Advancing explanatory and tonal dialectometry

Sung, H.W.M.

CHAPTER 6

Automatic Feature Detection for Yue Dialects[1]

6.1 Introduction

Feature extraction (in dialectometry) refers to the identification of important features which differentiate one dialect group from another. It is an important step to understanding the dialectal variation, a step which has traditionally been done manually. However, manual extraction of important features is susceptible to the following problems: it is a time-consuming task; there is a risk of overlooking certain features and lastly, every analyst can come up with a different set of features. This chapter proposes a solution to these problems, namely with a novel application of Normalised Pointwise Mutual Information on the Yue dataset, in order to understand the dialect groups detected in the Yue-speaking area further.

Traditionally, dialectologists rely on plotting carefully selected sets of dialect features on physical maps, drawing isoglosses and identifying dialect areas. As the previous chapter illustrates, dialectometry helps us to reduce biases in dialect classification. The dialectometric results, however, still requires the analysts' own interpretation and expertise in order to understand the underlying linguistic factors responsible for clusters

[1]This chapter (excluding the results) is based on Sung and Prokić (2024a).

automatically detected by the algorithms. Clustering and multidimensional scaling rely on distance matrices, which do not offer any details or explanations of the identified dialect partitions. During the conversion from the qualitative dialect data to dialect distances, the information on linguistic features is lost.

To overcome the problem, several approaches have been proposed to extract characteristic features in a dialect classification. One of these approaches is Prokić et al.'s (2012) method, which is based on Fisher's Linear Discriminant. It tries to seek variables which have the biggest homogeneity within the cluster, and are simultaneously distinctive from dialects outside the cluster. Another approach proposed by Pickl (2016) relies on a dimensionality reduction technique which proceeds from features rather than distance matrices, namely Factor Analysis. Before Sung and Prokić (2024a), there has been no systematic comparison of various feature extraction methods. Sung and Prokić (2024a) compared the two methods above, as well as proposing a third approach, the application of Normalised Pointwise Mutual Information (nPMI), and found that nPMI can extract the most exclusive features out of the three approaches. They argued that nPMI is the most suitable method for the task of automatic dialect feature extraction.

This chapter is structured as follows. In Section 6.2, a literature review of the previous feature extraction methods is presented. Section 6.3 contains the description of the methodology for using nPMI as a feature extraction technique. The extracted top features for each dialect group are presented in Section 6.4, and finally, a discussion and conclusion can be found in Section 6.5.

6.2 Previous approaches

Previous approaches in dialectometry which attempt to identify features characteristic for dialect groups can be divided into two categories: bottom-up approaches (e.g. Pickl 2016) and top-down (e.g. Prokić et al. 2012). Bottom-up approaches seek simultaneously the dialect groups and distinctive features, whereas top-down approaches require a pre-defined dialect classification before features can be extracted.

6.2.1 Bottom-up approaches

One of the more common bottom-down approaches in dialectometry is Factor Analysis (FA), which has been used in exploring dialect areas and their characteristic features (Pickl 2013; Pröll 2015; Pickl 2016). FA is a dimensionality reduction technique, like Multidimensional Scaling (MDS, Borg and Groenen 2005, see also Chapter 3), used in dialectometry (Embleton 1993; Heeringa 2004) to identify dialect groups. It condenses the variation of the categories in the data into a smaller number of patterns, or underlying factors, by grouping variants that "co-occur with a high frequency"(Pickl 2016:82). Within the dialectal context, FA detects (gradient) membership of dialect areas (factors) and at the same time finds out which features contribute to the make-up of these groups and their respective strength of association to the group. The features can be lexical, phonetic, morphological or combined, as long as they are categorical (see Pröll 2015).

Unlike top-down approaches, FA does not require predefined groups. Pickl (2016) has argued that dialect areas are 'fuzzy', and FA can capture this fuzziness by identifying condensations of co-occurring variants instead of hard clusters. The *Factor Loading* is one of the by-products when using FA, which indicates the relationship between the Factor (dialect group) and the location.[2] Unlike traditional assumptions or representations of dialect areas, where dialects are rather homogenous within the group, each location has a different degree of factor loading for each dialect group. Furthermore, Pickl (2016) illustrates the use of *Combined Factor Maps*, which represents the dominant factors, or the strongest dialect group that each locality is associated with, yielding the highest concentrations of each dialect area (the 'surface dialect landscape'), like the map below in Figure 6.1.

Another by-product of FA are *Factor Scores*. The dialect features most associated with a particular factor can be extracted based on the factor scores of the variant. The higher the factor score, the more it is associated with the respective factor.

[2]According to Pickl (2016), FA requires the elements to be more than the features, which is often not the case for dialectometry. Therefore, the transpose of the data matrix is used (Q-type FA). This is the version implemented in the GeoLing software. Hence, Factor Loading is associated with the locations, and Factor Score is associated with the features, not the other way round as one would expect.

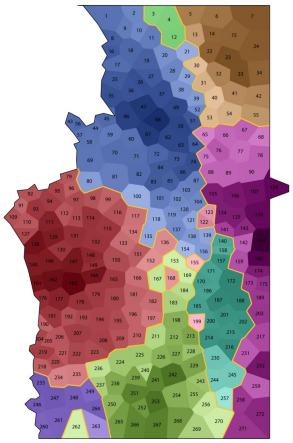


Figure 6.1: Combined Factor Map of the Sprachatlas von Bayerisch-Schwaben (SBS) survey sites (from Pickl 2016)

The earliest usage of FA to extract features can be found in Inoue and Kasai (1982b). Inoue and Kasai looked at 82 lexical variables from the *Linguistic Atlas of Japan* and calculated the percentage of Standard forms present in each prefecture. Inoue and Kasai used FA on this so-called 'Kasai dataset' and found 4 main factors as well as features associated with each factor.

Other uses of FA include Nerbonne's (2006) feature identification of American English dialects recorded in the *Linguistic Atlas of the Middle and South Atlantic States (LAMSAS)*, which uses vowel features and Grieve's (2014) analysis of American English vowels using vowel formants.[3]

A similar approach to FA is Principal Component Analysis (PCA). PCA is a set of mathematical procedures that seeks and groups sets of variables that strongly (positively or negatively) correlate to each other (Shackleton 2005:141). The analysis returns 'principal components', axes which group variables on the two poles, one being large positive values

[3]The vowel formants were pre-processed with a conversion into the Getis-Ord Gi z-score, an index for local spatial autocorrelation (Ord and Getis 1995). Local spatial autocorrelation calculates the degree a location is part of a high or low (formant 1 or formant 2) value cluster. The use of the Getis-Ord Gi z-score acts as a smoothing technique so that "the values of the smoothed variables only represent the underlying regional signals in the raw values of these variables" (Grieve 2014:74).

and the other being large negative values. The first principal component accounts for the most variance from the dataset, and the second principal component accounts for less variance than the first principal component. For each principal component, PCA could also reveal clusters of dialects and isolate sets of features that tend to co-occur. Shackleton (2005) has illustrated the use of PCA for identifying dialect features between British English and American English speakers, and Leinonen (2010) has applied PCA on Swedish Bark filtered vowel spectra in order to investigate dialect levelling.

Lastly, the use of bipartite spectral graph partitioning by Wieling and Nerbonne (2011) can determine dialect groups and their sound correspondences simultaneously. The result returns clusters which include both the dialect groups and their respective sound correspondences together. The importance of the sound correspondences is then calculated post-hoc by taking the average of the Representativeness and Distinctiveness indices given in Wieling and Nerbonne (2011:707).

6.2.2 Top-down approaches

There have been far fewer works done with the top-down approach in dialectometry. Prokić et al.'s (2012) method is a representation of this approach. Prokić et al. (2012) seeks features that differ little within a pre-defined group but differ enormously outside the group. This method was inspired by Fisher's Linear Discriminant (FLD), and since the authors did not give a name for this approach, this method is addressed as FLD throughout the chapter. The way it is done is by calculating the mean distance of a particular word or feature among all the pairs of dialects within the pre-defined group and outside the group with Levenshtein distance (Levenshtein 1966; Heeringa 2004). The pre-defined groups can be obtained by using cluster analysis (Prokić and Nerbonne 2008). Next, characteristic features are identified by seeking features with the largest differences between the within-group and outside-group differences.

The calculation of the within-group and between-group distance for one locality is illustrated in Figure 6.2. S represents a locality within the pre-defined group and the arrows represent the distances measured, which includes the pre-defined group (in blue) and the rest of the dialects outside the group (in yellow). This procedure is iterated for all the sites.

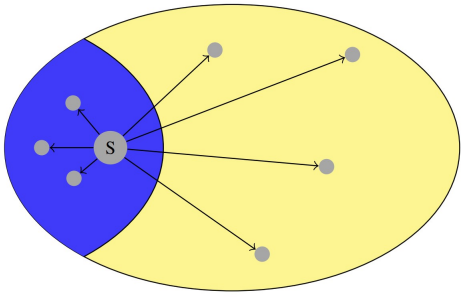


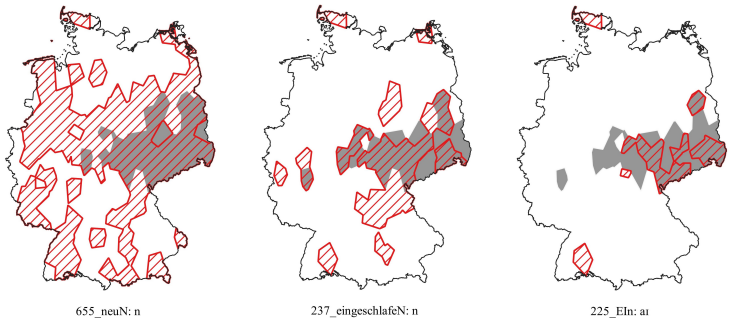
Figure 6.2: Illustration of distance calculation for the FLD method (from Prokić et al. 2012)

FLD has previously been tested on Dutch and German (Prokić et al. 2012). Both analyses have identified features which are rather homogenous within the pre-defined group. The authors have also found that the same word can show up as distinctive for more than one pre-defined group (with different variants). Lastly, this method is applicable to any feature type which can be defined with a numerical distance metric between elements. This includes words (as analyzed in Prokić et al. 2012), categorical data or vowel formants.

6.2.3 Comparison of feature extraction methods

Sung and Prokić (2024a) have compared three feature extraction methods, namely Factor analysis, Fisher's Linear Discriminant, and Normalised Pointwise Mutual Information. Using data from the *Phonetischer Atlas der Bundesrepublik Deutschland* (Göschel 1992), they have evaluated the top features extracted using each of the methods by how exclusive and how representative each feature is for each dialect group. They have found that nPMI is able to find the most exclusive (yet, not too localised) top features out of all three methods. This is most obvious when working with the Upper Saxon area. In Figure 6.3 below, the distribution map of the most important (Upper Saxon) feature extracted by the each of the 3 methods are shown. The shaded area is a dialect area we are interested in (in this case, Upper Saxon), and the areas in red are the dialects which the extracted feature is found. We can see that nPMI can identify a feature that is largely exclusive (though not

as representative). On the other hand, FA fails to extract an exclusive feature for the Upper Saxon area. When looking at more top features, nPMI remains being the method which can extract features with the highest average exclusivity score.



(a) FA **(b)** FLD **(c)** nPMI

Figure 6.3: Distribution map of the first feature extracted by various methods (from Sung and Prokić 2024a). The shaded area represents the Upper Saxon area, and the red area indicates the distribution of the most important feature of Upper Saxon extracted using the respective method.

Sung and Prokić (2024a) have noticed that when a decision has to be made between Exclusivity and Representativeness, nPMI seeks exclusive features at the expense of Representativeness, like in the case for Upper Saxon. This property of nPMI is something which becomes very useful in dialect feature extraction. This is because "a high representativeness of dialects with the identified feature found in a dialect group does not imply the feature is exclusive to the area" (Sung and Prokić 2024a:132), as shown in Figure 6.3a.

6.3 Methodology

In order to identify features which are characteristic for each dialect group, this chapter takes the data-driven approach from Sung and Prokić (2024a), which requires the following two steps: 1) *dialect classification* and 2) *feature extraction*. Dialect classification involves automatic detection of dialect groups by the means of unsupervised machine-learning

methods and feature extraction uses an association metric to seek features which are most associated with a particular dialect group. The following subsections will explain the steps above in more detail.

6.3.1 Classification of Yue Dialects

The classification of the Yue-Pinghua dialects consists of several steps: i) multiple sequence alignment, ii) distance calculation, iii) cluster analysis.

Multiple Sequence Alignment

There are numerous ways to calculate dialect distances. One of the factors in determining the method is data type. For instance, in the previous chapter, Levenshtein distance was used to calculate dialect distances for transcription data (in IPA). However, since the algorithm often only counts the number of operations to transform the transcription from one dialect to another (see Chapter 3), the features (or segments) that were transformed are often not recorded in the process. This means that while the dialect distances are represented by a number, the features that are not shared by the dialects in comparison (the operations) are lost during the process of distance calculation.

There are numerous ways to retrieve dialect features. However, unless the input data are categorical data (like dialect features typically used in Goebl's (1984) approach, multi-aligned segmental data (Sung and Prokić 2024a; Sung 2025) or numerical data (such as vowel formant data analysed with Factor Analysis (Grieve 2014)), the outputs of these methods will not return features which only consist of segments (or formants of certain vowels) which correspond to certain (historical) sound categories. This means that string data (e.g. IPA transcriptions of an entire word) does not return a feature value as specific as the data types mentioned above. For example, Heeringa's (2004:267) approach uses correlations between each dimension from the aggregate distance extracted with MDS and single-word distances (both calculated with Levenshtein Distance) to find the words which correlate to the overall aggregate pattern the most. Firstly, this method does not necessarily correlate to one particular dialect group, as a dimension might capture more than one dialect group. Secondly, this approach returns a number of words which still requires manual inspection and identification of the features which distinguish different dialect groups. Prokić et al.'s (2012)

approach also gives a word-based output when the raw transcription is used.

In order to get a segmental output, applying Multiple Sequence Alignment (MSA) to the transcription before distance calculation seems to be a solution. Sung and Prokić (2024a) applied multiple sequence alignment on their German data, before evaluating the three feature extraction methods (FA, FLD and nPMI). All three methods are able to return segmental features during the feature extraction process. This has shown that MSA is useful in giving the precise output during feature extraction. This yields a difference with other approaches such as FLD (when the input data are strings), where further inspections of the entire string are needed to find out precisely which segment the extraction is targeting. In addition, MSA also retains the full transcription, which does not require manual taxation (Goebl 1984).

In this chapter, transcriptions of each word in the data set were multi-aligned using the LingPy library (List et al. 2021). Yue, like many other Southeast Asian languages, has a fixed syllable template, hence the *template-based alignments* (Wu et al. 2020) has been used to multi-aligned the Yue data. As a result, each column in the MSA comprises of individual segments, i.e. consonants and vowels (monophthongs and diphthongs are counted as single vowels). This is illustrated in Figure 6.4.

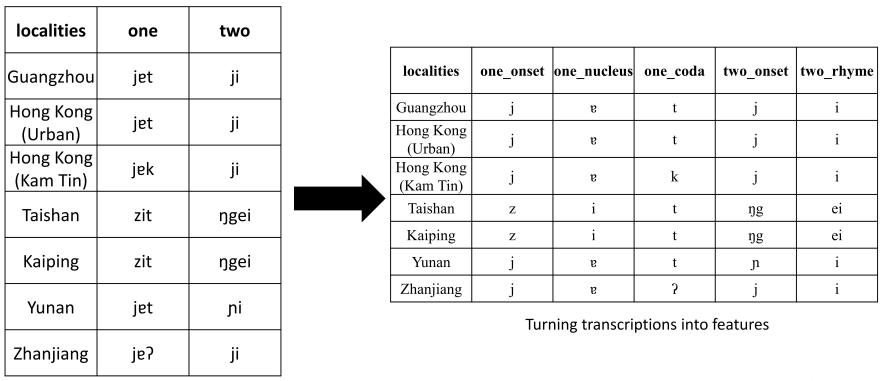


Figure 6.4: Illustration of Multiple Sequence Alignment

In MSA, each column contains different variants (different phonetic realisations) of the same variable (consonant or vowel phonemes or individual segments of a word), which is labelled as a dialect 'feature'. This

step has a lot in common with the way one establishes correspondences with the comparative method (Kessler 1995:134; Wu et al. 2020:8). The multi-aligned data is then manually checked for potential misaligned segments.

Distance Calculation

This procedure is a step which transforms qualitative data, i.e. columns of phonetic realisations of different segments of words in the multi-aligned data, into quantitative data, i.e. dialect distances. Unlike processing transcriptions with Levenshtein Distance, multi-aligned data are now categorical.

The calculation is based on the inverse value of *Relative Identity Value* (*RIV*, Goebl 1982, 1984), also known as the *Relative Distance Value* (*RDV*, Goebl 2018). The formula for RDV is provided in (1) below. To calculate RIV in a pairwise comparison (between two dialects), the number of matching features or *Co-identity* (*COI* in (1)) is divided by the total number of features compared, i.e. the number of matching columns (COI) plus the number of unmatching columns or *Co-difference* (*COD* in (1)). The resulting value is the similarity of the dialect pair ranging from 0 to 1. The RDV is 1 – RIV, which gives the pairwise distance instead of similarity.

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

The calculation of RDV was applied to all the dialect pairs in the data (Prokić and Nerbonne 2013). These distances are stored in a distance matrix and analysed by means of cluster analysis.

Since the methodology of distance calculation is different from the one in Chapter 5, the distance matrix is expected to be a bit different. The correlation of the two matrices are hence tested[4] and the two matrices are strongly correlated, with $r = 0.988$, $p < 0.0001$. This shows that despite the difference in the distance calculation, the combination of MSA and RDV can yield similar distances as Levenshtein distance.

[4]The Mantel test (Mantel 1967) was used to find the correlation between the distances matrices calculated using Levenshtein distance (Chapter 5) and Relative Distance Value (this chapter). More about the Mantel test can be found in Chapter 8.

Cluster Analysis

Introduced in Chapter 3, cluster analysis refers to the partition of objects (dialects in our case) into groups (Manning and Schütze 1999). In this chapter, Ward's method (Ward 1963) is chosen, and the 5-cluster solution is used (as illustrated in Figure 6.5).[5] These decisions are based on the results from the Levenshtein distance analysis in the previous chapter, for consistency.[6]

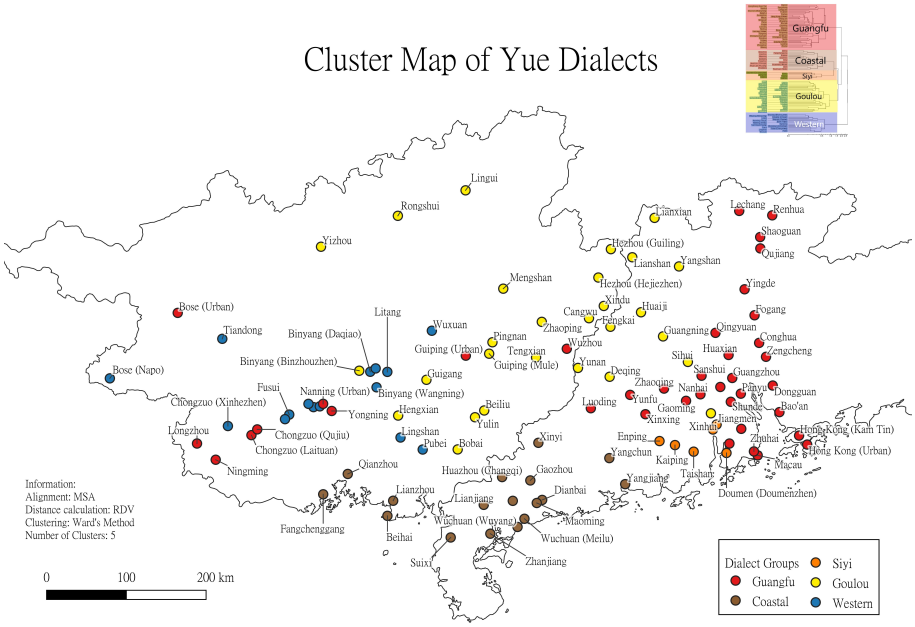


Figure 6.5: Cluster Map of Yue Dialects in Guangdong and Guangxi (MSA)

Despite the differences in the distance calculation, the classification based on cluster analysis for the MSA data has shown resemblance with the classification proposed in the previous chapter, as well as the *LAC* (Chinese Academy of Social Sciences (CASS) 2012). Firstly, the differ-

[5]Performed on *Gabmap* (Nerbonne et al. 2011; Leinonen 2010).

[6]The distance matrix from MSA highly correlates with the Levenshtein analysis in the previous chapter. Like the Levenshtein analysis, UPGMA also finds a lot of single- or double-member clusters. Since the two matrices are highly similar, the same clustering decisions are taken for the current MSA analysis.

ences with the previous chapter mainly lies between Cluster 1 (see red circles in Figure 5.15) or 'Central' dialects (see Table 5.2) in Chapter 5 and the 'Guangfu' dialects in this chapter (Figure 6.5). In this chapter, the traditional Yongxun dialects are now grouped with the 'Guangfu' dialects, whereas in the previous chapter, these dialects are classified as the 'Central' dialects in Cluster 1. In Chapter 3, it has been shown that the traditional Guangfu, Yongxun and Goulou dialects are part of the 'Inland' (green) continuum. The membership differences of the Yongxun dialects might be due to the differences in the distances calculated using two different metrics. Furthermore, traditional Yongxun dialects being classified as 'Guangfu' also makes sense in terms of its recent formation. As mentioned at the end of Chapter 5, Yongxun dialects were formed through recent migrations. There is no contradiction with the cluster analysis. It should be noted that the cluster membership does not affect the feature extraction process here. As shown below, nPMI only requires a pre-defined grouping of dialects as an input, it does not limit to only one grouping. However, since this study is meant to extract features using a data-driven approach, the grouping from the cluster analysis (of the MSA distances) will not be altered manually.

When comparing with the LAC, the differences lie in the number of groups, namely reducing from 8 dialect groups (Southern Pinghua as the eighth group) 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.

6.3.2 Feature Extraction

Pointwise Mutual Information, or PMI, is an association measure based on probabilities and co-occurrence (Church and Hanks 1990). Originally used in detecting word association based on their co-occurrences, it has also gained popularity in dialectometry for the use of automatic inference of segmental distances (Wieling et al. 2011). The idea behind PMI is comparing the probability of observing two categories, x and y, together (joint probability) and independently (by chance). The assumption is that if there is genuine association between x and y, 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 certain element occurs within the respective column) include probability of variant, x, or in the mathematical notation, *p(x)*, found within one column in the MSA data and the probability of dialect group, y, or in the mathematical notation, *p(y)*, within the classification column based on the cluster analysis in Section 6.3.1. Lastly, the final probability required is the co-occurrence of variant x given dialect group y, or in the mathematical notation *p(x,y)*.

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

The PMI scores are calculated using 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. All PMI scores are 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. This is illustrated with a toy example in Figure 6.6. 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.

An nPMI score of 1 indicates a perfect association between the variant and the dialect group. Lower nPMI scores can indicate either not all dialects show the respective variant or the variant is found elsewhere other than the dialect group under investigation or both. In general, the more we find these traits, the lower the nPMI score will be. To illustrate the idea, three distribution maps of three feature values are shown in Figure 6.7[7] based on the Flemish dialect group in Belgium, and their respective nPMI scores are given. The gray area is again the target dialect group (Flemish in this example), and the dots are the dialects which posses a particular feature value.

[7]The examples are created by the author using the GTRP Dutch dialect data available on *Gabmap* (Nerbonne et al. 2011; Leinonen et al. 2016).

Data

Varieties	Feature 1	Dialect Group
1	i:	A
2	e:	A
3	i:	A
4	i:	A
5	i:	A
6	e:	A
7	i:	A
8	e:	B
9	e:	B
10	e:	B
11	i:	B
12	e:	B

Co-occurrences

	Feature 1		
	***Group* A**	***Group* B**	
***Variant* i:**	5	1	*P(variant i:) = 0.5*
***Variant* e:**	2	4	*P(variant e:) = 0.5*
	P(Group A) = 0.58	*P(Group B) = 0.42*	

Calculate associations

Variants	Groups	Associations (PMI)	Normalisation
i:	A	$log_2\left(\frac{0.42}{0.5\times0.58}\right) = 0.543$	$\frac{0.543}{-log_2(0.42)} = 0.433$
i:	B	$log_2\left(\frac{0.08}{0.5\times0.42}\right) = -1.392$	$\frac{-1.392}{-log_2(0.08)} = -0.382$
e:	A	$log_2\left(\frac{0.17}{0.5\times0.58}\right) = -0.771$	$\frac{-0.771}{-log_2(0.17)} = -0.301$
e:	B	$log_2\left(\frac{0.33}{0.5\times0.42}\right) = 0.652$	$\frac{0.652}{-log_2(0.33)} = 0.407$

Figure 6.6: Calculation of nPMI scores with toy example. The probability of co-occurrence is calculated using the instances of the co-occurrence divided by 12 (total number of occurrences).

In Figure 6.7, when a dialect variant is found almost everywhere within the dialect area (West Flemish), but generally not outside (as shown in 6.7a), the nPMI score is rather high, which indicates a strong association. If a variant is found mainly within the dialect area, but it is not as representative, then the nPMI score will be lower (as shown in 6.7b). Lastly, if a variant is found within the dialect area but also very widely in the entire dataset, then the nPMI score is very low (even a negative score, as shown in 6.7c).

In addition, *Exclusivity* and *Representativeness* have also been calculated 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 of how distinctive these features are for a certain dialect group in relationship to the whole dialect area.

6.4 Top features in different Yue sub-dialect groups

Following the procedures described above in Section 6.3, the top 20 features of each of the 5 dialect groups have been identified. These are

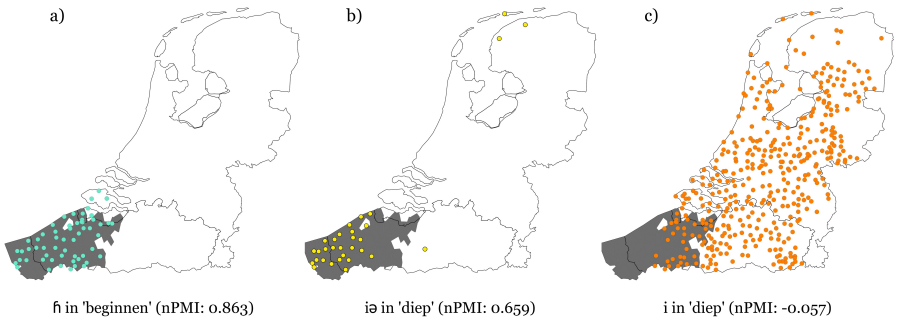


Figure 6.7: Comparison of associations (nPMI scores) between 3 Dutch dialect variants and the West Flemish area

the most strongly associated features per dialect group, ranked by the nPMI score. In addition, the exclusivity and representativeness scores are also included.

To interpret the tables, there are a number of things to keep in mind. The 'Variant' column refers to the value of the feature, i.e. the realisation of a certain linguistic variable (a column in the multiple sequence alignment). The 'Feature' column contains annotated features, instead of the raw MSA features (i.e. the position of the syllable in a word), as described in Section 6.3.1. The annotation contains the variable name (a segment), as well as the lexical item in which the segment is found. The variable names are divided into three types: synchronic, diachronic and medials. Synchronic variable names are based on the realisation of the variable in the Guangzhou dialect. The Guangzhou dialect here acts as a reference system for dialect comparison. This system seems to be adequate for the analysis of Yue dialects, as it: 1) contains innovations unique to the Guangfu dialects (which Guangzhou is part of), 2) preserves some 'archaic' features relative to other dialect groups, e.g. Guangzhou [y] vs. Coastal dialects [i] and 3) it is the most widely-spoken and used reference system[8] within Yue dialectology.

The second type of annotation is diachronic. These are marked with a '*'. The historical sound categories come from three time periods, namely Middle Chinese, Proto-Yue and Early Modern Cantonese (from sources in the 19th century). The three time periods are not differentiated in the

[8]See Zhan and Cheung (1987), Zhan and Cheung (1994) and Zhan and Cheung (1998).

annotations, but they are indicated in the analysis in each subsection when necessary.

The reconstructed values of the historical sound categories are used instead of the name of the sound categories. In the description of dialects in Chinese dialectology, very often the features are labelled with the Middle Chinese sound categories (represented by Chinese character). Studies following this tradition often state the sound categories (without sound values) and then provide their corresponding reflexes or the split-merger patterns in the dialects. The goal of having a diachronic annotation is not simply stating the correspondences. Reconstructed sound values are more useful for interpreting and explaining the existing variation in the Yue-speaking area, as illustrated in Sung (2020) and Sung and Prokić (forthcoming). They serve as a tool to explore the historical origins (explanations) of the present-day dialect landscape. Without the reconstructed sound values, discussions on the non-attested stages and exploration would not be possible.

There are numerous reconstructions for Middle Chinese. However, many tend to focus on maintaining the historical sound categories, rather than reconstructing the sound values, which could allow us to look at sound changes beyond correspondences (partly due to the difficulties in reconstructing the manner of articulations of some onsets as well as the values for the rhyme table categories). This is reflected in the choice of symbols for each sound category. For instance, Pulleyblank (1984) examines Karlgren's reconstruction, as well as evidence from dialects, Sino-Xenic languages, transcriptions of Sanskrit, and ancient rhyme dictionaries and rhyme tables in order to reconstruct Early and Late Middle Chinese. However, his reconstruction is not entirely IPA-like. For example, the symbol for the *Zhi* initial (知母) is written as <tr->, where <-r-> is used as a retroflex marker (Pulleyblank 1984:67). Sometimes, these notations can be converted to IPA, e.g. <-aă> represents [ɑ]. Norman (2006:233) on the other hand attempts to provide a reconstruction called *Common Dialectal Chinese* (CDC) by examining "the categories of the *Qieyun* (ancient rhyme dictionary) and systematically eliminating those features not reflected in the modern dialects" (Norman 2006:233). This method should "essentially yield the same results" as the comparative method. CDC provides a reference system for dialect comparison, but it cannot be counted as a reconstruction, since it is still not reconstructed from the bottom-up. Another notation representing Middle Chinese sound categories (from the ancient rhyme dictionaries or anno-

tations) rather than actual reconstruction based on present-day Sinitic languages, include Baxter (1992). Baxter (1992:27) states that the notation used in his book for Middle Chinese is "a convenient transcription which adequately represents all the phonological distinctions of Middle Chinese".

Zhu (2016) on the other hand reviewed 12 previous reconstructions of Middle Chinese, and through a comparison of Present-day dialects, additional rhyme books (rather than just *Qieyun*/ *Guangyun*) and Sanskrit transliterations, he has presented a reconstruction which is written in a broad IPA transcription. Despite there being a chance that the reconstruction is rather simplified (some issues may not have been addressed as in depth as some western reconstructions), the active use of dialectal evidence, evaluating the arguments from 12 previous reconstructions, and most importantly, making the reconstruction more phonetic makes Zhu's reconstruction more usable for this chapter than other reconstructions. For the reasons above, the reconstructions of Middle Chinese will follow Zhu (2016), unless specified (for more Yue-specific reconstructions). The correspondences between various reconstructions are provided in Appendix B.

The final category consists of the medials. As mentioned above, the reference system of the annotation is based on the Guangzhou dialect, which does not have medials (an -i- or -u- between the onset and the rhyme of a syllable). However, medials are common in many other Yue dialects. Therefore, it is necessary to establish a column for medials.

The following subsections will present the top features extracted for each dialect group.

6.4.1 Guangfu dialects

The majority of the features extracted in Table 6.1 are the presence of [œ] (Features 1, 4-6) and *ɲ-[9] > j- (Features 2-3, 7-9, 11-12 and 17). Multiple tokens of the same feature can be extracted in an analysis because there are different words which contain the same segments. The next few features that are associated with Guangfu dialects include the presence of [y] (Features 10, 13-14 and 18), *dz- > ts^h- (Feature 15), [ɔ] in the word 講 'to speak (colloquial)' (Feature 19) and absence of medial in 腳 'leg' (Feature 20).

[9] *ɲ- is from Early Modern Cantonese, which came from Proto-Yue *ŋj- according to Yue-Hashimoto's Yue-Hashimotoreconstruction.

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	tsʰ	*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

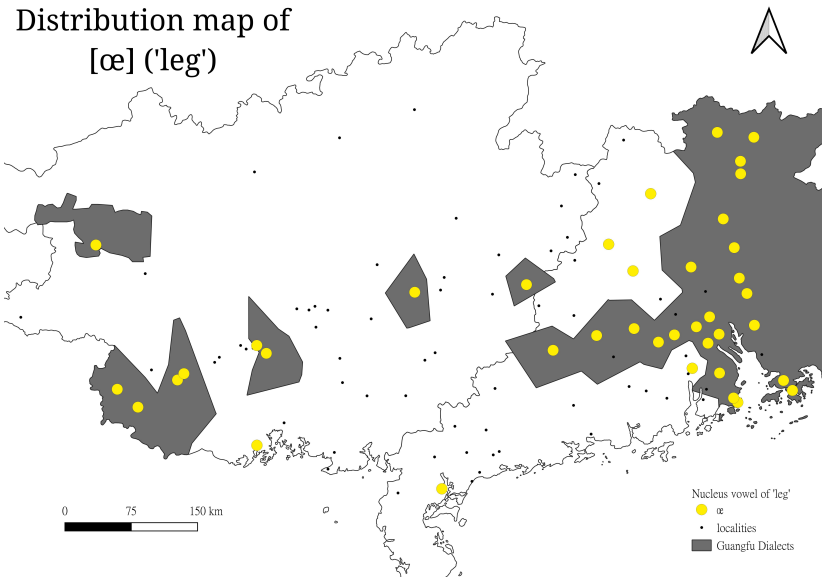
Table 6.1: Top 20 Features in Guangfu Yue Dialects. * represents reconstructed historical value of the segment (see Section 6.4).

The top 12 features can be said to be relatively exclusive to and representative of Guangfu dialects, as indicated by exclusivity and representativeness (above 0.7). The distribution maps in Figures 6.8a and 6.8b illustrate what relatively high exclusivity and representativeness look like, based on Features 1 and 2. For the rest of the features in Table 6.1, they are either less representative (but still exclusive), or not as exclusive, though the representativeness is high, like Feature 15 and 19.

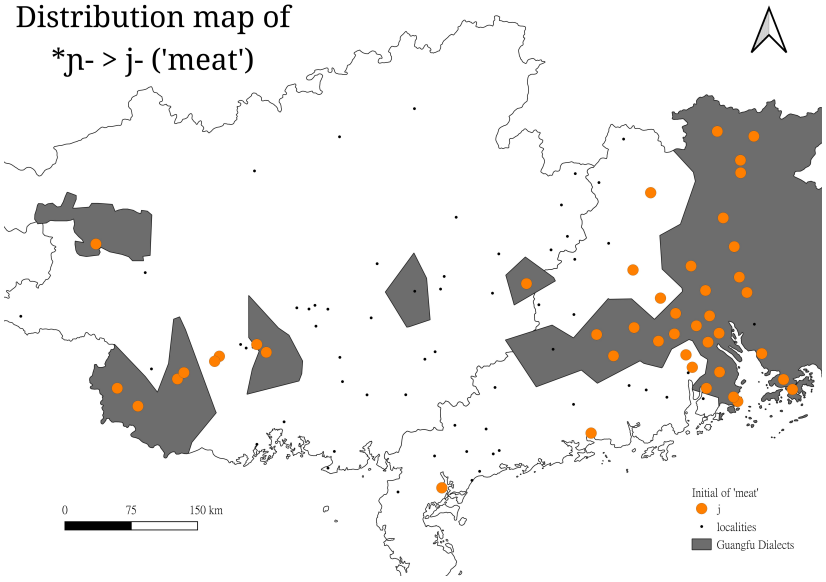
6.4.2 Siyi dialects

The top features which are closely associated with the Siyi dialects include *d- and *tʰ- > h- (Features 1-2), coda *-t > -p (Feature 3), *ɲ- > ŋg- (Features 4-7, 9-13), the presence of medial -i- in the word 'night' (Feature 8) and the presence of [z-] (Features 14-20).

All of these features show 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.



(a) Distribution map of the presence of [œ]



(b) Distribution map of *ɲ- > j-

Figure 6.8: Distribution maps of the top two Guangfu features. Gray areas indicate the Guangfu region, and the coloured circles indicate the dialects where the variant is found.

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 6.2: Top 20 Features in Siyi Yue Dialects. * represents reconstructed historical value of the segment (see Section 6.4).

6.4.3 Coastal dialects

In Table 6.3, the majority of the features involve the correspondence between Guangzhou [y] and Coastal [i] (Features 4-5, 7-10 and 12), which reflect the unrounding of *y. Other features include Guangzhou [w-] realised as a [ʋ-] (Feature 1), MC *ɣu- > ʋ- (Feature 2), MC *ɣ- > f- (Feature 3), *y (> *i) > ei (Features 6, 13, 15), development of *ɪ (Feature 14), *-t > -ʔ (Features 16-18), [au] as the rhyme in 鳥 'bird' (literary) and lastly, having a medial -u- for 肝 'liver' (Feature 20). The first three features have been found to be closely related in terms of their relative chronology with the sound changes involving MC *ɣ- (Sung and Prokić forthcoming). The same goes for the changes that involve the unrounding of *y and the diphthongisation of *i (Sung and Prokić forthcoming).

Rank	Variant	Feature	nPMI	Exclusivity	Representativeness
1	ʊ	w_cloud	0.767	0.917	0.647
2	ʊ	*ɣ_yellow	0.697	0.9	0.529
3	f	*ɣ_lake	0.672	0.6	0.882
4	i	y_tree	0.67	0.6	0.882
5	i	y_fish	0.664	0.65	0.765
6	ei	y_tool	0.635	1	0.353
7	i	y_rain	0.634	0.583	0.824
8	i	y_mouse/rat	0.632	0.583	0.824
9	i	y_village	0.623	0.75	0.529
10	i	y_pig	0.614	0.56	0.824
11	i	bird_lit_m	0.578	0.857	0.353
12	i	y_feather	0.561	0.484	0.882
13	ei	*y_woman	0.556	1	0.235
14	ei	*ɪ_teacher	0.555	1	0.235
15	ei	y_to go	0.555	1	0.235
16	ʔ	*-t_eight	0.537	0.833	0.294
17	ʔ	*-t_hair_head	0.537	0.833	0.294
18	ʔ	*-t_to kill	0.537	0.833	0.294
19	au	iu_bird_lit	0.534	0.833	0.294
20	u	liver_m	0.517	0.615	0.471

Table 6.3: Top 20 Features in Coastal Yue Dialects. * represents reconstructed historical value of the segment (see Section 6.4).

6.4.4 Goulou dialects

The top features of the Goulou dialect group show a major difference to the features extracted from the previous three dialect groups, specifically in their representativeness. Looking at the top 10 features (with nPMI scores lower than 0.56), we can see that most of them do not share a representativeness score above 0.33, despite the top 5 features (and Feature 7) having an exclusivity of 1. This pattern suggests that the entire Goulou dialect group does not have any unifying features which are both exclusive to and representative of this dialect group.

In Figure 6.9, the first token of the top 9 features in Table 6.4 are plotted on a multivariate distribution map.[10] A multivariate map visualises the distribution of multiple features simultaneously. Firstly, we can see that none of the Goulou dialects possesses all nine features shown on

[10]Feature 14 is grouped together with Feature 8 and 9 because they share the same nucleus and correspondence with ɐ in the Guangzhou dialect.

Rank	Variant	Feature	nPMI	Exclusivity	Representativeness
1	œ	ɔ_to speak_col	0.536	1	0.321
2	œ	ɔ_horn	0.534	1	0.321
3	∅	*x_red	0.512	1	0.286
4	∅	*ɣ_to merge	0.512	1	0.286
5	∅	*ɣ_summer	0.508	1	0.286
6	ts	*k_nine	0.5	0.737	0.5
7	θ	s_knee	0.486	1	0.25
8	au	ɐu_mouth	0.48	0.833	0.357
9	au	ɐu_hand	0.466	0.889	0.286
10	au	ɐu_autumn	0.464	0.889	0.286
11	au	ɐu_nine	0.464	0.889	0.286
12	θ	*d_to sit	0.458	1	0.214
13	p	f_tomb	0.456	1	0.214
14	au	ɐu_dog	0.448	0.769	0.357
15	t^h	*ts^h_autumn	0.446	0.684	0.464
16	p	*b_skin1	0.438	0.581	0.643
17	a	ɐ_enter	0.437	0.875	0.25
18	θ	s_four	0.437	0.875	0.25
19	θ	s_heart	0.437	0.875	0.25
20	θ	s_new	0.437	0.875	0.25

Table 6.4: Top 20 Features in Goulou Yue Dialects. * represents reconstructed historical value of the segment (see Section 6.4).

the map. If we look closely, we can see that the two dialects in Hezhou, Xindu, Zhaoping, Fengkai and Yulin possess six or more (out of the nine) most characteristic Goulou features. If we consider these features as characteristic features of the Goulou dialect group (even though they are not representative), then the area with the densest distribution of these features (a 'core' area) is around the north of the border between Guangdong and Guangxi (and Yulin). This leads to some follow-up questions: 1) why do many dialects cluster together as 'Goulou', despite not sharing the 'typical' Goulou features? This is perhaps due to the choice of the cluster algorithm. Is it appropriate, then, to use the 'typicality' of the Goulou features to analyse this proposed dialect group, since we are able to still find 'core' features to some extent; 2) What is the relationship between the Goulou and Western dialects? We see Feature 13 (onset of 'skin') and Feature 6 (onset of 'nine') are distributed where Western dialects are spoken. These questions are beyond the scope of the current thesis, and should be explored further in the near future.

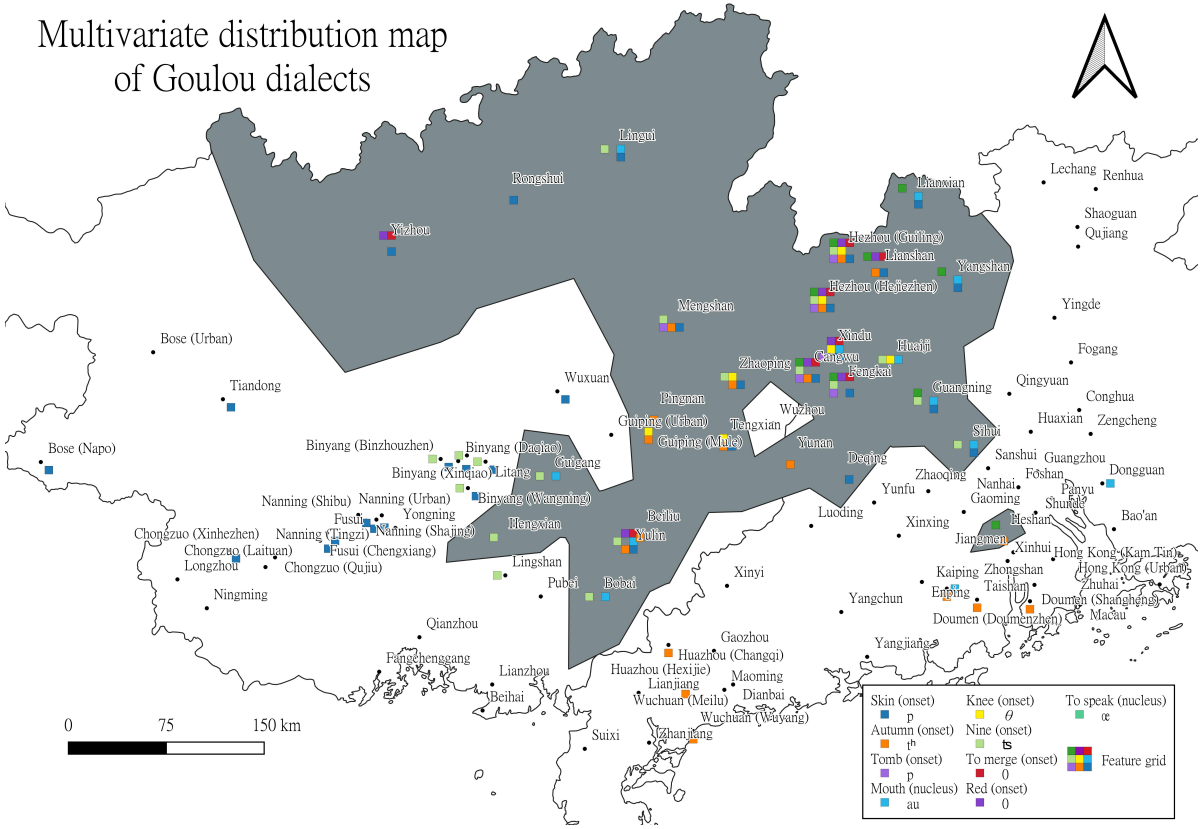


Figure 6.9: Multivariate distribution map of Goulou dialect features

6.4.5 Western dialects

Rank	Variant	Feature	nPMI	Exclusivity	Representativeness
1	i	ɛ_snake	0.902	0.882	0.938
2	j	*ɣ_summer	0.865	1	0.75
3	ɲ	*ŋ_to bite	0.865	0.929	0.812
4	ŋ	*ŋ_five	0.844	0.789	0.938
5	h	*ɣ_lake	0.83	0.727	1
6	h	*h_tiger	0.818	0.75	0.938
7	h	*h_fire	0.8	0.812	0.812
8	h	*ɣ_yellow	0.795	0.857	0.75
9	u	ɔ_to sit	0.794	0.917	0.688
10	ui	y_pig	0.765	1	0.562
11	ui	y_rain	0.765	1	0.562
12	ui	y_mouse/rat	0.764	1	0.562
13	ui	y_tree	0.764	1	0.562
14	o	œ_double	0.761	0.8	0.75
15	ɲ	*ŋ_tooth1	0.759	0.846	0.688
16	h	j_rain	0.746	0.652	0.938
17	a	ɔ_to speak_col	0.746	0.652	0.938
18	o	ɔ_yellow	0.733	0.75	0.75
19	əu	ɐu_cow	0.73	1	0.5
20	əu	ɐu_hand	0.73	1	0.5

Table 6.5: Top 20 Features in Western Yue Dialects. * represents reconstructed historical value of the segment (see Section 6.4).

The feature with the highest nPMI score for the Western dialects is [i] as the nucleus vowel in 蛇 'snake' (Feature 1). Other features with high nPMI scores include *ɣj- > j- (Feature 2), *ŋ- > ɲ- (Feature 3), *ŋ- remains non-syllabic in 火 'fire' (Feature 4), *ɣ- > h- (Feature 5, 8), *h- retention (Feature 6-7). Other features include [u] as the nucleus vowel for 坐 'to sit' (Feature 9), [ui] corresponding to Guangzhou [y] (Features 10-13), [o] corresponding to the [œ] in the Guangzhou dialect (Feature 14), the correspondence of [h-] to Guangzhou [j-] (Feature 16), the correspondence of [a] to Guangzhou [ɔ] in 講 'to speak' (Feature 17), the correspondence of [o] to Guangzhou [ɔ] in 黃 'yellow' (Feature 18) and the correspondence of [əu] to Guangzhou [ɐu] in Features 19-20.

6.5 Discussion and Conclusion

6.5.1 Features in the traditional classification

Using the LAC (Wu 2007, 2012; Tan 2012) classification as a reference, we can see that not many features used in the traditional classification are found in the feature extraction analysis using nPMI. This implies that the proposed features in the traditional classification are not closely associated with each dialect group.

The features that are identified as important in the traditional classification that are also identified by nPMI[11] are shown in Table 6.6:

Dialect group	Features
Guangfu	*None*
Siyi	*t^h- > h-
Goulou	*b- > p- ('skin')
Coastal	*None*
Western	*None*

Table 6.6: Features that are identified by nPMI and the traditional classification

As shown in Table 6.6, not all dialect groups' top features were used in the traditional partition of Yue sub-dialects. In addition, the only feature that can be said to be exclusive is *t^h- > h- in Siyi dialects.[12] Goulou's *b- > p-, on the other hand, is indeed a unique characteristic, but only if Western (Guinan) dialects were not considered (like in the LAC). Therefore, we cannot say that the LAC successfully identified a characteristic feature for Goulou dialects.

The partition of the dialect groups under the cluster analysis is still similar to the traditional classification, which suggests that the intuition (probably based on the experience and knowledge) of the dialectologists is somewhat correct. However, the failure of identifying the exclusive, closely associated features in most of the dialect groups might reflect influences from the broader classifications of Sinitic languages (e.g. Ting

[11] I.e. features that are ranked as important, by showing up as one of the top 20 features.

[12] Nucleus vowel [u] in 'to sit' in the Western dialects was identified by Li (2000), but since it was not included in the LAC, it is not included in Table 6.6.

(1982) proposes the use of aspiration patterns as a criterion for classifying major branches of Sinitic) and some subjectivity in the scholars' analyses. Hopefully the nPMI extraction method can help future dialectologists in solving the issue of manually finding associative features, which is susceptible to overlooking important features.

6.5.2 nPMI threshold

Out of the 5 dialect groups, it is clear that not all of the top 20 features are closely matching the geographical distribution of the dialect groups, i.e. not all features are equally exclusive and representative. Goulou dialects is a classic case for this, since none of the top 20 features are characteristic for the entire group. It is also the case that the nPMI scores for the top 20 Goulou features are rather low compared to the other dialect groups. Based on the manual inspection of the features for each group, is it possible to derive a rough indication for when a feature is not going to be useful, using the nPMI score?

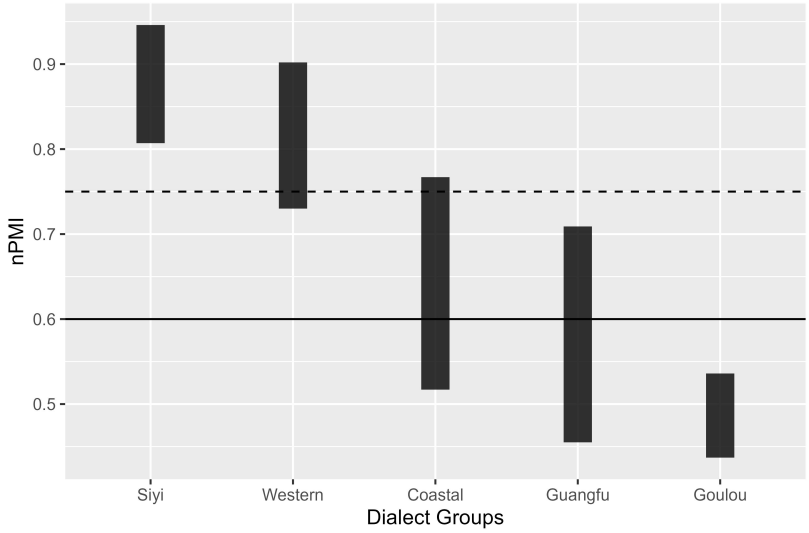


Figure 6.10: Ranges of nPMI of top 20 features for each dialect group

Figure 6.10 shows the range of nPMI scores of the top 20 features for each dialect group. Western and Siyi dialects have all top features above nPMI of 0.7 and 0.8 respectively. The first observation we can make (from these two groups) is that features with an nPMI score over

0.75 are excellent characteristic features that are both exclusive and representative of the dialect group. To some extent, this overlaps with the range of the Coastal dialects. The only feature that is above nPMI 0.75 is the labial approximant onset of 'cloud'. This feature is found in almost all Guangdong Coastal dialects (with some spill-overs to neighbouring dialects), but it is not found in the 4 Guangxi Coastal dialects. The lack of coverage of the western Coastal dialects suggests that perhaps 0.75 as a threshold is the lowest it can be. However, this threshold makes the features of the Guangfu dialects not very useful for distinguishing themselves from others, which is not true. We have seen on the maps in Section 6.4.1 that the top two features cover quite a lot of the Guangfu dialects, and are exclusive shared innovations.

Rather than setting one hard threshold, I would argue that it would be more useful to set 2 thresholds to help identifying features extracted using nPMI. For features above nPMI 0.75, they are excellent characteristic features which match the geographical distribution dialect groups closely. Features with an nPMI ranging from 0.6 to 0.75 (the solid line in Figure 6.10) are still characteristic, but one should expect more spill-overs (e.g. characteristic Guangfu features in Goulou dialects) or having more dialects lacking the feature in the group (e.g. Guangxi Coastal dialects lacking the labial approximant). For features lower than 0.6, they are either too localised or too widespread for a dialect group, as illustrated extensively from the Goulou dialects.

It should be noted that these thresholds are set as a guidance, they are by no means absolute and users of the method should always evaluate the features accordingly.

6.5.3 Historical interpretation

The analysis of the extracted features in the previous subsections remain synchronic, since it is beyond the scope of this chapter to go into the depth of the historical developments of these features. However, it by no means is the limit of the current analyses.

Although the features extracted are based on a synchronic comparison, these differences all arose from sound changes which are confined to a certain dialect area, which are rarely discussed. Results from the feature extraction must be interpreted further, especially with a historical dimension, if one wants to understand the dialects more.

Hypotheses on the historical developments can be formulated as a

continuation and a motivation for further research in the reconstruction of Yue sub-dialects, using a dialect geographical approach with the help of the feature extraction method. It should also be mentioned that the use of the current set of methods is not limited to the Yue dialects only. The methodology presented in this chapter can serve as an example of computer-assisted historical linguistics (Wu et al. 2020), when the historical dimension is also investigated. This is illustrated in e.g. Sung and Prokić (forthcoming), which investigates the relative chronology of some shared innovations in Yue dialects.

6.5.4 Conclusion

This chapter attempts to explore the problem of explaining the clusters in a dialectometric analysis by using a novel method in automatic feature extraction. The results show its usefulness in understanding the cluster groups. Furthermore, the methodology also offers a list of features for different dialect groups for future investigations in historical Yue dialectology. Last but not least, the set of methods introduced in this chapter is not Yue-specific. It is also not limited to phonetic variation only, given that categorical features are used.