Automatic classification of Veneto Dialects

Andrea Ceolin

August 16, 2018

Abstract

This project attempts to implement modern machine learning techniques to study dialectal variation. We first evaluate various feature representations of languages such as bag-of-words, sound correspondences, bag-of-phonemes and n-grams models, using supervised classification algorithms such as Logistic regression and Näive Bayes. Using the adequate feature representations, we further explore unsupervised clustering methods such as K-Means to examine the existing language classes based on linguistic investigation. The data analyzed are the varieties of Veneto dialects contained in the PLEDS database[1].

1 Introduction

Classifying languages into different groups with the aim of reconstructing the history of the speaker population is a task at the core of Historical Linguistics and Dialectology. Scholars interested in archaeology and genetics are also interested in proposals made on the basis of linguistic research because they can be used as independent evidence to reconstruct routes of migration. While the literature is full of classification proposals, the methods employed to support them are often opaque. For instance, while in certain cases isoglosses and sound correspondences are clear to identify the history of a dialectal variety, in some other cases local contact or the pressure of a language associated with the upper class might influence dialects to a point in which it is difficult to correctly classify them, and therefore geography and historical records are the only sources of reliable information. Furthermore, the evidence can also be controversial, and different scholars can come to different conclusions depending on their analysis of sound changes. In this project, we want to automatize the process of language classification by: i) designing a robust feature representation of a given language; ii) training a classifier using the feature representations we develop in order to test their accuracy in reconstructing the main groupings; iii) evaluate different proposals of clusters among languages.

[1] Many thanks to Hong Zhang and Nari Ree for the implementation of the algorithms and the descriptions of the problems, and Rolf Noyer, Don Ringe and Eric Eaton for important comments and suggestions

2 Data

The Penn Linguistics-Price Lab Etymological Database System (PLEDS) is an etymological database of Indo-European languages, coded with adequate linguistic information, such as proto-forms, glosses, and orthographic and phonetic representations (PLEDS 2017). From the database we extracted a set of 34 Veneto dialects (See Figure 1). The dialects have been grouped using five dialect labels, which represented the traditional classification of Veneto and are named after the primary regions where the dialects are spoken in the region: Central, Coastal, Venice, North, and West. This broad classification has been supported by several studies (cf. among others Zamboni 1974, 1979, Ferguson 2007). We selected a wordlist of 612 unique words matched by their meanings from PLEDS, based on the presence of variation and the lack of conflicting information. The data in PLEDS are mainly taken from the AIS (Jaber and Jud 1987) and from VIVALDI (Bauer 2010).

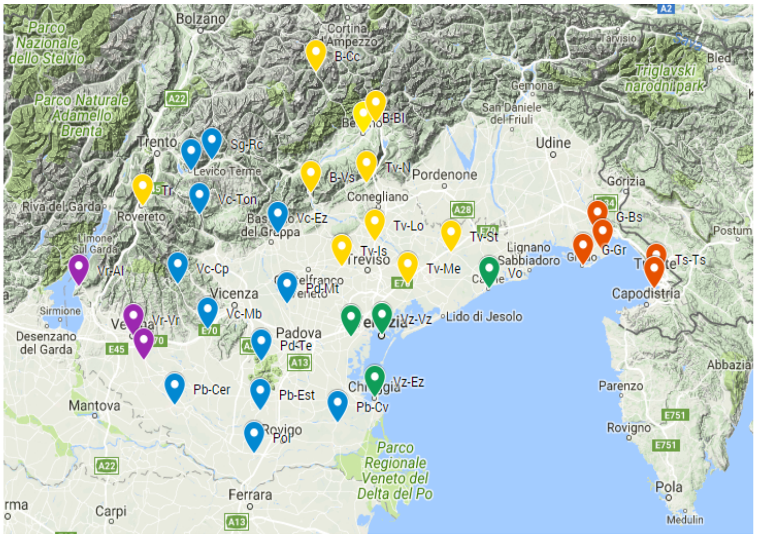

Figure 1: Map of the main groupings: blue-Central; red-Coastal; yellow-North; green-Venice; purple-West

3 Supervised Classification

3.1 Bag-of-Words

As a first step, we trained a Logistic Regression classifier on the full dataset using directly all phonetic representations of individual words as features. The classifier uses input data to estimate a weighting scheme for the features with the goal of maximizing the likelihood of each training instance belonging to its given class, and then can use the same weighting scheme to predict the class of unlabeled data. This approach captures when languages of the same dialect share the same phonetic representation of a word in contrast with the ones of other dialects. One interesting property of this approach is that by repeatedly training the classifier on a subset of the data and checking if the predictions on the unseen data match their labels, one can immediately identify the areas of lexical contact or those in which a high rate of sound change might have created some isolates that lost all the lexicon they had in common with their neighbors.

Surprisingly, the classifier yields a testing accuracy of 97% (calculated through cross-validation on a 24/10 train-test split on the dataset, for 100 randomized iterations), which basically means that the lexicon is quite distinctive of the five areas.

It is interesting to note that the accuracy does not reach 100% not because of random noise, but because of a specific point in the data which is consistently misclassified, i.e. the datapoint at the extreme south, Polesano ('Pol'), the dialect spoken in the location of Fratta Polesine, which the classifier puts with the 'North' or the 'Coastal' groups rather than with the 'Central' one. Looking at the dataset, it is to some extent surprising to find that there are indeed words shared by the speakers in Fratta Polesine and the speakers of other groups, but not with the other Central dialects. There are at least two patterns of contrast, which are summarized in the following table:

	Proto-Romance	Pol	Pd	Tv-Lo
'happy'	koŋtɛŋtu	koŋtɛŋto	koŋteŋto	koŋtɛŋto
'sunday'	dominika	domɛnega	domenega	domɛnega
'forever'	sɛmpre	sɛŋpre	seŋpre	sɛŋpre
'wool'	lana	lana	łana	lana
'monday'	lunis	luni	łuni	luni

First, it is reported that in Polesano some tense mid vowels are realized as lax before nasals when they are tense in the majority of the dialects of the region. This is reported in the Coastal dialects of Grado and Trieste and also in at least some varieties of Trevisano and Bellunese (like in the variety of Lovadina, 'Tv-Lo'), but not in the dialects of Padova ('Pd') and Vicenza ('Vc'). This is true in standard Italian as well.

Second, lenition of word-initial /l/ is absent in Polesano as it is in Trevisano (where it is transcribed with /l/), but it is active in Padova (where it is transcribed with the arciphoneme /ł/, that indicates lenition), probably under the pressure of the dialect of Venice (Zamboni 1974).

There is a third phenomenon which follows the same pattern, i.e. the reflexes of the palatization of /k/ before front vowels and /t/ before /j/ (which merge in the whole area):

	Proto-Romance	Pol	Pd	B-Cc
'wax'	kera	θera	sera	θera
'march'	martju	marθo	marso	marθo
'fifty'	kinqwaginta	θinqwanta	sinqwanta	θinqwanta

However, this phenomenon is not unique to Polesano, but also affects the other southern dialects of the region.

Given the evidence we provided, we can infer that the misclassification of 'Pol' results from two kinds of facts: i) the fact that while all the dialects of the central area show lenition of word-initial /l/ consistently and merger between proto-Romance /k/ before front vowels and /s/, ‘Pol’ did not participate in the changes. ii) the fact that ‘Pol’ midvowels are opened before /n/ + consonant clusters. Therefore, both shared retentions and shared innovations with other dialects made it look consistently different, at least from the lexical viewpoint, from the Central dialects, and closer to others.

A visual inspection on the variation of the data can be obtained through a Principal Component Analysis (PCA), that allows a 2D visualization of the data by retrieving the two main axes of variation. The PCA in Figure 2 shows that also languages of another three areas, all in the bottom part of the graph, are possibly subjected to similar phenomena: i) the other dialects spoken in the southern area, in spite of being recognized as Central by the classifier, appear close to Polesano (‘Pol’) in the graph, probably as a consequence of the similar development of the velar; ii) also in the north-west of the region, where there is a mixture of Central and North dialects, there must be some local convergence due to the closeness of the dialects of the Valsugana (‘Sg’, Central) and of Rovereto (‘Tr’, North): iii) the Coastal group seems also not to be well defined, with the dialects of Grado (‘G’) being close to various dialects from the western part. These are possible areas where in absence of the kind of specific evidence provided by these large wordlists a misclassification can be expected.

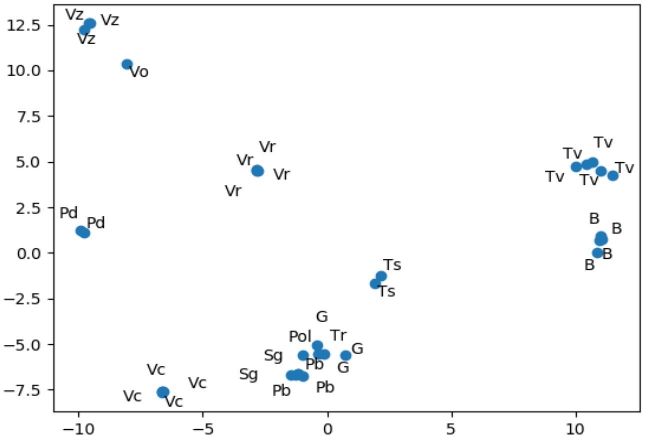

Figure 2: Principal Component Analysis on wordlists

3.2 Sound Correspondences

We then moved to sound correspondences, using 46 patterns of correspondences as features instead of 612 words. The number of correspondences is high because some correspondences are not clear, probably as result of different phonetic transcriptions, lexical exceptions or mistakes (for instance, there are 25 different patterns of midvowel correspondences). 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).

With this model, we trained a Logistic Regression classifier that achieved 88% estimated testing accuracy. In principle, before looking at the data we might have expected this model to perform even better than the first one, because it should be more robust with respect to loanwords. Instead, the results are slightly worse. First, it looks like moving to sound correspondences does not solve the problem of Polesano: the language is still classified as either a North or Coastal dialect, which means that the number of sound correspondences discussed in the previous section is not matched by a stronger signal that keeps the dialects together. Second, another cluster which becomes problematic is the Coastal cluster, because it looks like while there are words uniquely shared by the languages, there seems to be no clear sound correspondence that distinguish them from the other languages, and therefore they are attracted to the North group. This 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 classifier trained on sound correspondences as a baseline of how standard linguistic methods would perform in a dialectal area with geographical and historical properties close to these ones (i.e. the presence of both isolated areas like mountains and areas with a lot of traffic, like sea routes).

A visual inspection on the variation of the data through a Principal Component Analysis (Figure 3) confirms that there are some problems with the Central group, which looks scattered on the right part of the plot, and that the Coastal and North groups form a unique clade on the left. Another potential problem seems to be the dialect of Rovereto ('Tr'), displayed close to the Central dialects of the Valsugana ('Sg'), but this is not among the dialects misclassified by the classifier because it has many sound correspondences which are shared with the dialects from the Northern group and are probably lost in the dimension reduction of the PCA.

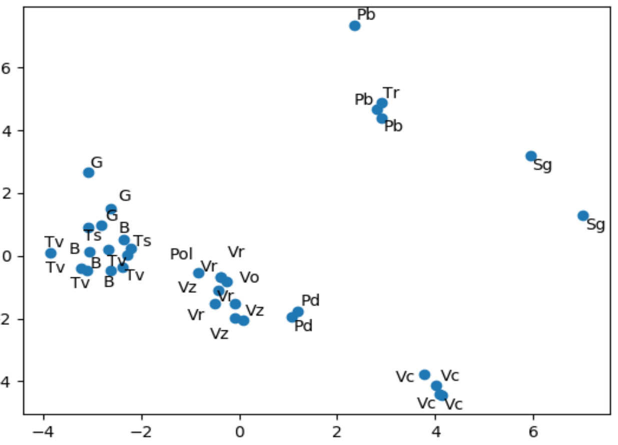

Figure 3: Principal Component Analysis on sound correspondences

3.3 Bag-of-phonemes and N-grams

The bag-of-words and sound correspondence representations, even though appealing in terms of their testing accuracy, are potentially controversial because of some properties.

As for the former, in order to perform well one needs to collect a lot of data to avoid sparsity. This turns out not to be a problem for this specific case: if we implement feature selection in the evaluation phase, we see that removing half of the 612 features would only hurt testing accuracy for a couple of percentage points (from 97% to 95%), which is a good indicator of the fact that our wordlists are highly redundant. However, 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 common words which differ for a word initial or word final syllable, or even a single consonant or vowel, will not provide any information more than two completely unrelated languages (cf. for instance French and Italian, which have almost no lexical overlap despite being closely related).

As for the latter, sound correspondences are just very expensive to obtain in terms of empirical work, and they typically require expert knowledge. In some cases, there can also be disagreement in the exact evaluation of some correspondence scheme.

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 phonemic counts, normalized for the number of total phonemes as to obtain phonemic frequencies. This model captures the intuition that even if the data is noisy because of accidental loanwords or language-specific sound correspondences, 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 Näive Bayes classifier trained on a term frequency-inverse document frequency (TF-IDF) of the n-gram features. This classifier estimates the posterior probability of the observations and, with labels representing prior hypotheses, it selects the most likely hypothesis that generated the data via Bayes' rule. Instead of estimating probabilities using directly term frequency, this estimate is corrected by inverse document frequency in order to reduce the weight given to very frequent (and thus uninformative) features.

Using a Gaussian Näive Bayes classifier with this feature representation we achieved 66% testing accuracy, calculated using the same cross-validation settings applied for the previous cases (100 iteration, 24/10 train-test split). 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, even though the two sound changes are independent from each other.

For this reason we also considered moving to more complex n-gram representations. We compared various n-gram representations of each dialect, starting from n=2 up to n=4. This kind of representation captures those sound correspondences that depend on a particular phonetic context. For instance, in all North dialects, word-final vowels tend to get deleted (e.g. [volpe] > [volp], 'fox'; [mare] > [mar], 'sea'), which can be captured in the absence of the sequence 'e#' 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 of 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. While models with $n<3$ do not reach good testing accuracy, a model of $n=3$ achieves a testing accuracy of 87%, 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.

A Principal Component Analysis of the trigram model (Figure 4) shows that the Central group (in the upper left), the North group (in the center) and the Venice group (in the upper right) are well recognized, with only one of the dialects from the Valsugana ('Sg') and Veneziano Orientale ('Vo') being misplaced, while there is not a clear centroid for the Coastal dialects and the West dialects, which are included among the North points.

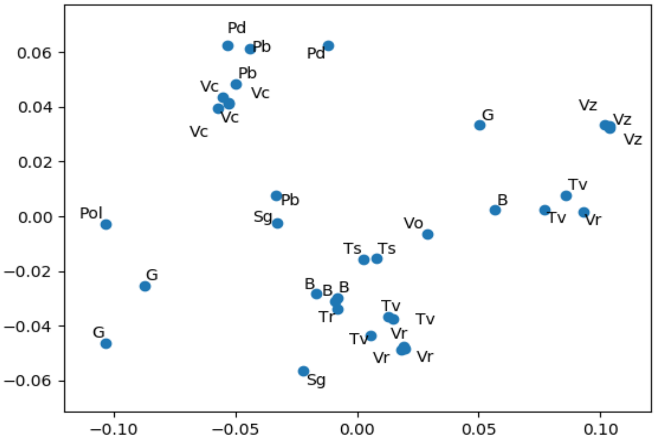

Figure 4: Principal Component Analysis on trigrams

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 (unigrams)	Gaussian Naive Bayes	66%
Bag of bigrams	Gaussian Naive Bayes	76%
Bag of trigrams	Gaussian Naive Bayes	87%
Bag of 4-grams	Gaussian Naive Bayes	86%

4 Unsupervised classification

So far we have not discovered anything new apart from the fact that the dialect at the south of the region, Polesano, shows a behaviour typical of dialects whose genetic signal has been covered by geographical contact or disturbed by parallel development, and the fact that the Coastal group turned out to be the most challenging to retrieve.

However, most of these conclusions were based on the assumption that the genetic affiliation of the dialects was given in advance, an assumption which does not hold in many real-world cases.

For this reason, we want to see what would happen in a scenario in which this information is not given. To do this, we move to unsupervised learning, i.e. learning without categorical labels. In this case, we are interested in clustering techniques, which address the question of which kind of clusters can be retrieved given a dataset of observations, a problem quite common in language classifications. The two most popular techniques are K-Means and Gaussian Mixture Models (GMM). K-means retrieves a number K of cluster, which must be given in advance, making the assumption that the clusters are similar in terms of their variance. GMM, on the other hand, allow clusters whose variance is not equal, but of course this property adds a lot of degrees of freedom, and therefore it means that convergence can be more difficult.

Running both techniques on the Veneto dataset, it turned out that K-means converges to a unique result, while GMM does not. Therefore, we investigate the results produced by K-means.

4.1 K-Means with Sound Correspondences

Since the number of possible clusters is a parameter that needs to be set for K-means, we explored various clusters proposed with K varying between 3 and 5. The result from K=3 applied on sound correspondences is displayed in Figure 5. As a visualization tool, we manually colored the areas corresponding to the different clusters proposed by the algorithm. Even though we tried to match the colors of the areas and those of the data points when there was a clear match, the colors do not have any particular significance.

In Figure 5, for K=3, we see that all the North varieties are grouped together, with the only exception of the dialect of Rovereto ('Tr'). This cluster also contains the Coastal varieties and Polesano ('Pol'), in agreement with the results provided by the Logistic Regression classifier on the dataset of sound correspondences.

The other two groups are less intuitive. One of the groups contain the dialects from the Bassa Padovana ('Pb') along with the dialects of the Valsugana ('Sg') and Rovereto ('Tr'), and is probably capturing the conservative part of the Central area, i.e. those dialects which have not been under the pressure of Veneziano. The other group, on the other hand, contains all the dialects around the cities of Venice, Padova, Vicenza and Verona.

Moving to K=4 (Figure 6) we see a correct split on this latter group that puts together the dialects of Verona ('Vr'). However, for K=5 we do not see a split that identifies the Coastal dialects or the dialects of Venice, but a split between the northern and the southern part of the Central group. This result hints at the facts that the dishomogeneity of the Central group can be one of the most challenging issues in the classification of the area. Note that this does not change if we increase K: some further tests showed that the Central group is the target of splits also for K=7 and K=8.

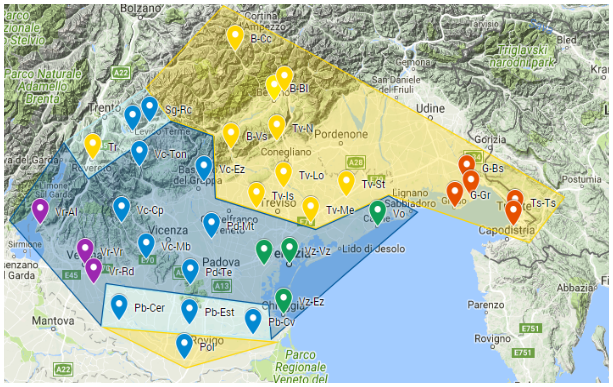

Figure 5: K-Means with K=3 on sound correspondences

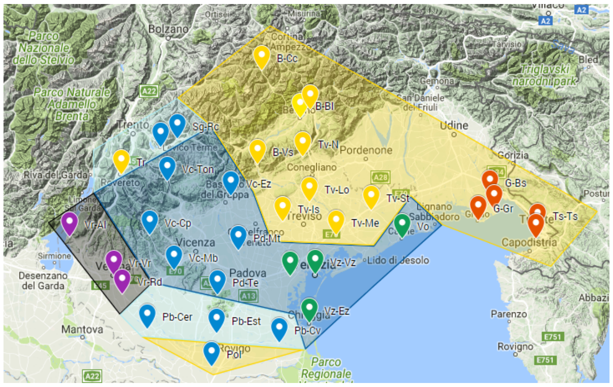

Figure 6: K-Means with K=4 on sound correspondences

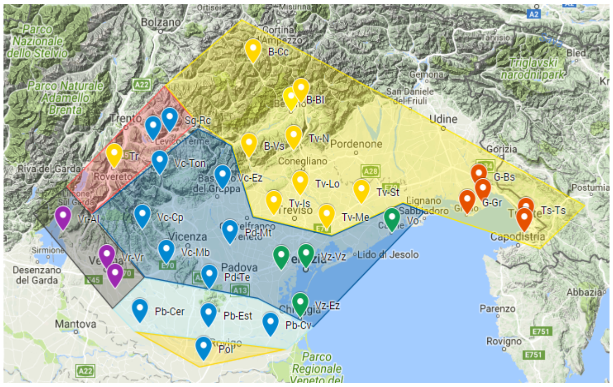

Figure 7: K-Means with K=5 on sound correspondences

4.2 K-Means with trigrams

Since the trigram representation provided an accuracy similar to the one provided by sound correspondences, it is interesting to see if the results in terms of clustering are also comparable.

In Figure 8 we have a K-Means analysis with $K=3$. We can see that there is a cluster which contains almost all the Central dialects, including Polesano, and leaves out just the two 'Sg' dialects from the Valsugana. This means that one of the three clusters identified by K-Means in this case almost overlaps with one of the real ones.

The second cluster identified by the algorithm contains most of the North dialects, with the exception of one 'B' dialect and two 'Tv' dialects. This group also contains the two 'Sg' dialects left out from the previous group, Veneto Orientale ('Vo') and also the varieties in Verona ('Vr'), which is not too surprising since given the constraint of only three clusters we might expect some groups to be merged.

Finally, the third cluster groups together the dialects of Venice with the exception of 'Vo' and the three Northern dialects which were not included in the previous group.

Interestingly, this time the Coastal varieties are grouped in the three different clusters for unclear reasons. It is clear from the map that some noise is clearly characterizing the eastern area.

Increasing K to 4 (Figure 9), we see that the Central group is affected, matching a result that we have seen in the previous case when K was increased to 5. A similar effect is reached if we increase K to 5 with this representation as well (Figure 10): in this case, we have a further split of the Central group that divides Polesano, the conservative areas (Valsugana, but only one of the two dialects, and Bassa Padovana) and the central areas (Vicenza and Padova).

These results match those obtained in the previous case: even changing the featural representation to phonemic trigrams, K-Means still splits the Central group before detecting the other clusters. Finally, like in the previous case, tests with higher values of K fail to retrieve the missing clusters.

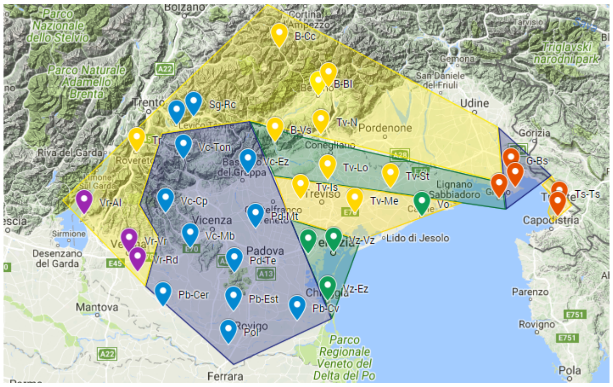

Figure 8: K-Means with K=3 on trigrams

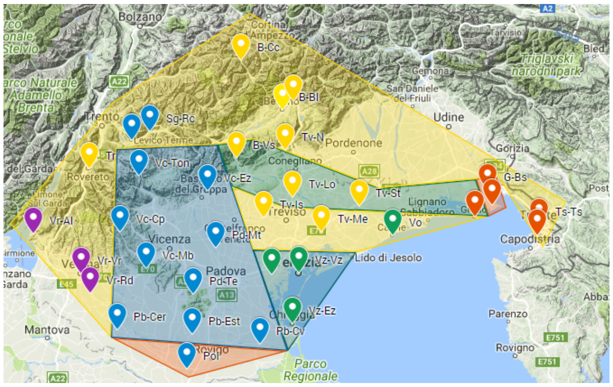

Figure 9: K-Means with K=4 on trigrams

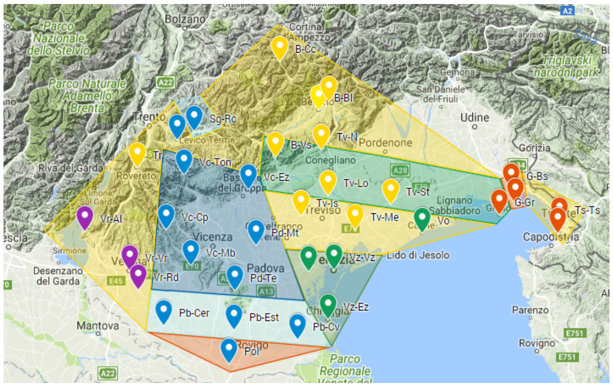

Figure 10: K-Means with K=5 on trigrams

5 Conclusions

In this project, we addressed the problem of classifying languages given a dataset of language varieties. Using supervised learning, we first defined the best feature representations of the data, those which were more robust and general for language classification, yet with high accuracy. Among the representations we have tested, sound correspondences and bag-of-trigrams with TF-IDF normalization turned out to have the best testing accuracy. Using these feature representations, we have employed K-Means to cluster the data.
Interestingly, both models converge in identifying the Central area as the most difficult one to characterize. Sound changes caused the area to display at least three different layers: a first layer which contains the conservative varieties (in Valsugana and Bassa Padovana), a central area which contains innovative varieties (Padova and Vicenza) and the area of Fratta Polesine which displays unique sound changes which also characterize some of the dialects in the North group.
This fact is one of the main challenges that computational methods which aim at clustering or cladistics have to face.
Potential ways to address the problem are the following:

1. **A more solid feature representation**. As we have seen, the trigram model suffers from the fact that it cannot correctly classify the Coastal dialects, which probably have undergone independent development and therefore contain a very weak signal. How-

ever, the sound correspondence representation yields more plausible results, therefore proving to be a more accurate representation, despite the testing accuracy reached by the Logistic Regression classifier. One question would be whether these techniques would perform better if 'spurious' correlations (i.e. those correlations which are not well supported or only partially supported by the data) are removed. This would require a more fine grained work on the specific varieties, one that was out of scope for this exploratory project.

2. **Different clustering techniques which are crucially dependent on spatial information**. It could just be that standard clustering techniques cannot work in a domain in which the data is noisy because of incorrect transcriptions or phenomena like independent development of the same linguistic change or borrowing. Therefore, ignoring geographical information is something that one cannot do. There are some machine learning techniques based on the concept of Random Field (cf. Conditional Random Fields, CRF, like in Huang (2006), and Gaussian Random Field, GRF), i.e. a space of variation where there is interaction between adjacent points driven by stochastic processes that do not affect, if not indirectly, points far apart in the space. This technique would bias the results as to penalize clusters which are not adjacent in space (like it is the case, in Veneto, of the dialects of Valsugana and the dialects of Bassa Padovana and Fratta Polesine), a problem which we had to face using K-Means. On the other hand, these techniques introduce another bias, i.e. they can give more weight to local contact, which in this area plays a major role because of the prestige of the dialect of Venice. As a consequence, these methods are more likely to retrieve recent phenomena rather than genetic affiliations.

3. **Directionality of change**. Another property that was not encoded in our model was the directionality of change, which is another kind of information that requires arbitrary judgment. However, it is not clear how directionality would solve the problems we have faced in this case, where technically the case of the Central dialects of Veneto was not different from a case in which different migrations correspond to different sound changes (like it is the case with many classical sound correspondences in Indo-European, as clearly shown in Ringe et al. 2002). Perhaps one should think about directionality as an additional output of the algorithms (e.g. applying techniques such as Random Fields or Hierarchical Clustering) rather than a requirement for clustering techniques to perform better. Once the clustering problem is solved, it should be straightforward to apply cladistics techniques or hierarchical clustering models to retrieve the phylogenetic history of the area.

To conclude, this means that the peculiarity of dialect variation cannot be simply addressed by general machine learning techniques, but requires some *ad hoc* methods that make use of more detailed information. Given the results presented in this paper, encoding this kind of information looks necessary to avoid problems such as contact and parallel development, which turned out to vastly characterize diachronic change, especially at the local level.

References

[1] Bartoli M., Introduzione alla neolinguistica. (Vol. 12)., LS Olschki, 1925.

[2] Bartoli M., Saggi di linguistica spaziale, V. Bona, 1945.

[3] Bauer R., Le projet Vivaldi: présentation d'un atlas linguistique parlant virtuel, Anuario del Seminario de Filología Vasca 'Julio de Urquijo', 2010.

[4] Ferguson R., A linguistic history of Venice, Biblioteca dell'«Archivum Romanicum» - Serie II: Linguistica, vol. 57 2007.

[5] Huang R., Automatic Dialect Classification: Advances for Read and Spontaneous Speech, and Printed Text, Doctoral Dissertation, University of Colorado, 2006.

[6] Kondrak G., Algorithms for language reconstruction. PhD dissertation, Toronto: University of Toronto, 2002.

[7] Jaberg K. and Jud J., AIS: atlante linguistico ed etnografico dell'Italia e della Svizzera meridionale, Unicopli, 1987.

[8] Lei Y. and Hansen J. H. L., Dialect Classification via Text-Independent Training and Testing for Arabic, Spanish and Chinese, IEEE Transactions on audio, speech, and language processing Vol. 91, n.1., 2011.

[9] PLEDS,
Penn Price Linguistics Lab Etymological Database System, https://n411.fmphost.com/fmi/webdPLEDS_ONLINE, 2017.

[10] Ringe D., Warnow T. and Taylor A. Indo-European and Computational Cladistics, Transactions of the Philological Society 100 (1) 59-129, 2002.

[11] Zamboni A., Veneto, Pacini editore, 1974.

[12] Zamboni A., Le caratteristiche essenziali dei dialetti veneti, in Manlio Cortelazzo (a cura di), Guida ai dialetti veneti, Padova, CLEUP, 1979.