Veneto Dialect Classification - Final report

Andrea Ceolin CEOLIN@SAS.UPENN.EDU
Nari Rhee NRHEE@SAS.UPENN.EDU
Hong Zhang ZHANGHO@SAS.UPENN.EDU

Abstract

This project attempts to implement machine learning techniques for automatic language clustering. We first evaluate various feature representations such as bag-of-words, sound correspondences, bag-of-phonemes and bag-of-ngrams using supervised classification algorithms such as Logistic regression and Naive Bayes. We further explore unsupervised clustering models (Spectral clustering and Hierarchical clustering) to examine the existing language classes based on linguistic investigation, and explore unnoticed clusters. The analysis is based on a dialect variation dataset of Veneto, Italy.

1. Data

The Penn Linguistics/Price Lab Etymological Database System (PLEDS) has released an etymological database of Indo-European languages, coded with adequate linguistic information, such as its proto-forms, gloss, orthography, and phonetic representation (Noyer, 2017). From the database we extracted a subset of the Romance languages, specifically the Veneto dialects, from 34 villages with five dialect labels (See Figure 1). The five dialect labels are: Central, Coastal, Venice, North, and West. This broad classification has been supported by several studies (Zamboni, 1974; 1979). A wordlist of 612 unique words matched by their meanings was recorded from each village, taken from AIS, (Jaberg & Jud, 1987), and VIVALDI, (Bauer, 2010). We note, however, that wordlists are not complete for all villages. All the words are represented in the same consistent phonetic transcription. In this project, we i) test the established language classification scheme with regard to Veneto varieties, and ii) examine what previously unnoticed language clusters can be discovered by features generated from our dataset.

University of Pennsylvania, CIS 419/519 Course Project.

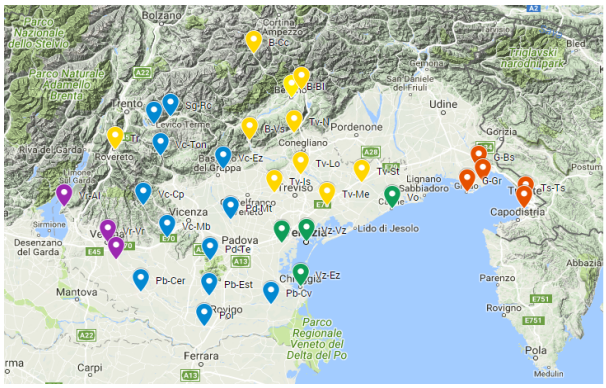

Figure 1. Map of five Venetian dialects: blue-Central; red-Coastal; yellow-North; green-Venice; purple-West

2. Supervised Classification

Bag-of-Words We first trained a Logistic Regression Classifier on the full dataset using directly all phonetic representations of individual words as features. This approach classifies languages in the same group according to some shared vocabulary. A testing accuracy of 97% (calculated through an average of 100 cross-validations on a 70/30 training/testing split) was obtained.

This result can be misleading, however, since languages that are traditionally grouped together tend to share a fair amount of vocabulary, or have regular lexical matches, both of which can be the result of contamination (e.g. loanwords) through contacts between groups of people. To deviate from this problem, we further consider factors that are relatively independent from such interactions, such as sound correspondences between languages and the phonemic categories.

It interesting to note that one specific point is consistently misclassified, i.e. the village at the extreme south, Pol (which stands for the village of Fratta Polesine). Looking at the dataset, it is to some extent surprising that there are indeed words shared by the speakers in Pol and other speakers, in particular those in the northern mountains (in the area of Belluno, labeled 'B'). Since a convergence in geographical distant spaces cannot be explained by horizontal transmission, there can only be two explanations:

either there is a parallel development in the same direction for the same trait (cf. homoplasy in evolutionary biology) or the convergence is a result of an archaism (Bartoli, 1925; 1945). On phonetic grounds, we can rule out the first option, and therefore we can argue that the misclassification of Pol must results from the fact that while all the dialects of the central area innovated some specific sounds, Pol didnt participate in the change, and therefore it is to some extent closer to the extreme northern dialects spoken in the mountain, which are conservative because of geographical isolation.

Sound correspondences Sound correspondences are often regarded as the gold standard in Historical Linguistics. In order to perform cladistics, linguist experts track the development of each phoneme that displays variation in the languages and use the correspondences to estimate the nodes which caused the split among two or more populations. This feature representation should be correlated with the previous one. However, the process cannot be automated: most of the correspondences are context-dependent or noisy, and therefore the process cannot be handled automatically (but cf. (Kondrak, 2002) for an attempt to do it).

With this model, we trained a Logistic Regression Classifier that achieved 88% testing accuracy. In principle we could expect this model to better than the first one, because it should be more robust with respect to contamination. However, the results are slightly worse. First, sound correspondences does not solve the problem of Pol, which means that even according to the strict criterion of sound correspondences, the amount of those shared with these groups is more significant (according to the classifier) than the amount of those shared with the other Central dialects. Second, the Coastal cluster becomes problematic, because there seems to be no clear sound correspondence that distinguish them from the others, and therefore they are attracted to the North group. This result has a possible explanation: since the speakers of the eastern coasts moved to the area not as a result of a single migration, but through the sea, it is plausible that they share a lexicon without having developed their own sound changes (which is a phenomenon more typical of clear-cut population splits). For this reason, we can consider the 88% accuracy of the Logistic regression trained on sound correspondences as a baseline for how standard linguistic methods would perform in a dialectal area with geographical and historical properties similar to these ones.

Bag-of-phonemes The bag-of-word and sound correspondence representations have some problems. As for the former, in order to perform well one needs to collect a lot of data. Additionally, we can easily imagine how in an area in which phonological change is more frequent, the bag-of-words model suffers from sparsity: two languages which share the same lexicon but happen to have all the common words which differ for a word initial consonant or vowel will not provide any more information than two completely unrelated languages. As for the latter, sound correspondences are just very expensive to obtain, and they typically require expert knowledge.

In order to create a more abstract feature representation with more generalizability, we created a classifier based on what we call a 'Bag-of-phonemes' model. Instead of having words as features, each dialect was represented with a set of phonemes. This model captures the intuition that even if the data is noisy, the frequency of use of every phoneme should still be a good estimator of the origin of the languages. Since we are moving from counts to frequency, we decided to switch to a Gaussian Naive Bayes classifier trained on a term frequency-inverse document frequency (TF-IDF) of the n-gram features. We achieved 66% testing accuracy with this feature representation. This approach is simpler and more general; however, this representation loses too much information about not only the lexical correspondences but also the sound correspondences. In fact, one of the reasons why the accuracy of this representation is lower is that a word-initial lenition of a particular vowel would have the same effect as a word-final lenition of the same vowel in a completely different context. Sound correspondences, on the other hand, would consider these two as two independent phenomena. For this reason we decided to move to a more complex representation.

Bag-of-ngrams We compared various n-gram representations of each village dialects, starting from n=2 up to n=4. While this featural representation allows us to have more abstract and general representations than the bag-of-words model, it can still captures interesting linguistics properties like possible consonantal clusters or context-dependent sound changes. For instance, in some North dialect of Veneto, word-final vowels tend to get deleted in certain dialects after a consonant (e.g. /koare/ [koar]; /fare/ [far]), which can be captured in the absence of the sequence re in these dialects for any n-gram of $n>2$.

Another potential advantage of any model which is trained on $n>2$ n-grams is that it can retrieve the full root of the word, which is typically composed by a sequence of 3-4 phonemes. The only potential disadvantage of a model which looks at long sequence is that we might expect it to have some troubles in handling sparse data, for instance data in which similar languages share only a particular consonant-vowel sequence for the majority of their words.

We trained a Gaussian Naive Bayes classifier using cross-validation, and we found that models of n=3 or n=4 achieves a testing accuracy of 82%, which approaches the standard set by sound correspondences. For this reason, we

decided to use trigrams as the basic representation to train our clustering algorithms.

Table 1. Summary of the classifier/feature representation implemented so far.

Feature Representation	Algorithm	Testing Accuracy
Lexical Representation	Logistic Regression	97%
Sound Correspondences	Logistic Regression	88%
Bag of Phonemes	Gaussian Naive Bayes	66%
Bag of 2-grams	Gaussian Naive Bayes	76%
Bag of 3-grams	Gaussian Naive Bayes	82%

From supervised to unsupervised model Our result from the text-classification approach shows that the provided class labels can be somewhat informative about the relationship between Veneto varieties. However, these classifiers are either constrained by the small number of instances with regard to a large number of features or mediocre performance due to more complex feature engineering. Furthermore, in real life a scenario in which every data point is labeled using historical information is rare. For this reason, in the next stage of the project, we will explore other algorithms that can capture the internal characteristics of language varieties without or with limited human judgment. We will explore models such as spectral and hierarchical clustering. These geometric models can help verify the proposals regarding language classifications relatively independent of human bias.

3. Clustering

Language variation and change can be modeled by graphical models, because of its close relation with population migration and contact. It can be expected that regions along the historical migration path, or connected by trade or other forms of human interaction, share more linguistic features in the dialect variety spoken therein. With this background, both spectral clustering and hierarchical clustering are suitable for finding the internal relations among dialect varieties.

3.1. Spectral clustering

We define the linguistic features using trigrams TF-IDF frequency. Euclidean distance between pairs of dialect varieties was used to construct the affinity matrix. Figure 2 plots the affinity matrix using 34 observations. It shows that the dialect varieties can be roughly partitioned into 3 to 4 clusters.

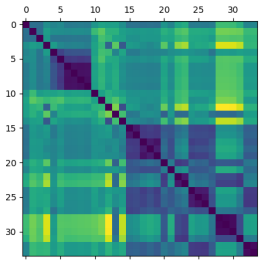

Figure 2. Affinity Matrix

In terms of data visualization, we tried to represent the cluster space by drawing lines on a map as to match the clusters returned by the algorithm. Since these classes come with no labels, the colors of the shaded areas were selected according to the majority class that the cluster is correctly identifying.

For instance, in the case of n=3 (Figure 3) the algorithm is clearly identifying the class of the Central dialects (the blue class), even though the two northern dialects are missed. The second natural class is the Venice class, which contains the dialects spoken in the urban area of Venice. This area expands to the north capturing also some of the yellow points, a result which can be explained by geographical contact. It is less intuitive to understand why also some of the dialects of the extreme west end up in the green class, but we are probably in presence of noise or casual similarities (i.e. parallel development). The final class, the yellow class, contains in addition to the North dialects also some leftovers from the other classes, and given that it goes through all the mountain cities it might be the most conservative class. Interestingly, it includes also the Coastal dialects, hinting again at a possible convergence between Coastal and North dialects.

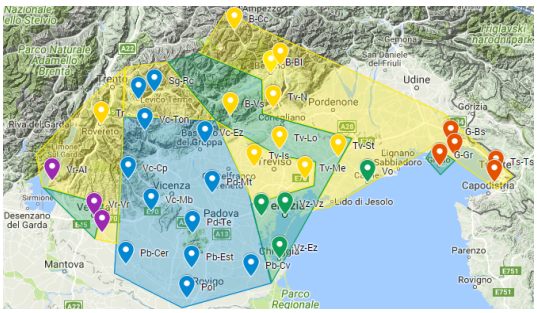

Figure 3. Spectral clustering for n=3

For n=4 (Figure 4), we notice a very interesting pattern. In this case, the clustering algorithm focuses on breaking the blue class in three different pieces, and puts together almost all of the other dialects (with the exception of some Coastal dialects). This means that the algorithm is sensible

to the fact that the dialects in the extreme south are showing a peculiar behavior (as detected by the sound correspondence Logistic Classifier) and a similar thing is true in the dialects at the extreme north of the area, which again might be displaying some conservative feature because of their geographical isolation. This somehow raises awareness towards the whole blue class and hints at the need of further investigation.

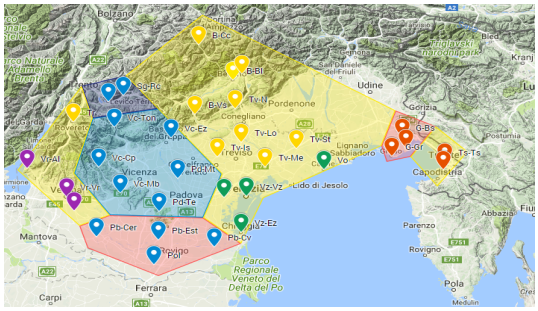

Figure 4. Spectral clustering for n=4

3.2. Hierarchical clustering

Hierarchical clustering is another good representation of the problem of language classification which provides an insight not only about the relationship between dialect groups, but also about the direction of language change. Hierarchical clustering allows us to draw a dendrogram, which is comparable to a tree representation of language cladistics. The dendrogram of reference of our tri-gram representation is the one in Figure 5.

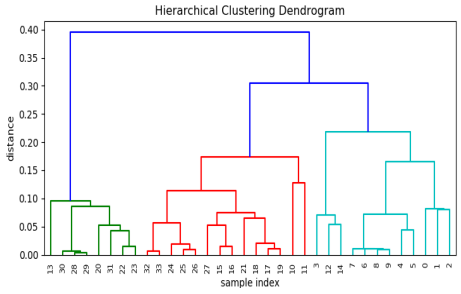

Figure 5. Dendrograms obtained through a Hierarchical clustering analysis on our trigram representation. Ward distance has been used to derive the similarity matrix.

From the dendrogram, we can see that we can easily identify 3 clusters. The result that we get (Figure 6) is similar to the one obtained through spectral clustering: we can clearly identify the 'blue' group (with the exception of the two northern dialects) while the rest of the languages are divided in two other groups that tries to capture a split between locations in the mountains and locations in the flat land. However, two of the 'North' locations are clustered together with the mountain locations, which can be explained as a sort of attraction of the 'North' group. The same thing is happening for one 'Venice' and two 'Coastal' locations, probably a result of the general noise contained in the data from the villages on the sea.

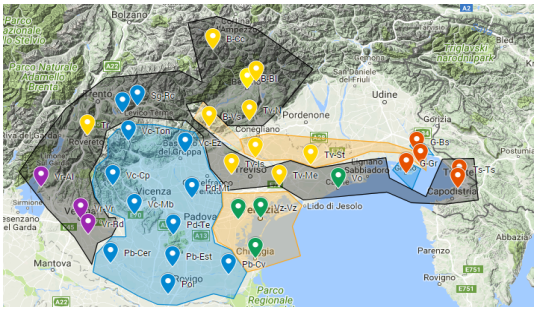

Figure 6. Hierarchical Clustering analysis for n=3.

If we increase n to 4, we notice a similar pattern to the one previously identified: rather than clustering together other groups, the dialect of 'Pol' is isolated. This means that both our clustering algorithms are able to immediately identify the data point which is missed by expert classifications, just by using a very abstract representation.

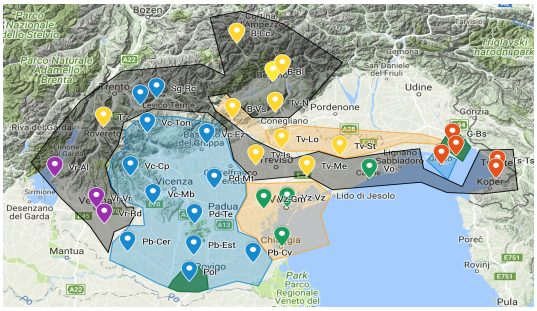

Figure 7. Hierarchical Clustering analysis for n=4.

4. Conclusions

In this project, we addressed the problem of classifying languages given a dataset of language varieties. Using supervised learning models, we first defined the best feature representation of the data, one that is more robust and general for language classification, yet with high accuracy. Among the representations we have tested, bag-of-trigrams with TF-IDF representation has turned out to have the best testing accuracy, especially for its automatibility. Using this feature representation, we have tried spectral and hierarchical clustering models to do unsupervised classification of the data. Both spectral and hierarchical clustering models largely corresponded to the big language groupings recognized by historical linguists; however, the mistakes of the clusters also provide us insight in the language classification problem. Future work remains to define how much data is necessary in order to achieve a similar performance, and whether our feature representation would work for datasets with more variation in the languages.

References

Bartoli, Matteo. *Introduzione alla neolinguistica:princìpi–scopi–metodi*, volume 12. LS Olschki, 1925.

Bartoli, Matteo Giulio. *Saggi di linguistica spaziale*. V. Bona, 1945.

Bauer, Roland. Le projet vivaldi: présentation dun atlas linguistique parlant virtuel. *Anuario del Seminario de Filología Vasca" Julio de Urquijo"*, pp. 71–88, 2010.

Jaberg, Karl and Jud, Jakob. *AIS: atlante linguistico ed etnografico dell'Italia e della Svizzera meridionale*. Unicopli, 1987.

Kondrak, Grzegorz. *Algorithms for language reconstruction*. University of Toronto Toronto, 2002.

Noyer, Rolf. *Penn Price Linguistics Lab Etymological Database System (PLEDS)*. https://n411.fmphost.com/fmi/webdPLEDS_ONLINE, 2017.

Zamboni, Alberto. *Veneto*, volume 1. Pacini, 1974.

Zamboni, Alberto. Le caratteristiche essenziali dei dialetti veneti. *Guida ai dialetti veneti*, 1:9–43, 1979.