone to another. One could speak of a mixed dialect when dialects use variants from their neighbours from both sides; a fudged dialect on the other hand is a dialect which uses a variant which is midway between the neighbouring variants (Chambers and Trudgill 1998, 110). One more possibility is to have both mixed variants and the midway variant of their neighbours together. In addition, mixed dialects can also show lexically, phonologically or socially conditioned usage differences of one variant over the other (Taeldeman 1989).

Many studies on the types of transitional dialects, however, are based on single variables. There are limited discussions of dialect transition on a multivariate level, with the possible exception of Simmons (2012), which examined various features in detail. Despite the insightful results we have obtained for the typology of transitional dialects, single-variable studies do not necessarily reflect the general tendencies of the way dialects transition from one dialect area to another, as these studies often carefully select one feature out of thousands of other possible features, and that poses biases to the analysis (Nerbonne 2010). A multivariate analysis, on the other hand, can offer us a somewhat more representative view of the issue; the more features considered, the more representative it can get. However, due to the limits of how many features one can process manually, like in Simmons (2012), often it is still difficult to avoid a subjective selection of features in the process in the manual multivariate analyses, which is understandable.

For the past few decades, computational methods, or more specifically, distance-based approaches in unsupervised learning have constributed to the major breakthroughs in dialectology. The quantitative branch of dialectology, known as dialectometry (Séguy 1971), makes use of techniques such as cluster analysis (e.g. Goebl 1984), multidimensional scaling (e.g. Embleton 1993) and calculating Levenshtein distance (e.g. Heeringa 2004) to process a vast amount of dialect transcriptions published in the dialect surveys or linguistic atlases, which were collected from the most conservative dialect speakers known as NORMs (non-mobile, rural, old males, Chambers and Trudgill 1998). One of the major goals which dialectometry delves into is automatic dialect classification. By taking account of more features in the analysis together (by aggregation, into dialect distances) with the help of cluster analysis and multidimensional scaling, dialectometrists are able to find dialects which are more similar to each other, yielding a more objective classification of dialects.

The distance-based approach, however, received criticisms in not being able to return the "details or explanations of the identified dialect partitions" (Sung and Prokić 2024a). This is because qualitative linguistic differences between dialects were reduced to numerical dialect distances, and methods such as cluster analysis cannot retrieve the qualitative features afterwards. To address this problem, several approaches have been proposed. For instance, Pickl (2016) made use of Factor Analysis (FA), a dimensionality reduction technique that seeks the relationship between the dialect and a dialect group (the *factor loadings*) and the degree of association of features to a dialect group (the *factor scores*) simultaneously. Another technique proposed by Sung and Prokić (2024a) utilises normalised pointwise mutual information, an association measure based on co-occurrence statistics. This technique requires a predefined grouping of dialects (e.g. from cluster analysis), and it extracts the features post-hoc by calculating the association between each variant and the dialect groups.

Despite the advances in quantitative dialectometry with feature extraction, the application of feature extraction beyond explaining dialect groups is still limited. At the same time, the examination of dialect transitions have mostly been focusing on single-variants. Even in cases where multiple variables are considered, like in Simmons (2012), not all the features in the data are being utilised, due to the limit of manual analysis. This paper attempts to apply an innovative method extended from automatic dialect feature extraction in order to understand dialect transitions. Unlike previous methods, not only multiple features are considered (opposed to the single-feature transition studies) and not only the association to a particular dialect group is stated (as in FA), this novel approach aims to look at which exact (typical) features are present and absent in a group of transitional dialects. The current case study is based on dialects found in the Yue-speaking area in Southern China.

This paper is structured as follows. Section 2 gives an overview to the current issues in the study of transitional dialects, and Section 3 introduces the data used in this study and the methodology. Section 4 dives into the analysis of transitional dialects of Yue-dialects followed by a discussion in Section 5. Lastly, the paper concludes in Section 6.

2. Transitional Dialects

Speech varieties in space are often found not having abrupt boundaries, but rather a gradual transition from one area to another, forming a continuum (Chambers and Trudgill 1998, Taeldeman 2013). Since dialects[1] are not discrete varieties in a continuum, it is not an easy task to distinguish a group of relatively homogenous dialects from another. One should be cautious when they attempt to divide groups of dialects into discrete categories, as this is considered as an act of deliberation (Pickl 2016). Dialectologists speak of 'core' dialects to refer to dialects which are "structurally co-

[1] In this study, a *dialect* is defined as a the variety of the locality in which we have data for, i.e. a locality in dialect survey. This is based on the view of the linguistic geographers, that each location has its own dialect (Wiesinger 1983). Original text from Wiesinger (1983): "Nach sprachgeographischer Anschauung besitzt daher jeder Ort seinen eigenen Dialekt." A collection of dialect is then referred to as a dialect group. Sociolinguists might prefer 'variety' (Chambers and Trudgill 1998, 5) or 'lect' (Macaulay 2010, 63) instead of 'dialect', as these are considered as more neutral terms.

herent and maximally distinct", opposed to transitional dialects which are mixed and features from two (or more) core areas can co-occur (Taeldeman and Hinskens 2013).

2.1 Core vs. Transitional dialects

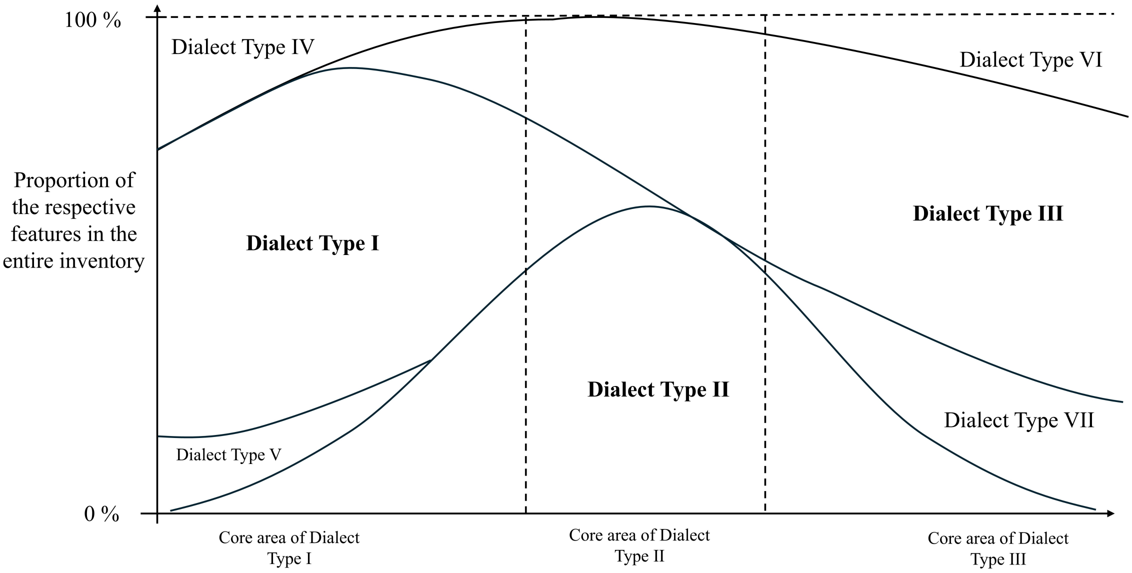

100 %
0 %
Proportion of the respective features in the entire inventory
Dialect Type IV
Dialect Type VI
Dialect Type I
Dialect Type III
Dialect Type II
Dialect Type V
Dialect Type VII
Core area of Dialect Type I
Core area of Dialect Type II
Core area of Dialect Type III

Figure 1: Transition and overlap of dialect types (reproduced based on Pickl 2013, 68)

In order to understand dialect transitions, it is necessary to understand what cores are, otherwise, there is nothing to 'transit' from and to. 'Structural coherence' (Taeldeman and Hinskens 2013) can be understood as possessing a (relatively) distinct set of features of its own.[2] Berruto (2010, 236) argues that varieties are "concentration areas", where they are "identified by a particular frequency of certain variants, by the co-occurrence of several features and possibly by some diagnostic traits, which appear in that variety only". However, such a variety is not "clearly-cut from other varieties". Bach (1950) also argued for using co-occurrence to define (regional) 'dialect', but he also added that the co-occurrence can include both features which are exclusively found in the core region, as well as features which have a much wider distribution.

The idea of 'core' vs. 'transitional' are also recognised through the mathematical, cognitive and perceptual perspectives. From the mathematical perspective, one could categorise dialects as elements, and the groups which the elements belong to are the sets (from set theory, Girard and Larmouth 1993, 108). In an ideal world, where dialects do not show any transitions (for example, not having mixed properties), the set membership of the dialects are either '0' or '1'. The traits which dialects possess are cues to the categories. However, when the elements display heterogeneity and mixture (of features from different sets), which is what we often find in the dialect data, heterogeneity and mixture of dialects suggest the interpretation of dialect areas as 'fuzzy sets', meaning the set membership of any element can now range from '0' to '1'. The more traits of one set a dialect has, the more likely the dialect belongs to a particular set. The view of dialect areas as fuzzy sets has

[2] Linguistic coherence is also a big topic in sociolinguistics. Often referred to as *coherence* alone in sociolinguistics, it concerns to what extent "multiple co-existing linguistic variables have similar distributions, both internally and in the speech community" (Hinskens and Guy 2016). In the case of the latter, sociolinguists are interested in whether the 'sociolinguistic isoglosses' coincide between different sociolects. There is a lot in common between the research in coherence and dialectometry, which proposes an aggregate analysis that "encompasses as much of the variation between language varieties as possible rather than concentrating on single linguistic features" (Nerbonne 2010, 476).

some resemblance with prototype categories in Cognitive Linguistics. The prototype view of lectal gradience sees some varieties being 'typical', 'central' or 'better examples' of a given variety than others (Kristiansen 2008, 59). Preston (1988), on the other hand, attempted to find the limit of Southern dialects in the US through asking informants to draw a map, delimited the area which they consider as 'Southern'. By overlaying all the responses (the drawn area of perceived 'Southern' dialects) from the informants, the aggregate perceptual map also shows an increasing density of agreement of a 'core' dialect area. The attempts above in capturing dialectal transition in space all point to the direction that something is shared in a core region, and as we move away from this region, the dialects gradually show less resemblance of the core dialects.

Given that individual dialects can have different degrees of association with a certain dialect group, with some dialects being more 'typical' than others (based on the idea of fuzzy sets and prototypes, Girard and Larmouth 1993, Kristiansen 2008), Pickl (2013, 2016) established a concept called *dialect type* to model transition in geographical variation from one area to another. A dialect type is constructed on the basis of co-occurring features, and based on dialects which are "more or less like a *typical* variety" (similar to the concept of a dialect area). A dialect can display features from multiple dialect types, and a dialect is not required to belong 100% to a dialect type. This is illustrated in Figure 1. In the core area of Dialect Type II, Dialect Type I and II (and a lesser extent Type VI) are still present. In this model, core areas are defined by the dominant dialect type that a dialect belongs to (some may not have a dominant). In Figure 1, Dialect Type I, II and III each shows dominance in a certain confined region. Hence, in this figure, three core areas are identified, and labelled based on these dominant Dialect Types. When projected to a map, dialect types which dialects belong to become fuzzy dialect areas, as the membership of dialect types are graded.

The full set of features (which in reality may not be fully found in any dialect) form a prototypical variety of the dialect type. From the perspective of dialect types, any local dialect (dialect of one location) can then belong to one or more dialect types (with varying proportion), depending on which features it possesses from each dialect type. Transition, then, is the (gradual) varying proportion of dialect types (in overlap) which local dialects belong to. The area (group of dialects) which has the highest content of a particular dialect type is the core area, as illustrated in Figure 1.

2.2 Types of transitional dialects

Dialects can show transition in more than one way. Chamber and Trudgill (1998) for instance examined several variables of British English, and found that there are several ways dialects can transition from one area to another. The following examples come from transitional dialects which are spoken between two 'pure' dialect areas, i.e. dialects which show consistent usage of one variant of the same variable.

Mixed dialects show the co-occurrence of variants from both sides of the pure dialect areas. In Chamber and Trudgill's (1998, 110) example, South-eastern British English dialects show variation in the variable (u), with mainly two variants: [ʊ] and [ʌ]. In the southern pure dialect region of this area, [ʌ] predominates and in the north, [ʊ]. Some dialects in between the pure dialect regions show a mixture of both variants, this suggests the mixed dialect pattern. *Fudged dialects* on the other hand are transitional in a sense that a phonetically-medial form of the variants in the two pure dialect areas are found in such dialects. This is also found in Chamber and Trudgill's (1998) study of (u), where a third variant [ɤ] is identified. [ɤ] is phonetically intermediate between [ʊ] and [ʌ]. Furthermore, the combination of the mixed and fudged dialect variants form a third type of transitional dialects, namely *scrambled dialects* (Chambers and Trudgill 1998, 117). The types of transitional dialects are summarised in Table 1 below.

In addition to the types of transitional dialects mentioned in Table 1, the distribution of variants can also be conditioned by phonological rules, the lexicon, as well as social factors (Taeldeman 1989). For phonological conditioning for instance, in the intermediary dialects between East and West Flemish dialects, ɡ-deletion occur before -ən, but not -ə (e.g. in Waregem, O 80). In lexically-

Type	Pure Dialect X	Transitional Dialect	Pure Dialect Y
Mixed	Variant 1	Variant 1 & 2	Variant 2
Fudged	Variant 1	Variant ½	Variant 2
Scrambled	Variant 1	Variant 1, 2 & ½	Variant 2

Table 1: Types of Transitional Dialects (based on Taeldeman 1989)

conditioned transitions, also known as *lexical diffusion*, the distribution of a variant is determined by the lexical item. For instance, in West Flemish (e.g. in Oedelem, I 153), West Germanic *a before r is realised as [ɑ], whereas in East Flemish (e.g. in Adegem, I 155), it is [æ]. In Kleit (I 154a), an intermediary dialect between Oedelem and Adegem, *spar* 'firtree' is pronounced with [ɑ], but *kar* 'kart' is with [æ]. Lastly, social factors can also condition the usage of certain variants. In East Flemish, West Germanic *au before a labial or velar is reflected as [yə], and in West Flemish, [uə], e.g. E.Fl. [ryək] vs. W. Fl. [ruək] *rook* 'smoke'. In Deerlijk (O 82), another intermediary dialect between East and West Flemish, it has been found that both variants, [yə] and [uə] are found, by 60% and 40% respectively. The two variants are found even in the same lexical items, and the distribution appears to depend on the social profile of the speakers.

2.3 Multivariate Approaches

The studies in the previous section explored transitionality based on single features. Since dialect varieties do not only consist of one single feature, it is questionable how representative the transition patterns found based on single-variant studies are in a variety.

Simmons's (2012) analysed 10 dialects in the transition zone between Mandarin and Wu in the Danyang area in China. Based on the analysis of 8 features, Simmons (2012, 290) concluded that as we move closer to the core region, dialects "conform more strongly to the core type". In addition, sound changes also show more consistency, being a neater reflection of a Neogrammarian sound change (sound change without (lexical) exceptions). Between the two core regions (of Mandarin and Wu), since there is less consistency of sound changes (lexical diffusion is found more often), such mixing resulted in dialect transition. Simmons's study is a step forward from the single-variant analysis, as evidence of dialect transition is not only found in one variable, but multiple. However, this approach still suffers from two major pitfalls, namely the need to select features, which might lead to a bias in the phenomenon as well as potentially overlooking some features which are relevant. Moreover, there is a limit to how much one can analyse with a big pool of features. These are the motivations to turn to using computational approaches to process mass dialect data.

Dialectometry is a computational branch of dialectology. Dialectometrists argued for an aggregate approach to the analysis (Nerbonne 2010), which encourages making use of more features in an analysis in order to avoid a generalisation of dialectal variation based on a handful of carefully-selected features. The use of computational methods in dialectometry allows analysts to process a large amount of data in one single analysis, in a relatively short period of time. The visualisation techniques, including quantitative maps, enable dialectometrists to understand the underlying patterns within the mass amount of data.

On the line of distance-based approaches, Heeringa (2004) uses multidimensional scaling (MDS) maps[3] to visualise the continuum-like variation of Dutch dialects by projecting the coordinates of the first three dimensions with the RGB colour spectrum. Such a map can be found in Figure 2.[4] Although an MDS map can visualise transitions from one dialect area to another (e.g. yellowish area in Flemish gradually becoming green in the Belgian Brabants area in Figure 2), it cannot tell us which features are present in these transitional dialects. Thus, we are unable to speculate how

[3] MDS is a dimension reduction technique which reduces high dimensional data (of variation, as in Heeringa (2004), to a lower dimension.

[4] Map created using one of the demos on *Gabmap* (Nerbonne et al. 2011, Leinonen et al. 2016), based on the GTRP data.

dialects transition from one area to another. This also implies that it is unclear whether we are seeing a gradual disappearance of a set of features which are closely associated to one group (like mixed dialects). Heeringa (2004) additionally uses stepwise correlation with word segmental distances to find which groups of words are responsible for the variation we see on the MDS map. However, this method does not immediately indicate the individual segmental features, but the entire word as an output characteristic feature. One still has to inspect multiple words and identify the transitioning features, which brings us back to the problem of manual analysis of features.

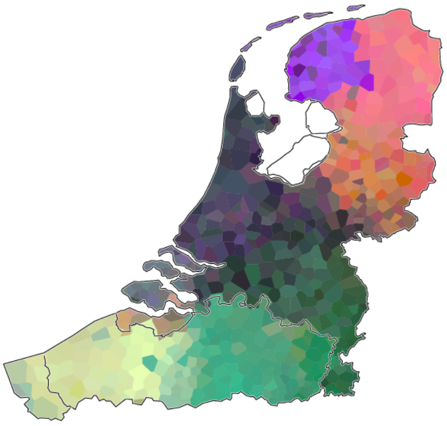


Figure 2: Continuum (MDS) Map of Dutch Dialects

Pickl (2013) and Pickl (2016) on the other hand used Factor Analysis (FA), a dimensionality reduction technique, to study dialect transitions.[5] FA identifies condensations of co-occurring variants and returns factor loadings for each location, which indicate the relationship between the locality and the factors, as well as the factor scores, which indicate the degree of association each variant of a feature has to the factors.

Another motivation for the development of a novel method for studying dialect transition is that it is not immediately clear which features are found in the dialect directly from the output of FA, nor from the original feature matrix. Hence, it is still unclear whether there is a link between the importance (ranked by the Factor score) of the feature and their behaviour in the transitional dialects without further procedures with the data file and analyses. For instance, are more typical features first to be lost as we get away from the core area? Or rather, typical features tend to stay persistent, even in dialects that are further away of the core dialects? Such questions cannot be answered with FA directly, and more analyses must be done. Moreover, FA does not extract the most exclusive dialect features as the most important features of a dialect group, as Sung and Prokić (2024a) found. If we are interested in the behaviour of the most typical dialect features in transitional dialects, one should use an alternative method for feature extraction, such as normalised pointwise mutual information (nPMI), which prioritises exclusiveness in the process (Sung and Prokić 2024a).

Based on the current research gaps, this paper shall address the following research question: 1) what does dialect transition look like based on their top characteristic features? and 2) do transitional dialects always adopt the most typical features from a certain dialect group? The questions stated above are explored through a novel method, namely *dialect typicality decay analysis* (dialect typicality analysis[6] hereafter for brevity), which is an extension of the feature extraction

[5] Although Pickl (2013,2016) focus on lexical variation, Pröll (2015) has illustrated that the same method can be applied to phonetic data, given that they are categorical.

[6] The term ‘dialect typicality analysis’ was originally used in Sung and Prokić (2024b) for using the count of characteristic features of a certain dialect group in all the dialects to estimate the kernel or core area. Despite the goals are very similar to the dialect typicality decay analysis, Sung and Prokić (2024b) only consider the typicality based on one threshold of

technique with nPMI by Sung and Prokić (Sung and Prokić 2024a). The procedures of dialect typicality analysis are provided in the following section.

3. Data and Methodology

The data of the case study used in this paper comes from the dialects of the Yue language, spoken in Southern China. These dialect data were transcribed during various dialect surveys and individual studies since the 1980s to the 2010s. In terms of the methodology, a series of procedures are required for a dialect typicality analysis. These include dialect classification, feature extraction and typicality analysis. In the following subsections, the data used in the current study are introduced, followed by a description of the methodology.

3.1 Data

The data of the current study comes from Yue dialects. Yue is a collection of dialects from the Sinitic family, mainly spoken in Southern China, namely Guangdong and Guangxi provinces (CASS 2012). Yue has been found to share general variation patterns with European languages on the segmental level, namely dialects form a continuum, and variation can also be explained by geographical and historical-social correlates (Sung and Prokić 2023, Sung forthcoming).

Recently, a dataset consisting of 104 Yue dialects has been digitised (Sung et al. 2024). This dataset contains IPA transcriptions of around 130 monosyllabic words per dialect. The sources of these transcriptions include dialect surveys (*Survey of Dialects in the Pearl River Delta* (Zhan and Cheung 1987), *Survey of Yue Dialects in Northern Guangdong* (Zhan and Cheung 1994), *Survey of Yue Dialects in Western Guangdong* (Zhan and Cheung 1998), *The Phonological Study of the Yue Dialects spoken in the Zhan-Mao area in Western Guangdong* (Shao 2016), *Chinese Dialect Research in the Guangxi Province* (Xie 2007), *Yue, Pinghua and Tuhua Dialect Survey Collection Part 1* (Chen and Lin 2009)), and other individual studies (Liu 2015, Zhong 2015, Huang 2006, Chen 2009, Yang 2013, Tan 2017, Shi 2009 and Chen and Weng 2010).

This dataset also contains transcriptions for tones, but this study only considers the segmental part of the data, and thus addresses issues in segments. Lastly, the geographical location of the dialects in the data can be found in Figure 4.

3.2 Methodology

The procedures involved for this study can be taken apart into the following steps: 1) multiple sequence alignment, 2) distance calculation, 3) cluster analysis, 4) feature extraction and 5) typicality comparison. The first four steps follow the workflow in Sung and Prokić (2024a). The final step is an extension to the feature extraction approach.

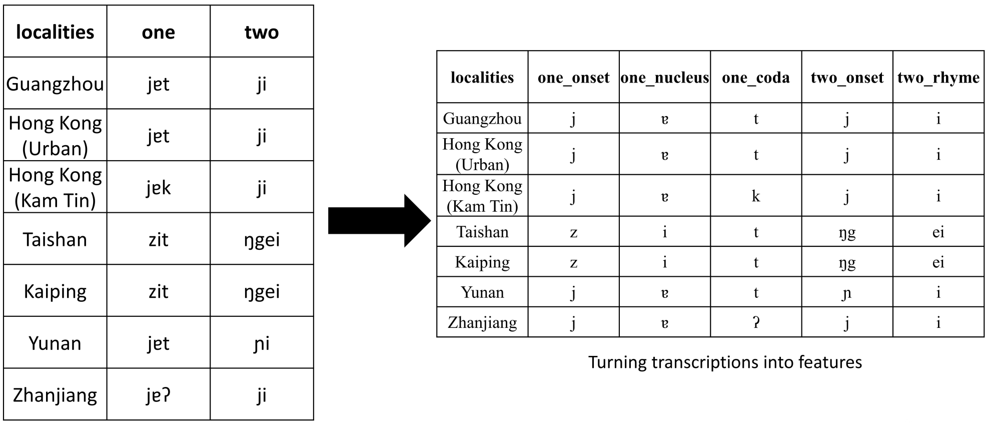
3.2.1 Multiple Sequence Alignment

Multiple sequence alignment (MSA) is an important step in dialectometry if we are going beyond simply seeking a classification of dialects without explanation. MSA is a semi-automated procedure which breaks segments apart and aligns (diachronically) corresponding segments. An illustration of MSA is given in Figure 3, using examples from Yue dialects. To start with, this example shows two words from the Yue data, 'one' and 'two' in the columns on the left, in seven different Yue dialects. The transcriptions of these two words are broken down into individual segments or clusters of consonants (e.g. [ŋg] in Taishan) and vowels (e.g. the diphthongs in Taishan), shown in the columns on the right. Insertions and deletions (related to historical epenthesis and elision of sounds

feature counts (e.g. all the features with an nPMI score above 0), whereas in the current paper, typicalities based on different number of features considered are analysed in the same analysis (see Section 3.2.5 for the explanation).

respectively) can also be represented. When a sound is missing (due to the lack of epenthesised segment or an elision of a segment), a '-' is added to the slot.

MSA is performed using the Python *LingPy* library (List et al. 2021), followed by manual checking for potential misaligned segments. The end result (as shown on the table on the right in Figure 3) contains columns of aligned features which are referred to as synchronic 'dialect features', but they are also diachronically related.



localities	one	two
Guangzhou	jɐt	ji
Hong Kong (Urban)	jɐt	ji
Hong Kong (Kam Tin)	jɐk	ji
Taishan	zit	ŋgei
Kaiping	zit	ŋgei
Yunan	jɐt	ɲi
Zhanjiang	jɐʔ	ji

localities	one_onset	one_nucleus	one_coda	two_onset	two_rhyme
Guangzhou	j	ɐ	t	j	i
Hong Kong (Urban)	j	ɐ	t	j	i
Hong Kong (Kam Tin)	j	ɐ	k	j	i
Taishan	z	i	t	ŋg	ei
Kaiping	z	i	t	ŋg	ei
Yunan	j	ɐ	t	ɲ	i
Zhanjiang	j	ɐ	ʔ	j	i

Turning transcriptions into features

Figure 3: Example of Multiple Sequence Alignment

Sung and Prokić (2024a) have argued that using MSA in feature extraction with certain methods, e.g. normalised Pointwise Mutual Information (nPMI, see Section 3.2.4 below) can return a precise feature which corresponds closely with a certain dialect group. Since segments are all diachronically related within each column, it also allows further diachronic interpretation of each dialect feature.

3.2.2 Distance Calculation

Distance calculation is a step which converts qualitative data (columns of multi-aligned dialect features), into quantitative data (dialect distances). This step is also necessary for feature extraction, because the method for feature extraction, normalised pointwise mutual information (nPMI), is a top-down feature extraction method (Sung and Prokić 2024a), and it requires an existing dialect classification as an input, before it can extract features. As the traditional classification of Yue dialects is not justified either methodologically or in terms of their criteria (Sung forthcoming, Sung and Prokić accepted), a more objective approach in seeking dialect groups of Yue is taken, namely with cluster analysis based on the dialect distances with more features considered. Cluster analysis requires pairwise distances as an input, and thus distance calculation for the Yue dialects is necessary. Since dialect features are categorical (each segmental value is a category of its own), a suitable method for the multi-aligned transcriptions is *Relative Distance Value* (*RDV*, Goebl 2018).

The formula for RDV is provided in (1) below. To calculate RDV in a pairwise comparison (between two dialects), the number of non-matching features or Co-difference (COD in (1)) is divided by the total number of features compared, i.e. the number of matching columns or Co-identity (COI in (1)) plus the number of unmatching columns or Co-difference (COD in (1)). The resulting value is the difference of a pair of dialects ranging from 0 to 1.

$$RDV_{jk} = \frac{\sum COD_{jk}}{\sum COI_{jk} + \sum COD_{jk}} \quad or \quad \frac{no.\ of\ unshared\ features\ in\ both\ dialects}{total\ no.\ of\ features\ compared} \tag{1}$$

The calculation of RDV was applied to all the pairs of dialects in the dataset, and the distances are stored in a distance matrix and analysed by means of cluster analysis.

3.2.3 Cluster Analysis

Cluster analysis refers to the partition of objects (dialects in our case) into groups (Manning and Schütze 1999). Cluster analysis helps us to group dialects which are similar enough to be considered as the same group. There are numerous cluster algorithms which can be used for this purpose, but out of these algorithms, agglomerative hierarchical clustering algorithms are most extensively used within dialectometry. These algorithms find successive clusters based on previously established clusters, creating a hierarchical representation of the clusters in a dataset.

For the current study, Ward' s method (Ward 1963) has been chosen. This algorithm is also known as the minimal variance method. Ward's algorithm merges clusters which will yield the smallest increase in the sum of the square distances of each element from its cluster' s mean. There are a number of reasons why the Ward's method is used in the current analysis. First of all, out of the three popular agglomerative hierarchical cluster algorithms (UPGMA, Complete Linkage and Ward's method), UPGMA prioritises finding island-like dialects as clusters. In the current dataset, the first few clusters that split often contain only one member, sometimes two. While the UPGMA representation of dialect distances is not wrong, it is not useful for the purpose of understanding transitional dialects (as it requires at least a couple of dialects to see transitions). On the other hand, complete linkage and Ward's method yield somewhat more similar results.[7] However, Ward's method shows a higher degree of similarity with the traditional classification (LAC, CASS 2012, based on the author's interpretation and judgement). Using Ward's cluster solutions can shed lights in both the dialectometric classification, as well as the traditional classification (see below for the similarity between the cluster solution and the traditional classification). In addition, in Prokić and Nerbonne' s (2008) evaluation of a number of cluster algorithms (including those mentioned above), Ward' s method yields relatively good performance in external validation using *Modified Rand Index* (Hubert and Arabie 1985) and *Entropy* (Zhao and Karypis 2001).

In this study, the 5-cluster solution has been chosen[8], as illustrated in Figure 4.[9] The current classification has shown resemblance with the traditional classification, as depicted in the *Language Atlas of China* (LAC, CASS 2012), and the differences largely fall into merging some traditional dialect groups into one bigger cluster. The number of groups in the current analysis reduces from 8 dialect groups to 5 groups. The *Guangfu* dialects under the new classification consists of 'Guangfu' and 'Yongxun' dialects in the *LAC*; *Coastal* dialects include 'Gaoyang', 'Wuhua' and 'Qinlian' dialects in the old classification. 'Guinan Pinghua' is now called *Western* dialects, since they are located in the western side of the Yue continuum. Lastly, *Goulou* dialects largely overlap with the LAC Goulou area, hence the same label remains. The preference to a dialectometric classification is justified by the following reasons: 1) the dialectometric approach makes use of as many features available in the transcriptions as possible, avoiding a subjective selection of features and 2) it allows us to see the global patterns of the dialect landscape.

3.2.4 Feature Extraction

There are other existing approaches in automatic dialect feature extraction, in addition to Heeringa (2004) mentioned in Section 2.3. For instance, Prokić et al. (2012) use an approach similar to Fisher's linear discriminant to find characteristic words of different dialect areas. Wieling and Nerbonne (2010, 2011) on the other hand use bipartite spectral graph partitioning to identify sound correspondences (between a reference variety and a dialect) which are characteristic to different

[7] For the following analysis, two dialect groups, Guangfu and Siyi, are both identified by both cluster algorithms.

[8] It should be noted that the feature extraction method (nPMI) can be used for other cluster solutions, as well as other classifications and non-phonetic features too. Other cluster solutions (with other algorithms) could also be explored.

[9] Distance calculation and cluster analysis were performed using *Gabmap* (Nerbonne et al. 2011, Leinonen et al. 2016).

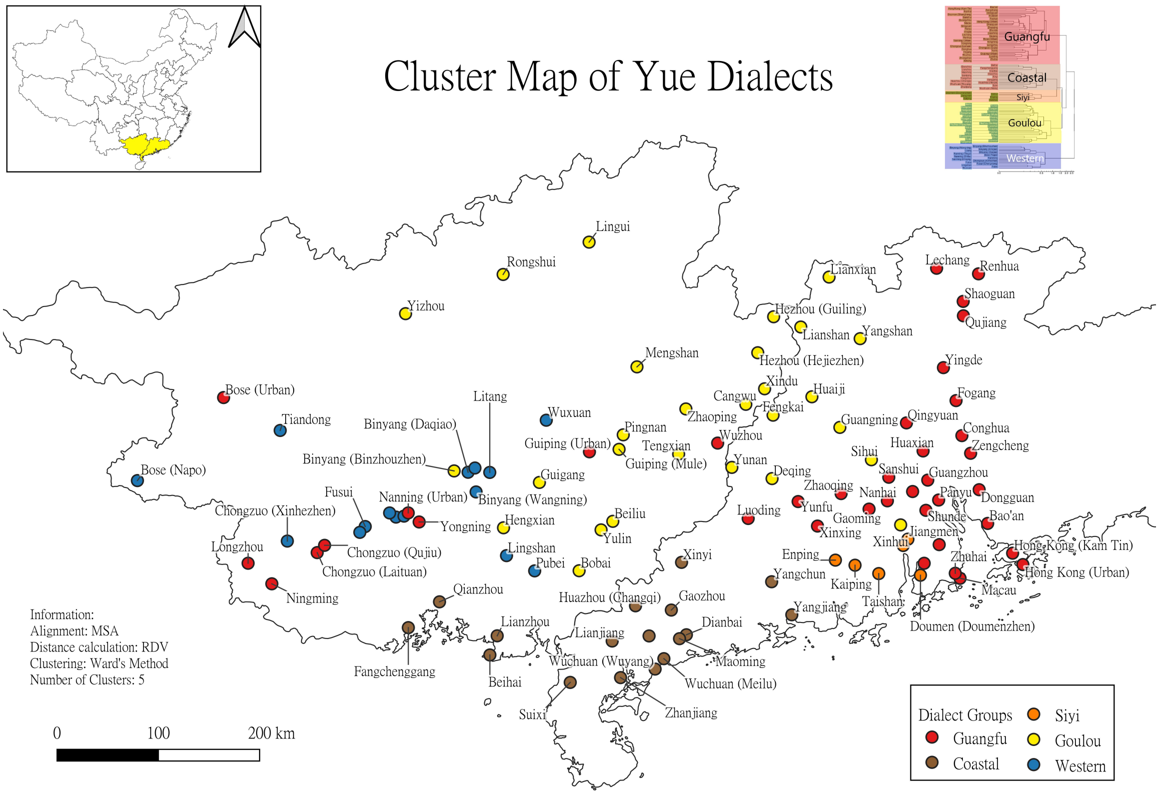


Figure 4: Cluster Map of Yue Dialects in Guangdong and Guangxi.

dialect groups. More recently, Sung and Prokić (2024a) and Rubehn et al. (2024) both use multiple sequence alignment and normalised pointwise mutual information in their feature extraction procedures. However, both approaches have a major difference: the former uses dialect variants directly to find the characteristic features of a certain dialect group, whereas for the latter, sounds correspondences are used as the input features. The difference produces different outputs in the results.

The feature extraction procedure follows Sung and Prokić (2024a). Sung and Prokić's (2024a) method is less computationally heavy, and the results are easier to process and be further applied for the following typicality analysis. The choice in method does not rule out that the results from Rubehn et al.'s (2024) methodology can be adapted to the dialect typicality analysis below. However, an additional advantage with Sung and Prokić's (2024a) approach is that it can also be applied to categorical features from other linguistic levels (e.g. lexis, morpho-syntax). An illustration of the proposed typicality analysis is therefore more suitable with Sung and Prokić's (2024a) approach, which has a broader applicability.

Pointwise Mutual Information (PMI) is an association measure based on co-occurrence and probabilities (Church and Hanks 1990). This association measure was originally used in seeking word associations (such as collocations and keyword analysis), but it has also gained popularity for the automatic inference of segment distances in dialectometry (Wieling et al. 2011). The idea behind PMI in feature extraction is the comparison of the probability of observing two categories, the dialect variant and the dialect group, together (joint probability) and independently (expected or by chance probability). If there is an actual association between the variant and dialect group, then the joint probability would be much greater than their probability together by chance (Church and Hanks 1990, 23).

The required probabilities (how often a variant and a dialect group co-occur together) include probability of variant x or *p(x)* found within one column in the MSA data; the probability of dialect group y or *p(y)* within the classification column based on the cluster analysis in Section 3.2.3. The final probability that PMI requires is the co-occurrence of variant x given dialect group y, or *p(x,y)*.

$$pmi(x,y) = log_2 \frac{p(x,y)}{p(x)p(y)} \tag{2}$$

The PMI scores are calculated using the formula given in (2) (Church and Hanks 1990), which is the log base 2 of the probability of the co-occurrence of a given variant and a given group, out of all the possible instances that they could co-occur in the data. The score is then normalised following Bouma (2009), presented in (3).

$$npmi(x,y) = \frac{pmi(x,y)}{-log_2 p(x,y)} \tag{3}$$

The steps described above are iterated for all the variants found in the same column, and for each dialect group in the classification label column. When the nPMI score for all the variants and dialect groups have been processed for the first column, the same procedures are iterated until the last column of the MSA data has been processed.

In addition, two additional indices are calculated, namely the *Exclusivity*[10] and *Representativeness* for each of the features extracted (Sung and Prokić 2024a). *Exclusivity* concerns the extent to which a specific variant is only found within the given cluster and *Representativeness* on the other hand calculates the number of dialects within the cluster which has the specific variant. These metrics give us a rough idea how distinctive these features are for a certain dialect group in relationship to the whole dialect area.

The output of the extracted features is a table of features, ranked by the nPMI scores, from high to low. An example is given for the Guangfu dialects (red circles in Figure 4) in Table 2 below.

3.2.5 Dialect Typicality Calculation

Dialect typicality refers to how much a particular dialect belongs to a particular dialect group (Pickl 2016, 78). In the last volume of the *Atlas Linguistique et ethnographique de la Gascogne*, Séguy (1973) has included numerous dialectometric maps, one of them being the *Champ gradient de la Gasconité [Gradient map of Gasconity]* (Map 2530, also Figure 5). Séguy had manually identified 40 specific features of Gascon (which contrast with the rest of the Occitan and Gallo-Roman dialects), and on this map, Gasconity or Gasconness of each dialect is represented by the number of features identified by Séguy it possesses. Séguy's gradient Gasconity map shows a 'core' area near the Southwestern tip of France, where the dialects have the densest amount of typical Gascon features. As we move away from this core area, the density of typical Gascon features decreases in the dialects.

Based on Séguy's (1973) procedures, to induce the dialect typicality of dialects, we need 1) a defined dialect area and 2) features that are exclusive or typical for the dialect area. Cluster analysis from Section 3.2.3 offers us the dialect areas and feature extraction using nPMI from Section 3.2.4 returns a list of typical features[11] for each dialect group. Both automatic procedures were not possible back in Séguy's time. The automation of the two steps above is data-driven, which

[10] 'Exclusivity' is different from 'Distinctiveness' found in Wieling and Nerbonne (2011). Sung and Prokić (2024a) have compared the two indices and they are highly correlated, although the calculation for Distinctiveness is more computational heavy, hence Exclusivity is used instead.

[11] 'Typical' is used interchangeably with 'characteristic' in the current article. Both terms refer to features that are characteristic (as identified by nPMI) to a certain dialect group, without implying being recognisable to the dialect speakers.

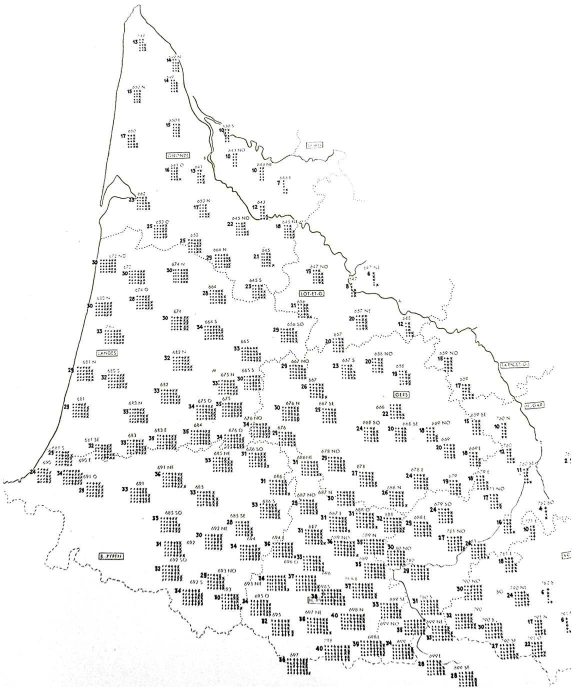


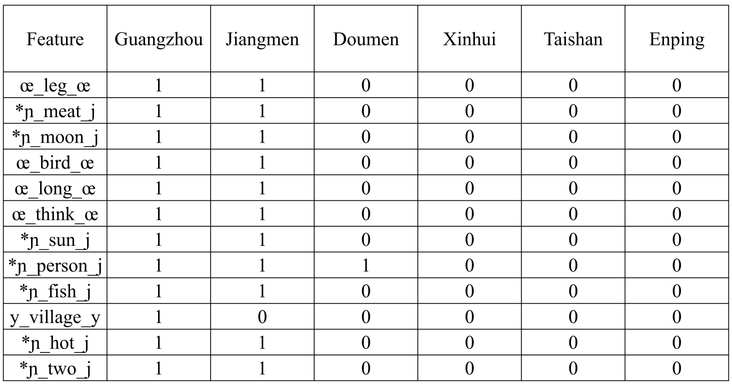
Figure 5: Gradient map of Gasconity from Séguy (1973), Map 2530

avoids subjectively picking features which are considered as important and potentially missing some typical features. Instead, the features extracted are justifiable with the nPMI score, as well as their exclusivity and representativeness indices.

The final step is to calculate dialect typicality for each dialect (for a particular dialect group). This translates to counting how many typical features (identified by nPMI) are found in each dialect in the dataset. The typicality score is normalised by dividing the number of typical features a dialect has by the total number of features considered, which returns a score ranging from 0 to 1, with 1 being the dialect contains all the typical features.

To do the steps above, we first need a ranked table of typical features of a particular dialect group (the output from Section 3.2.4), then we create an m x n binary matrix, which I shall refer to as the *typicality matrix*. In this matrix, m stands for the features in the ranked table, from the most associative feature to the least associative feature, and n stands for the dialects in the data. A subset of the typicality matrix is given below in Figure 6.

Rather than directly calculating the sum of the total number of typical features for each dialect, a typicality matrix gives us more flexibility for a dialect typicality analysis. For instance, given that we are considering 40 features in total, if a dialect has a 50% dialect typicality, that implies it possesses 20 features out of the total 40 features. However, which 20 features are those? This is a very important question to ask for dialectologists, as the table of features is ranked for their association with the dialect group, a possession of different sets of features will tell us different information of the dialects. For example, if a dialect only has the most important features of dialect group A (e.g. the top 10 features, which are also shared innovations), but the rest of the features are different from other group A dialects, this suggests that this dialect is inherently from a different dialect group, but it has adopted features from dialect group A, perhaps due to contact. Since dialect features have different amounts of importance to a particular dialect group, speculating features in the fashion provided above is absolutely key to understanding ways in which dialects transition from one dialect



Feature	Guangzhou	Jiangmen	Doumen	Xinhui	Taishan	Enping
œ_leg_œ	1	1	0	0	0	0
*ɲ_meat_j	1	1	0	0	0	0
*ɲ_moon_j	1	1	0	0	0	0
œ_bird_œ	1	1	0	0	0	0
œ_long_œ	1	1	0	0	0	0
œ_think_œ	1	1	0	0	0	0
*ɲ_sun_j	1	1	0	0	0	0
*ɲ_person_j	1	1	1	0	0	0
*ɲ_fish_j	1	1	0	0	0	0
y_village_y	1	0	0	0	0	0
*ɲ_hot_j	1	1	0	0	0	0
*ɲ_two_j	1	1	0	0	0	0

Figure 6: Typicality matrix of the top 12 Guangfu dialects for the Guangzhou dialect and Siyi dialects

area to another, or more specifically, the ways they adopt or abandon the most characteristic or even less characteristic features. This is linked to the issue the second research question is trying to address, whether the most important feature from another dialect group is always adopted first.

Based on the discussion above, I propose to use a new visualisation technique called a *dialect decay plot* to understand the core dialects in dialect group, as well as transitional dialects. A dialect decay plot does not show the total dialect typicality immediately, but rather, cumulatively. An example is given in Figure 7. On this plot, dialect typicality of each dialect is shown by an increase of 10 features at a time[12] (starting from 10, and stopping at 70 in Figure 7). By showing the cumulative typicality, we can then inspect the changes (if any) with the possession of more or less important features in the typicality matrix.

Figure 7 shows the typicality of the top 5 most Guangfu dialects, from considering 10 to 70 features. It shows that all six Guangfu dialects possess the top 50 Guangfu features extracted using nPMI (hence the overlap of the curves). However, when 70 features are considered, we start to see discrepancies. At 70 features, we see that Foshan retains the highest number of Guangfu features, and the runner up is Macau, followed by Panyu. It should be noted that Guangzhou and Hong Kong (Urban) have the exact same number of features as Macau, hence they are not seen on the plot, as Macau overlaps with these two dialects. The tail we see starting at the 60 feature mark is there because nPMI does not only extract the most exclusive features, but also less exclusive features (the features that are ranked lower, with a lower nPMI score). Hence the plot is named a 'decay' plot, signalling the tail that is found for the most typical dialects. However, despite the slight differences with the curve at the 70 feature mark, the same pattern can still be seen with these five dialects. We can see this shared pattern as a sign for the most typical Guangfu dialects, which can be used as a reference for dialect (typicality) comparison.

4. Dialect Typicality Analysis

In this section, I will demonstrate the dialect typicality analysis on a small area of the Yue-speaking region. There are two dialect groups spoken in the studied area, namely Siyi and Guangfu, and there is potentially a dialect boundary between the two dialect groups in between, because in the literature, Siyi is known as a rather different dialect group in the area (see below). In this section,

[12] The normalised typicality is calculated first by 10 features, then 20, then 30 etc.

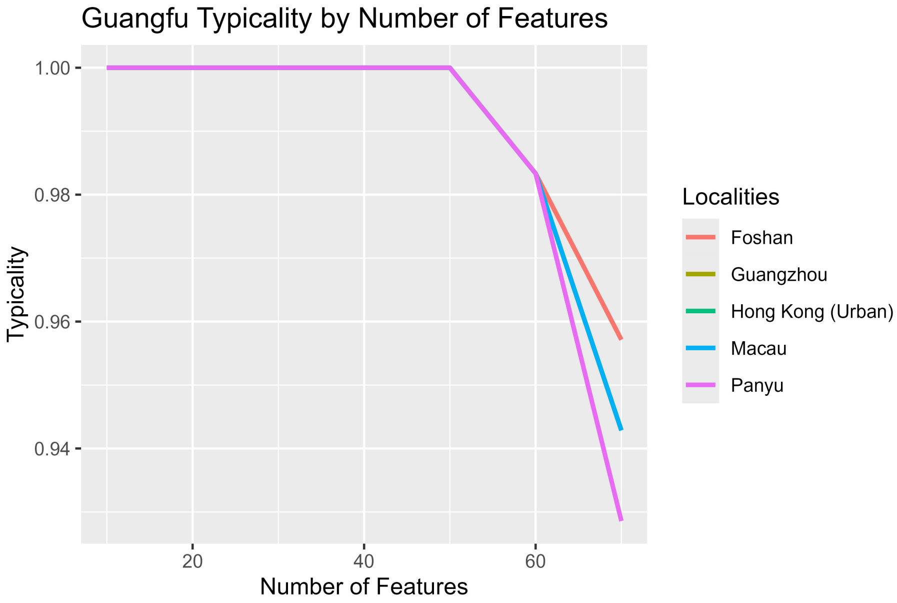


Figure 7: Dialect decay plot of 5 most typical Guangfu dialects

the backgrounds of this area are first given, then the issues for dialect transitionality are explored using the approach introduced in Section 3.2.

4.1 The Siyi area and the surroundings

The Siyi dialects are spoken on the left of the Pearl River Delta in Guangdong, indicated by the orange circles in Figure 8. Siyi dialects are surrounded by the Guangfu dialects (in red) in the North and the East, and Coastal dialects in the West. The Heshan dialect is traditionally considered as part of the Siyi dialect group, but in the cluster analysis, it is grouped closest to the Goulou dialects due to its island like property to its surrounding dialects. In this study, the Heshan dialect is not included in the typicality analysis. The potential transition area is in Figure 8 goes from the Enping-Kaiping area (the west-most extent of the Siyi area) towards the Guangzhou area (the heartland of Guangfu dialects) near the top of the map.

In the literature, the Siyi dialects are often described as a group of dialects with distinctive features from the rest of Yue dialects (e.g. Zhan and Cheung 1990, 18). Despite the difficulties of scholars coming to an agreement in the classification of Yue dialects, agreements have been reached more easily for the Siyi dialects as a separate Yue dialect group (Shao and Gan 1999, 128), which reflects Siyi dialects' distinctiveness. Gan (2003) even proposed that Siyi dialects did not come from the common ancestor of Yue dialects, but the more distantly-related Min and Hakka dialects. The Yue features were adapted by a later Yue-isation[13], due to the later economic and cultural developments of the Guangfu area.

There are not too many discussions on the transitionality of Siyi dialects to Guangfu dialects, but the dialectometric analysis suggests otherwise. In complement to the cluster analysis, which returns a crisp image of dialect partitions, multidimensional scaling is used as a dimensionality reduction technique to visualise the high dimensional distance data in a lower dimension plot, which is suitable for visualising dialect distances. MDS is a method which represents "measurements of similarity (or dissimilarity) among pairs of objects as distances between points" (Borg and Groenen 2005, 3). In

[13] Yue-isation is translated literally from Gan's (2003) terminology '粵化'.

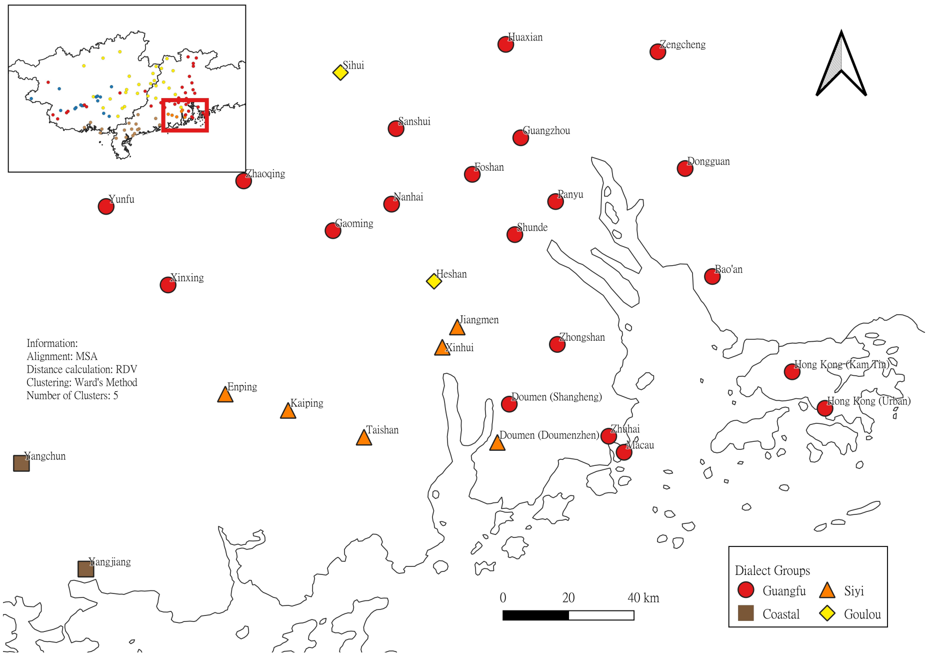


Figure 8: Cluster map of Yue dialects spoken in the Siyi and surrounding areas (zoomed in version of Figure 4, with modified symbols for clarity)

dialectometry, dialect distances are condensed into lower dimensions, often in 2- or 3-dimensions (as they often give a good representation of the original distances), and they are visualised in a scatter plot. Dialects are represented by points in an MDS plot; the further the points are from each other, the more different they are.

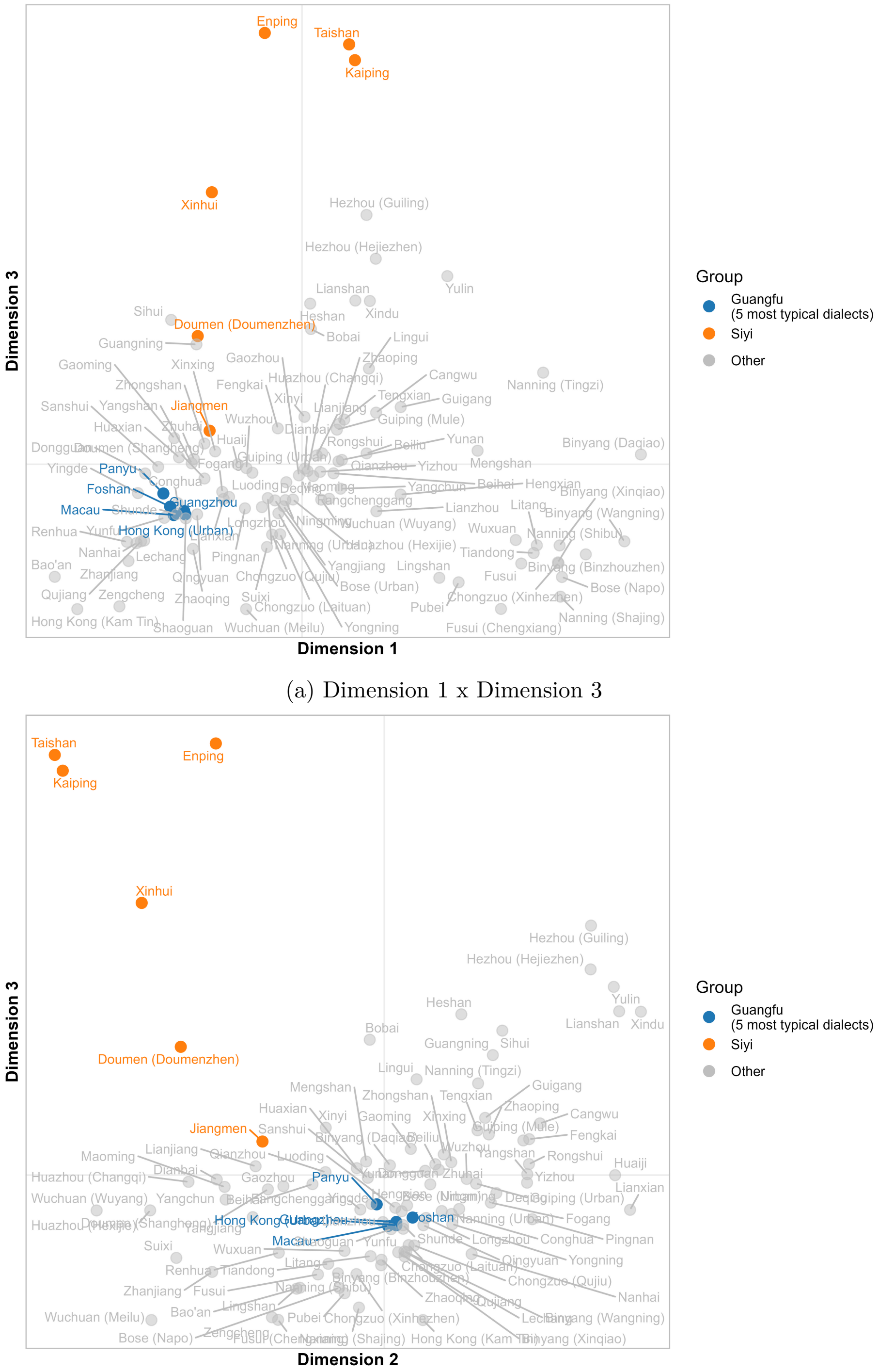
Figure 9a and 9b are 2-dimensional MDS plots of the Yue dialects. These two plots represent a 3-dimensional view[14] of the dialect distances, which explains 66.2% of the variation of the original distance matrix. The Siyi dialects are highlighted in orange in Figure 9a and 9b, whereas the top 5 most Guangfu dialects (as shown in Figure 7) are highlighted in blue, while other dialects are all in gray. We can see from these plots that the top Guangfu dialects are clustered rather close to each other, whereas the Siyi dialects do not really form a cluster, with the possible exception for Taishan and Kaiping. Other Siyi dialects form a continuum towards the blue cluster (as well as other dialects). This suggests a transition between Siyi and Guangfu dialects.[15]

4.2 Dialect Typicality: Guangfu

Since there is a pattern that some Siyi dialects are closer to the most typical Guangfu dialects than others, the first analysis focuses on the dialect typicality based on Guangfu features.

[14] It is difficult to project a 3D plot on a 2D surface, hence two 2-dimensional plots are used in complementary to each other to show the dialect distances.

[15] There are other possibilities for the transition to go towards other dialects, but since Siyi is adjacent to Guangfu dialects, Guangfu is chosen as the initial investigation of transitional dialects. See Figure 8.



(a) Dimension 1 x Dimension 3

(b) Dimension 2 x Dimension 3

Figure 9: Multidimensional Scaling (MDS) plots showing relationships between different dimensions.

First of all, the top 20 Guangfu features can be found below in Table 2.[16] In the first half of the table, the recurring variants are [œ] and [j], which stand for the possession of [œ] before velar codas and the sound change (*ŋj >) *ɲ > j.[17] These two features are the most prominent Guangfu features, indicated by their relatively high exclusiveness and representativeness. Other features (with a lower nPMI score) include having a nucleus vowel [y], Middle Chinese voiced obstruent became devoiced and aspirated (conditioned by the level tone, Feature 15), having a nucleus vowel [ɔ] in 講 'to speak' (Feature 19) and no medial glide (Feature 20).

Rank	Variant	Feature	nPMI	Exclusivity	Representativeness
1	œ	œ_leg	0.709	0.821	0.865
2	j	*ɲ_meat	0.694	0.773	0.919
3	j	*ɲ_moon	0.687	0.8	0.865
4	œ	œ_bird_col	0.685	0.816	0.838
5	œ	œ_long	0.682	0.8	0.865
6	œ	œ_to think	0.682	0.8	0.865
7	j	*ɲ_sun	0.67	0.767	0.892
8	j	*ɲ_person	0.65	0.75	0.892
9	j	*ɲ_fish	0.623	0.756	0.838
10	y	y_village	0.607	0.727	0.865
11	j	*ɲ_hot	0.577	0.744	0.784
12	j	*ɲ_two	0.575	0.757	0.757
13	y	y_all	0.51	0.66	0.838
14	y	y_pig	0.494	0.618	0.919
15	ts^h	*dz_all	0.493	0.6	0.973
16	y	y_feather	0.493	0.6	0.973
17	j	*ɲ_ear	0.472	0.684	0.703
18	y	y_tree	0.465	0.593	0.946
19	ɔ	ɔ_to speak_col	0.462	0.581	0.973
20	∅	leg_m	0.455	0.6	0.892

Table 2: Top 20 Features in Guangfu Yue Dialects

Figure 10 is the typicality decay plot (see Section 3.2.5) based on the Guangfu features (70 features in total). The order of varieties in the legend follow the end points of the typicality curves. Here, the Guangzhou dialect (in red) is used as a reference dialect for the most typical Guangfu dialects, as shown in Figure 7, since they all share very similar patterns. Having one curve for the most typical Guangfu dialects avoids the confusion in the plot.

In Figure 10, we can distinguish four patterns of curves. The first one is Guangzhou, where the typicality remains at 1 until the 50-feature mark, which we have already seen in Figure 7. The second pattern goes to Jiangmen, where it starts off with a high Guangfu typicality, and at the 20-feature mark, it drops sharply, but the typicality score remains above 0.5 all the way to the 70-feature mark. Doumen, Xinhui and Enping form the third pattern. These dialects starts off with a very low typicality score, but after the 20-feature mark, the typicality increases, and they all end with a score of above 0.25 by the 70-feature mark. Lastly, the typicality score of Taishan and Kaiping remain 0.12 or below through the 70 features.

This plot suggests that the Jiangmen dialect contains the most typical Guangfu features, but other than that, the other features present in Jiangmen seem quite different from the most typical

[16] The 70-feature cut off point is chosen because the most typical Guangfu dialects have all top 50 features, and with 70 features, different tails belonging to different dialects start to emerge in Figure 7. The number of features is kept consistent for the Siyi analysis. The full list of the 70 features from both dialect groups can be found in Appendix 1 and 2.

[17] There are multiple instances of the same segments in different lexical items, hence the recurring variants.

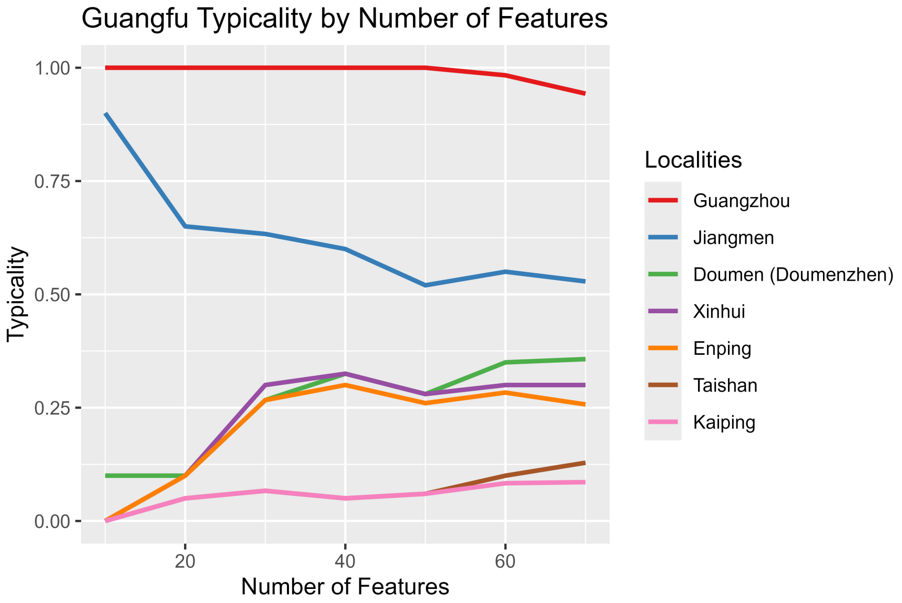


Figure 10: Typicality Decay Plot (based on Guangfu features)

Guangfu dialects. For Doumen, Xinhui and Enping, they seem to possess the less exclusive features of Guangfu (since the typicality only rises later, implying that only when more (less important) features are considered, we see the increase of Guangfuness). Lastly, Taishan and Kaiping barely have any Guangfu features.

In order to understand precisely which features these dialects have, we need to go back to the typicality matrix (see Figure 6). The Jiangmen dialect has remained high in the Guangfu typicality until the 30-feature mark. It turns out that the Jiangmen dialect has both [œ] before velar codas and the sound change (*ŋj- >) *ɲ- > j-, but not the other features for the first 20 features. If we look further, we can see that it also has an [s] instead of a lateral fricative. Up to the top 50 features (which the most typical Guangfu dialects all have), these three features that Jiangmen possesses are very categorical (i.e. no patterns of lexical diffusion or exceptions found). The results are summarised in Table 3 below. However, it should be noted that Jiangmen does not have 24 (lower ranked) features out of these 50 features. Based on these results, Jiangmen seems to be a type of dialect which has adopted the most exclusive Guangfu features. However, this is only one side of the story. Without examining how many typical Siyi dialects it also has, we cannot observe its transitionality fully. The analysis of the Jiangmen dialect with the perspective of the Siyi dialects can be found in Section 4.3 below.

Feature	Tokens in dialect	Total tokens
œ before velar coda	5	5
ɲ > j	10	10
s onset	9	9
Other shared features	3	3

Table 3: Guangfu features found in the Jiangmen dialect

In terms of Doumen, Xinhui and Enping, both Xinhui and Enping almost have a complementary set of Guangfu features (out of top 50) with Jiangmen, except for having the [s] onset instead of lateral fricatives. For Doumen, in most cases, the features it has on this list are identical with Xinhui and Enping, except it has three instances of *ɲ > j (one of it can be seen in Figure 6). This explains why it has a higher starting Guangfu typicality than the other two dialects (at the 10-feature mark). These three dialects are rather different from Jiangmen. The only similarity is that they all share the [s] onset, and not the lateral fricative.

Lastly, out of the top 50 Guangfu features, both Taishan and Kaiping only have three of these features. This explains why the two curves in Figure 10 stay low throughout.

4.3 Dialect Typicality: Siyi

Rank	Variant	Feature	nPMI	Exclusivity	Representativeness
1	h	*d_head	0.946	0.857	1
2	h	*tʰ_soil/earth	0.946	0.857	1
3	p	*-t_knee	0.94	1	0.833
4	ŋg	*ɲ_meat	0.94	1	0.833
5	ŋg	*ɲ_sun	0.94	1	0.833
6	ŋg	*ɲ_hot	0.88	0.833	0.833
7	ŋg	*ɲ_moon	0.88	0.833	0.833
8	i	night_m	0.875	1	0.667
9	ŋg	*ɲ_person	0.875	1	0.667
10	ŋg	*ɲ_to drink	0.875	1	0.667
11	ŋg	*ɲ_ear	0.829	0.714	0.833
12	ŋg	*ɲ_fish	0.829	0.714	0.833
13	ŋg	*ɲ_two	0.829	0.714	0.833
14	z	*ɲ_enter	0.807	0.8	0.667
15	z	j_feather	0.807	0.8	0.667
16	z	j_leaf	0.807	0.8	0.667
17	z	j_night	0.807	0.8	0.667
18	z	j_one	0.807	0.8	0.667
19	z	j_rain	0.807	0.8	0.667
20	z	j_round	0.807	0.8	0.667

Table 4: Top 20 Features in Siyi Yue Dialects

In order to gain a fuller picture of the relationship between the Siyi and core Guangfu dialects, an analysis should be done from the perspective of the Siyi dialects. Other than seeing whether Siyi dialects (as identified by the cluster analysis) adopted Guangfu features (as we have seen in Section 4.2), whether some Siyi dialects lose some typical features should also be investigated.

The top features which are closely associated with the Siyi dialects can be found in Table 4. These include Middle Chinese *d- and *tʰ- > h- (Feature 1-2), coda *-t > -p (Feature 3), *ŋj- > ŋg- (Feature 4-7, 9-13), the presence of medial -i- in the word ‘night’ (Feature 8) and the presence of *j- > z- (Feature 14-20).

All of these features have high exclusivity, and half of them have very high representativeness as well. This indicates that these top features are quite unique for this dialect group.

In Figure 11, instead of seeing bundles, it is more like we see layers of lines, deviating from the tight bundle of Siyi dialects consisting Enping, Kaiping and Taishan. These three dialects have high typicalities throughout the 70 features, and can hence be considered as the most typical varieties of the Siyi dialects. Other dialects except Guangzhou, on the other hand, show a similar line pattern, except their starting point varies (gradually lower, from Xinhui to Jiangmen). This suggests they all shared some most typical Siyi features to start with, but they all have a different amount of Siyi typicality. Lastly, Guangzhou does not have any top Siyi features.

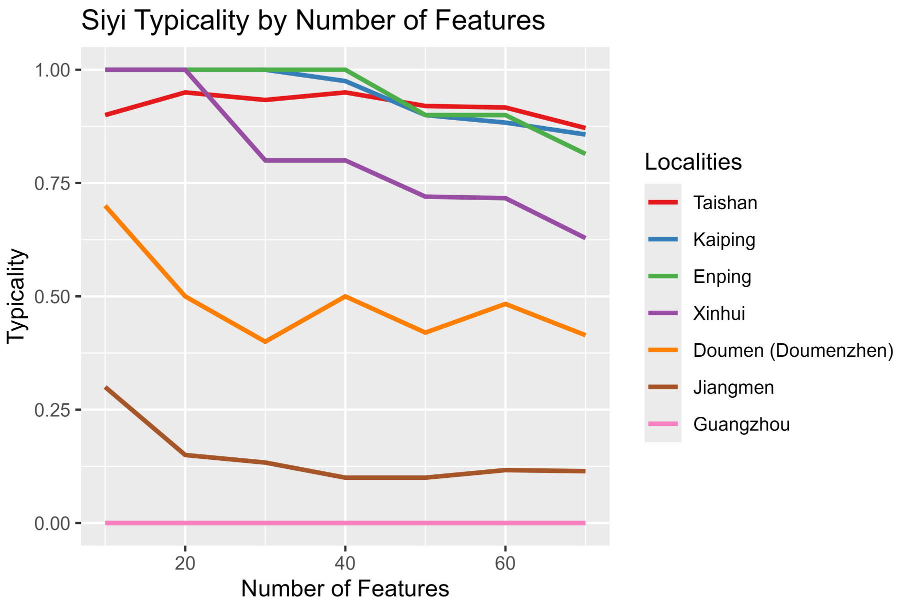


Figure 11: Typicality Decay Plot (based on Siyi features)

Turning to the typicality matrix, out of the top 50 features, 30 of the features Enping and Kaiping have can be classified as *t^h- > h-, nasal > nasal-stop cluster (m-, n- and ŋ-) and *j- > z-. Three features can be defined as the most characteristic features that Siyi dialects have, as reflected in Enping and Kaiping dialects. For the Taishan dialect, the slightly lower typicality score compared to the Enping and Kaping dialects is caused by the absence of a handful of features which are not recurrent (like *-t > -p), but the three features mentioned above are intact. The Xinhui dialect also has all the main Siyi features (the three listed above), which is also reflected in the decay plot at the 20-feature mark. However, unlike Enping, Taishan and Kaiping, Xinhui has fewer non-recurring Siyi features on the list, hence the lower typicality curve in Figure 11.

The Doumen dialect is different from Xinhui (and the most typical Siyi dialects), because out of the top 50 features, it only has *t^h- > h-, nasal > nasal-stop cluster (m-, n- and ŋ-), but not *j- > z-. In addition, there are two exceptions for nasal > nasal-stop cluster, namely in the words 人 'person' and 飲 'to drink'. The absence of *j- > z- might explain why the typicality drops from near 0.75 to below 0.5 after the 20-feature mark in Figure 11.

The Jiangmen dialect has one feature of Siyi, *t^h- > h-, and pretty much none of the other top 50 features, except two minority features. This is reflected in the sustained low typicality after the 20-feature mark.

Lastly, the Guangzhou dialect (together with the other top Guangfu dialects) does not possess any of the top 70 features of Siyi. That is reflected in Figure 11 where the typicality remains at 0 throughout (the other Guangfu dialects are not included in the plot since they are not visible).

4.4 Schematic view of the Guangfu-Siyi transition

Table 5 summarises the characteristic features found in the typical Guangfu dialects (represented by the Guangzhou dialect) and all the Siyi dialects. In this table, 7 recurrent characteristic features discussed in Section 4.2 and 4.3 are included. Since these are the most important features, and they have a relative abundance in the data (can explain a big portion of the features analysed in the previous two sections), this section will focus on this set of features from the two dialect groups only.

<table>
<tr><th>Features</th><th>Guangzhou</th><th>Jiangmen</th><th>Doumen</th><th>Xinhui</th><th>Enping</th><th>Taishan</th><th>Kaiping</th></tr>
<tr><td>*y</td><td>y</td><td colspan="6">i</td></tr>
<tr><td>*t^h-</td><td>t^h-</td><td colspan="6">h-</td></tr>
<tr><td>Pre-velar œ</td><td colspan="2">œ</td><td colspan="5">non-œ</td></tr>
<tr><td>*N</td><td colspan="2">N</td><td colspan="5">NC</td></tr>
<tr><td>*ŋj-</td><td colspan="2">j-</td><td colspan="5">ŋg-</td></tr>
<tr><td>*j-</td><td colspan="3">j-</td><td colspan="4">z-</td></tr>
<tr><td>*s-</td><td colspan="5">s-</td><td colspan="2">ɬ-</td></tr>
</table>

Table 5: Overlap of characteristic features between typical Guangfu dialects and Siyi dialects

Furthermore, the feature values are presented in Table 5, instead of showing presence or absence, as does a typicality table.

The order of the features presented in Table 5 is not by dialect group, but their transitionality. The reflexes of *y (e.g. found in 羽 'feather') mainly include [y] and [i]. The rounded reflex is a Guangfu feature, which is not found in all of the Siyi dialects.[18] The reflexes of *t^h-[19] being [h-] is found in all Siyi dialects, but not Guangfu dialects. Hence, Guangzhou has a retention of *t^h-. Having a pre-velar œ is a strong characteristic of Guangfu dialects. However, Jiangmen also has this feature, with no exceptions. Another strong Guangfu feature which Jiangmen has, *ŋj- > *ɲ- > j-, is also present in Jiangmen. In the other Siyi dialects, there seemed to be no palatals following the *ŋ- in the earlier stage, which led to a ŋg- in present-day dialects. This is caused by the sound change *N (nasal) > NC (nasal-stop sequence), which Jiangmen and Guangfu dialects also did not participate in. It should be noted that a few words in the Doumen dialect undertook *ŋj- > *ɲ- > j- (which could also be due to borrowing). There are exceptions to the sound changes, unlike other dialects, which the exact reason is currently unclear. Next, Siyi dialects except Jiangmen and Doumen undertook *j- > z-. The Doumen dialect also shows the Guangfu variant j- instead of the Siyi z-. Last but not least, only Taishan and Kaiping possess the lateral fricative as a reflex of *s-. Other dialects have the Guangfu variant s-.

The schema in Table 5 presents a shift from co-occuring Guangfu features to Siyi features, with gradual shift of occurrence of features from one dialect group to the other.

5. Discussion

The examination of the Siyi dialects have shed some insights to how dialects can show transitionality. In terms of the type of transitional dialects we see in the data, we tend to see categorical presence or absence of features (different variables). Transitional Siyi dialects are mixed, in the sense that some dialects contains features which are both typical for Guangfu and for Siyi. The more Siyi features a dialect has, the more Siyi a dialect is, vice versa for Guangfu. This mirrors Berruto's (2010) characterisation of 'core' dialects (and the implied transitional dialects).

What is also interesting is that there are a few instances of e.g. *ŋj- > *ɲ- > j- in the Doumen dialect, which could be due to lexical diffusion of the sound change or borrowing from Guangfu. These handful of tokens are showing a mixed dialect pattern, but other than those in this one dialect, everything else is still very categorical. The categorical presence and absence of different features in the transitional dialects shows that dialects do not necessarily have to show lexical diffusion-like patterns in order to be transitional. Sometimes, the number of typical features can also define how typical a dialect is. If a dialect does not have all the (top) typical features from one dialect group,

[18] However, it should be noted that *y > i is not exclusive to Siyi dialects, since Coastal dialects also had this sound change (Sung and Prokić accepted).

[19] It also implies the reflex of *d- here, since in Siyi dialects, *d- became *t^h- in certain contexts, and these *t^h- also became [h-].

or it has the most typical features from two (or more) dialect groups, then it can be considered transitional too. Since they are categorical, these features were probably not considered or missed in the single-variant analyses. The aggregate analysis followed by a dialect typicality analysis allow a multivariate view of dialect transition and simultaneously, allow us to inspect and not missing the features which are categorical in these dialects in order to evaluate their transitional status.

The second research question concerns whether transitional dialects would always contain features that are the most typical for one dialect group, and as we get further away from the core area, we start losing less-typical features first. It turns out, this is not necessarily the case. When we look at the Siyi typicality decay plot in Figure 11, we can see that as we move away from the core Siyi area (where Taishan, Enping, Kaiping is), the typicality in general drops for the other dialects (the gradual drop of typicality shows the transitionality of Siyi dialects). However, from the Guangfu typicality decay plot in Figure 10, we can see that Doumen, Xinhui and Enping have rather low typicality to start with. But then, at the 30-feature mark, the typicality increased. This is because at the 30-feature mark, the variant [s-] is then considered. [s-] is not the most typical feature in Guangfu dialects, but these three Siyi dialects have adopted it, but not the other two more typical features. The motivation behind why a less-typical feature is adapted to the Siyi dialects is unclear at the present. Perhaps perception is involved with the adoption of dialect features, but perceptual data is not available in the dialect survey data, meaning the question awaits for further research.

Another interesting discovery during the typicality analysis is that multiple top features that have been identified as highly exclusive to Guangfu and Siyi dialects is that many of them are shared innovations. While the automatic detection of shared innovations is highly useful for historical linguistics, it also raises issues in linguistic classification. Historical linguists speak of *subgrouping*, when classifying varieties which are related with each other. Historical linguists often criticise dialectologists for using both innovations and retentions as criteria for dialect classification. In historical linguistics, the historical relationship between varieties is determined by the notion of *shared innovation*, i.e. a change that occurred to the parent of a subset of daughter varieties before they diversified (Campbell 2013, 175). The current study has shown that there might be some pitfalls if we only look at shared innovations alone. First of all, as we have seen in the current study, a dialect can possess shared innovations from two dialect groups (e.g. the Jiangmen dialect). How should Jiangmen be subgrouped then? Secondly, as we have seen in the typicality decay plot, the typicality of a dialect changes depends on how many feature we consider. The more features we consider, the more we can see their underlying structure and their belonging to different dialect groups. I would argue that the aggregate analysis would yield a better picture for the relationship between dialect varieties, although that does not automatically imply phylogenetic relationships of the dialects. Dialect typicality analysis is able to identify key dialect features which have been adopted, perhaps as a result of contact, and that could add a powerful explanation to the MDS plot we see, which is used to show (synchronic) relationships that is not based on a handful of features only.

Last but not least, it should be noted that the current study is based on 130 lexical items only. There is a possibility that the items chosen happen to show a categorical presence or absence of features, and this might change with the inclusion of more words. Moreover, there are not too many Siyi dialects, opposed to Guangfu dialects in the data (due to the lack of surveyed sites). To further understand Siyi dialects (and dialect transitions in general), more dialects should be sampled. Further investigations should be done by looking into more lexical items than the current dataset, and perhaps with more dialects in the Siyi-Guangfu transition area.

6. Conclusion

This study has explored transitional dialects under the scope of an aggregate analysis. In addition, a dialect typicality decay analysis is performed on top of automatic dialect feature extraction, which allows us to look at the most typical features from the Guangfu and Siyi dialects of Yue in detail, regarding to how dialect transitions work. Previous methods, like Factor Analysis, were able to extract the proportion of different dialect contents from the data, whereas dialect typicality decay analysis provides more to the analysis of transitionality in terms of identifying the actual features.

Throughout the analysis, it has been found that most dialects do not show a mixed dialect pattern (for one single variable), but rather, the gradual difference in the categorical presence or absence of multiple features that are typical to each dialect group. Additionally, it has been discovered that dialects do not necessarily need to adopt the most typical feature of one dialect group first before adopting something less typical. Despite more research is necessary to understanding the motivation of the adoption of certain features, which could be linked to perception, the current study has illustrated how an aggregate analysis and dialect typicality analysis can help us explore transitional dialects on a multivariate level.

Acknowledgements

I would like to thank both anonymous reviewers for their feedback, especially Reviewer 1's extensive comments. I would also like to thank Dr. Jelena Prokić for her comments on the typicality decay plots.

Supplementary material

The data was digitised by Sung et al. (2024). It can be found in the OSF repository here: `https://osf.io/m9g2a/`.

The script for nPMI feature extraction (from Sung and Prokić 2024a) can be found in the Github repository here: `https://github.com/dialmatt123/feature_extraction_nPMI`.

Lastly, the scripts for the dialect typicality analysis described in the current article are available in: `https://github.com/dialmatt123/dialect_typicality_decay_analysis`.

References

Bach, A. (1950), *Deutsche Mundartforschung : ihre Wege, Ergebnisse und Aufgaben*, Germanische Bibliothek. 3. Reihe, Untersuchungen und Einzeldarstellungen, 2. aufl. ed., Winter, Heidelberg.

Berruto, G. (2010), *13. Identifing dimensions of linguistic variation in a language space*, De Gruyter Mouton, Berlin, New York, pp. 226–241. https://doi.org/10.1515/9783110220278.226.

Borg, I. and P. J. F. Groenen (2005), *Modern multidimensional scaling: Theory and applications*, Springer Science and Business Media.

Bouma, G. (2009), Normalized (pointwise) mutual information in collocation extraction, *Proceedings of the Biennial GSCL Conference*, Vol. 30, pp. 31–40.

Campbell, L. (2013), *Historical linguistics*, Edinburgh University Press.

Chambers, J. K. and P. Trudgill (1998), *Dialectology*, 2nd ed., Cambridge University Press.

Chen, H. and Y. Lin (2009), *粵語平話土話方音字彙第 1 編: 廣西粵語、桂南平話部分 [Yue, Pinghua and Tuhua Dialect Survey Collection Part 1]*, Shanghai Educational Publishing House.

Chen, X. (2009), 廣西賀州八步 (桂岭) 本地話音系 [the phonology of the hezhou babu (guiling) dialect in guangxi], *方言 [Dialect]* (1), pp. 53–71.

Chen, X. and Z. Weng (2010), *粵語西翼考察–廣西貴港粵語個案研究 [Investigating Western Yue - A case study on Guigang Yue in Guangxi]*, Jinan University Press.

Chinese Academy of Social Sciences (CASS) (2012), 中國語言地圖集 *[Language Atlas of China]*, 2 ed., Commercial Press, Beijing.

Church, K. and P. Hanks (1990), Word association norms, mutual information, and lexicography, *Computational linguistics* **16** (1), pp. 22–29.

Embleton, S. (1993), Multidimensional scaling as a dialectometrical technique: Outline of a research project, *Contributions to Quantitative Linguistics: Proceedings of the First International Conference on Quantitative Linguistics, QUALICO, Trier, 1991*, Springer Netherlands, pp. 267–276.

Gan, Y.-E. (2003), 四邑話: 一種粵化的混合方言 [siyi vernacular: A yue-ised mixed dialect], 中國社會語言學 *[Chinese Sociolinguistic]*.

Girard, D. and D. Larmouth (1993), Some applications of mathematical and statistical models in dialect geography, *American dialect research* pp. 107–132, John Benjamins Philadelphia.

Goebl, H. (1984), *Dialektometrische Studien: Anhand Italoromanischer, Rätoromanischer und Galloromanischer Sprachmaterialien aus AIS und ALF*, Vol. 3, Niemeyer, Tübingen.

Goebl, H. (2018), Dialectometry, *in* Boberg, C., J. Nerbonne, and D. Watt, editors, *The Handbook of Dialectology.*

Heeringa, W. (2004), *Measuring dialect pronunciation using Levenshtein distance*, PhD thesis, University of Groningen.

Heeringa, W. and J. Nerbonne (2001), Dialect areas and dialect continua, *Language Variation and Change* **13** (3), pp. 375–400.

Hinskens, F. L. M. P. and G. R. Guy (2016), Linguistic coherence: Systems, repertoires and speech communities, *Lingua* **172**, pp. 1–9, Elsevier BV.

Huang, Q. (2006), 賀州市賀街本地話同音字匯 [homonymic syllabary of the hezhou hezhoujie local vernacular], *Journal of Guilin Normal College* **20** (3), pp. 6–13.

Hubert, L. and P. Arabie (1985), Comparing partitions, *Journal of classification* **2** (1), pp. 193–218.

Kristiansen, G. (2008), Style-shifting and shifting styles: A socio-cognitive approach to lectal variation, *Cognitive sociolinguistics: Language variation, cultural models, social systems*, Mouton de Gruyter Berlin/New York.

Leinonen, T., Ç. Çöltekin, and J. Nerbonne (2016), Using gabmap, *Lingua* **178**, pp. 71–83.

List, J.-M., R. Forkel, S. Greenhill, T. Tresoldi, C. Rzymski, G. Kaiping, S. Moran, P. Bouda, J. Dellert, T. Rama, and F. Nagel (2021), *LingPy. A Python library for historical linguistics*, Max Planck Institute for Evolutionary Anthropology, Leipzig. https://lingpy.org.

Liu, C. (2015), 廣東兩陽粵語語音研究 *[Research in the Phonetics of Yue in the Guangdong Liangyang Area]*, PhD thesis, Jinan University.

Macaulay, R. K.S. (2010), Dialect, *Variation and Change: Pragmatic Perspectives*, John Benjamins Publishing Company, pp. 61–72.

Manning, C. and H. Schütze (1999), *Foundations of Statistical Natural Language Processing*, MIT Press.

Meyer, P. (1877), Archives des missions scientifiques et littéraires. 3e série, t. iii, 2e livraison, *Romania* **6** (24), pp. 630–633.

Nerbonne, J. (2010), Mapping aggregate variation, *An International Handbook of Linguistic Variation*, Vol. 2, pp. 476–495.

Nerbonne, J., R. Colen, C. Gooskens, P. Kleiweg, and T. Leinonen (2011), Gabmap-a web application for dialectology, *Dialectologia: revista electrònica* pp. 65–89.

Paris, G. (1888), Les parlers de france, *Revue des patois gallo-romans* **2**, pp. 161–175.

Pickl, S. (2013), *Probabilistische Geolinguistik : geostatistische Analysen lexikalischer Variation in Bayerisch-Schwaben*, Zeitschrift für Dialektologie und Linguistik. Beihefte ; Bd. 154, Franz Steiner Verlag, Stuttgart.

Pickl, S. (2016), Fuzzy dialect areas and prototype theory: Discovering latent patterns in geolinguistic variation, *The future of dialects* pp. 75–98.

Preston, D. R. (1988), Methods in the study of dialect perceptions, *Methods in Dialectology. Clevedon: Multilingual Matters* pp. 373–395.

Prokić, J., Ç. Çöltekin, and J. Nerbonne (2012), Detecting shibboleths, *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pp. 72–80.

Prokić, J. and J. Nerbonne (2008), Recognising groups among dialects, *International journal of humanities and arts computing* **2** (1-2), pp. 153–172.

Pröll, S. (2015), Raumvariation zwischen muster und zufall, BiblioScout.

Rubehn, A., S. Montemagni, and J. Nerbonne (2024), Extracting tuscan phonetic correspondences from dialect pronunciations automatically, *Language Dynamics and Change* **14** (1), pp. 1–33, Brill.

Séguy, J. (1971), La relation entre la distance spatiale et la distance lexicale, *Rev. Ling. Romane* **35** (139-140), pp. 335–357.

Séguy, J. (1973), *Atlas linguistique et ethnographique de la Gascogne*, Vol. 6, Centre National de la Recherche Scientifique, Paris. [ALG].

Shao, H. (2016), 粵西湛茂地區粵語語音研究 *[The Phonological Study of the Yue Dialects spoken in the Zhan-Mao area in Western Guangdong]*, Sun Yat-Sen University Press.

Shao, H.-J. and Y.-E. Gan (1999), 廣東四邑方言語音特點 [the phonetic characteristics of guangdong siyi dialects], 方言 *[Dialect]*.

Shi, R. (2009), 廣西防城區粵語音系 [the phonology of the fangcheng yue dialect in guangxi], 百色學院學報 *[Journal of Baise University]* **22** (2), pp. 106–116.

Simmons, R. V. (2012), *Dialect Transition Zones in Southern Jiangsu - A Close Range Examination*, pp. 253–294.

Sung, H. W. M. (forthcoming), How can digital and computational methods benefit the study of yue dialects, *in* Lau, C. W. Y., S.-K. Cheng, and Y.-P. Lai, editors, *Cantonese and Digital Humanities*, Linguistic Society of Hong Kong.

Sung, H. W. M. and J. Prokic (2023), What are guangfu dialects?, *27th International Conference on Yue Dialects*, Ohio State University, Online Presentation. https://u.osu.edu/yue2023/.

Sung, H. W. M. and J. Prokić (2024a), Detecting dialect features using normalised pointwise information, *Computational Linguistics in the Netherlands Journal* **13**, pp. 121–145. https://www.clinjournal.org/clinj/article/view/177.

Sung, H. W. M. and J. Prokić (2024b), Identification of dialect typicality and kernels, *12th International Conference on Language Variation in Europe (ICLaVE|12)*, University of Vienna, Oral Presentation. https://pretalx.dioe.at/iclave12/talk/review/3SPCKT8TUCFMH3LYP3LZ7KPRTAHJABRK.

Sung, H. W. M. and J. Prokić (accepted), Relative chronology of dialectal phonetic features, *in* Wandl, F., T. Olander, and J.-M. List, editors, *Relative chronology in historical linguistics*, Language Science Press.

Sung, H. W. M., J. Prokic, and Y. Chen (2024), A new dataset for tonal and segmental dialectometry from the Yue- and pinghua-speaking area, *in* Hahn, M., A. Sorokin, R. Kumar, A. Shcherbakov, Y. Otmakhova, J. Yang, O. Serikov, P. Rani, E. M. Ponti, S. Muradoğlu, R. Gao, R. Cotterell, and E. Vylomova, editors, *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*.

Taeldeman, J. (1989), *A typology of dialect transitions in Flanders*, De Gruyter Mouton, Berlin, Boston, pp. 155–164. https://doi.org/10.1515/9783110883459-015.

Taeldeman, J. and F. L. M. P. Hinskens (2013), The classification of the dialects of dutch, *Language and Space. An International Handbook of Linguistic Variation. Volume 3: Dutch*, De Gruyter Mouton, pp. 129–142.

Tan, Y. (2017), 廣西賓陽縣 (賓州鎮) 本地話音系 [the phonology of binzhouzhen in the binyang county in guangxi], 梧州學院學報 *[Journal of Wuzhou University]* **27** (5), pp. 58–71.

Ward, Jr., J. H. (1963), Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association* **58**, pp. 236–244.

Wieling, M. and J. Nerbonne (2010), Hierarchical spectral partitioning of bipartite graphs to cluster dialects and identify distinguishing features, *Proceedings of TextGraphs-5-2010 Workshop on Graph-based Methods for Natural Language Processing*, pp. 33–41.

Wieling, M. and J. Nerbonne (2011), Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features, *Computer Speech & Language* **25** (3), pp. 700–715, Elsevier.

Wieling, M., E. Margaretha, and J. Nerbonne (2011), Inducing phonetic distances from dialect variation, *Computational Linguistics in the Netherlands Journal* **1**, pp. 109–118. https://clinjournal.org/clinj/article/view/10.

Wiesinger, P. (1983), Die einteilung der deutschen dialekte, *in* Besch, Werner, editor, *Dialektologie: Ein Handbuch zur deutschen und allgemeinen Dialektforschung*, Vol. 1.2 of *Handbücher zur Sprach- und Kommunikationswissenschaft*, Berlin/New York: de Gruyter, Berlin, New York, pp. 807–900. http://www.degruyter.com/view/books/9783110203332/9783110203332.807/9783110203332.807.xml.

Xie, J. (2007), *Studies on the Chinese dialects in Guangxi*, Guangxi People’s Publishing. In Chinese.

Yang, S. (2013), 廣西藤縣濛江方言音系 [the phonology of the tengxian mengjiang dialect in guangxi], 方言 *[Dialect]* (1), pp. 71–85.

Zhan, B. and Y.-S. Cheung (1987), *A Survey of Dialects in the Pearl River Delta, Vol. 1, Comparative Morpheme-Syllabary*, People’s Publishing House of Guangdong.

Zhan, B. and Y.-S. Cheung (1990), *A Survey of Dialects in the Pearl River Delta, Vol. 3, A Synthetic View*, People’s Publishing House of Guangdong.

Zhan, B. and Y.-S. Cheung (1994), *A Survey of Yue Dialects in North Guangdong*, Jinan University Press.

Zhan, B. and Y.-S. Cheung (1998), *A Survey of Yue Dialects in West Guangdong*, Jinan University Press.

Zhao, Y. and G. Karypis (2001), Criterion functions for document clustering: Experiments and analysis, *Technical Report 01-40*, Department of Computer Science, University of Minnesota, Minneapolis, MN.

Zhong, Z. (2015), 廣西蒼梧本地話音系 [the phonology of cangwu local vernacular in guangxi], 方言 *[Dialect]* (2), pp. 177–192.

Appendix 1: Top 70 Features in Guangfu Yue Dialects

Feature	Variant	nPMI	Exclusivity	Representativeness
œ_leg	œ	0.642	0.795	0.816
*ɲ_meat	j	0.626	0.75	0.868
*ɲ_moon	j	0.621	0.775	0.816
œ_bird_col	œ	0.619	0.789	0.789
œ_long	œ	0.615	0.775	0.816
œ_to think	œ	0.615	0.775	0.816
*ɲ_sun	j	0.603	0.744	0.842
*ɲ_person	j	0.584	0.727	0.842
*ɲ_fish	j	0.559	0.732	0.789
y_village	y	0.542	0.705	0.816
*ɲ_hot	j	0.515	0.718	0.737
*ɲ_two	j	0.513	0.73	0.711
y_all	y	0.488	0.66	0.816
y_pig	y	0.471	0.618	0.895
*dz_all	ts^h	0.468	0.6	0.947
y_feather	y	0.468	0.6	0.947
*ɲ_ear	j	0.452	0.684	0.684
ɔ_to speak_col	ɔ	0.436	0.581	0.947
y_tree	y	0.427	0.576	0.919
*d_to sit	ts^h	0.424	0.583	0.921
ɔ_horn	ɔ	0.424	0.569	0.974
y_rain	y	0.414	0.574	0.921
y_moon	y	0.41	0.593	0.842
œ_double	œ	0.401	0.649	0.649
y_fish	y	0.4	0.565	0.921
y_boat	y	0.395	0.582	0.842
leg_m	0	0.395	0.582	0.842
y_mouse/rat	y	0.394	0.565	0.921
ui_to sleep	ɵy	0.393	0.762	0.421
*d_head	t^h	0.392	0.567	0.895
ui_water	ɵy	0.389	0.762	0.421
s_four	s	0.389	0.609	0.737

Feature	Variant	nPMI	Exclusivity	Representativeness
ui_water	ɵy	0.389	0.762	0.421
s_four	s	0.389	0.609	0.737
s_heart	s	0.389	0.609	0.737
s_new	s	0.389	0.609	0.737
s_to die	s	0.389	0.609	0.737
s_to think	s	0.389	0.609	0.737
s_west	s	0.389	0.609	0.737
ɵ_spring	ɵ	0.384	0.812	0.342
y_round	y	0.384	0.585	0.816
bird_col_m	0	0.379	0.571	0.842
*ɲ_enter	j	0.379	0.571	0.842
y_blood	y	0.373	0.574	0.816
s_small_lit	s	0.372	0.596	0.737
s_three	s	0.372	0.596	0.737
*d_abdomen	tʰ	0.369	0.566	0.811
s_small_col	s	0.368	0.596	0.737
long_m	0	0.365	0.561	0.842
to think_m	0	0.365	0.561	0.842
tsʰ_long	ʦʰ	0.363	0.537	0.947
*ɲ_to drink	j	0.35	0.552	0.842
*ı_teacher	y	0.344	0.889	0.211
*b_skin1	pʰ	0.343	0.521	0.974
ɵ_out	ɵ	0.338	0.786	0.289
ɔ_to sit	ɔ	0.336	0.541	0.868
*ŋ_cow	ŋ	0.328	0.522	0.921
stone_n	ɛ	0.327	0.611	0.579
s_knee	s	0.325	0.625	0.526
*g_strange	kʰ	0.317	0.507	0.974
*ts_bird_col	ʦ	0.31	0.568	0.658
ɔ_I	ɔ / ɔi	0.309	1	0.105
*-t_seven	k	0.309	1	0.105
*-n_to see_col	ŋ	0.309	1	0.105
ɛ_snake	ɛ	0.301	0.507	0.921
h_to see_lit	f	0.293	1	0.083
*u_soil_earth	ou	0.289	0.588	0.526
*-n_all	ŋ	0.284	1	0.079
*-t_bone	k	0.284	1	0.079
a_eight	a / ɛ	0.284	1	0.079
*-t_lice	k	0.284	1	0.079
*-n_smoke	ŋ	0.284	1	0.079

Appendix 2: Top 70 Features in Siyi Yue Dialects

Feature	Variant	nPMI	Exclusivity	Representativeness
*d_head	h	0.946	0.857	1
*t^{h}_soil/earth	h	0.946	0.857	1
*-t_knee	p	0.94	1	0.833
*ɲ_meat	ŋg	0.94	1	0.833
*ɲ_sun	ŋg	0.94	1	0.833
*ɲ_hot	ŋg	0.88	0.833	0.833
*ɲ_moon	ŋg	0.88	0.833	0.833
night_m	i	0.875	1	0.667
*ɲ_person	ŋg	0.875	1	0.667
*ɲ_to drink	ŋg	0.875	1	0.667
*ɲ_ear	ŋg	0.829	0.714	0.833
*ɲ_fish	ŋg	0.829	0.714	0.833
*ɲ_two	ŋg	0.829	0.714	0.833
*ɲ_enter	z	0.807	0.8	0.667
j_feather	z	0.807	0.8	0.667
j_leaf	z	0.807	0.8	0.667
j_night	z	0.807	0.8	0.667
j_one	z	0.807	0.8	0.667
j_rain	z	0.807	0.8	0.667
j_round	z	0.807	0.8	0.667
j_smoke	z	0.807	0.8	0.667
j_to swim	z	0.807	0.8	0.667
ash_m	u	0.804	1	0.5
ɐu_autumn	ɛu	0.804	1	0.5
ɐu_hand	iu	0.804	1	0.5
ɐ_ten	i	0.804	1	0.5
ɐu_to swim	iu	0.804	1	0.5
ɐ_tomb	u	0.803	1	0.5
*ŋ_cow	ŋg	0.785	0.625	0.833
*ŋ_eye	ŋg	0.785	0.625	0.833
*m_hair_body	mb	0.785	0.625	0.833

Feature	Variant	nPMI	Exclusivity	Representativeness
*m_horse	mb	0.785	0.625	0.833
*ŋ_I	ŋg	0.785	0.625	0.833
*n_man	nd	0.785	0.625	0.833
*ŋ_to bite	ŋg	0.785	0.625	0.833
*ŋ_tooth1	ŋg	0.785	0.625	0.833
many_m	u	0.751	0.667	0.667
*n_woman	nd	0.746	0.556	0.833
bone_n	u	0.723	0.75	0.5
ɐ_enter	i	0.723	0.75	0.5
full_m	u	0.723	0.75	0.5
ɵ_out	u	0.723	0.75	0.5
boat_m	u	0.722	1	0.333
*ŋ_five	ŋ̍ / m̩	0.722	1	0.333
to drink_m	i	0.722	1	0.333
*d_floor/ground	0	0.721	1	0.333
i_leaf	a	0.721	1	0.333
one_m	i	0.721	1	0.333
round_m	u	0.721	1	0.333
*m_full	mb	0.711	0.5	0.833
*m_tail	mb	0.711	0.5	0.833
*n_you	nd	0.711	0.5	0.833
*ɣ_lake	v	0.703	0.429	1
to sit_m	u	0.703	0.571	0.667
ɛ_snake	a	0.68	0.455	0.833
ɐ_cloud	u	0.661	0.6	0.5
I_m	u	0.627	0.444	0.667
*n_bird_lit	nd	0.625	0.444	0.667
fire_m	u	0.625	0.385	0.833
*d_big	0	0.619	0.667	0.333
*t_cold	0	0.619	0.667	0.333
u_full	a	0.619	0.667	0.333
ɐ_lice	ɛ	0.619	0.667	0.333
*t_many	0	0.619	0.667	0.333
*t_winter	0	0.619	0.667	0.333
*d_abdomen	0	0.618	0.667	0.333
ɐu_nine	ɛu	0.618	0.667	0.333
ɐ_one	i	0.618	0.667	0.333
y_all	ɛ	0.614	1	0.167
*ts_bird_col	s	0.614	1	0.167