Empiricism and sociolinguistic cladistics: analysis of a Salentinian idiolects network

Antonella Gaillard-Corvaglia

Sorbonne Nouvelle LPP-UMR 7018 / INALCO / LaLIC- Paris La Sorbonne

antoc75@gmail.com

Abstract

This study is an attempt to classify Italo-Romance Southern dialects using cladistics, a method developed around 1950 by the German entomologist Hennig. A description of the methodology will be followed by the application of cladistic analysis to data collected in Salento. Phonological variables are identified according to diachronic criteria in order to proceed with cladistic analysis. The results are presented as an unrooted cladogram, providing insights into sociolinguistic patterns and social networks. Instead of attempting to capture the global resemblance between languages, our research prefers to focus on a cladistic approach, which has been previously applied to other dialects and data.

1. Introduction

Cladistics (from Greek *kládos* ‘branch’), also called phylogenetic systematics, is a classificatory method that emerged around 1950 when the German entomologist W. Hennig contrived a classificatory method for genetic analysis of living species based on typological clues ordered by derivational chains [1], [2]. Phylogenetic construction is based on the principle of "descent with modification": the characters observed in two or more species that indicate a close relationship are those inherited from their common ancestor [3]. Attempts to reconstruct the evolution of language have been proposed since the middle of the 20th century. One of these approaches, *Numerical Taxonomy*, consists of estimating the linguistic distance between pairs of languages, and calculating evolutionary trees or networks to produce linguistic classifications. This approach is generally used in dialectometry [4], [5], [6]. The cladistic approach, inherited from 19th century linguists, is more recent and uses various methodologies [7], [8], [9], [10], [11]. For proponents of cladism such as Hennig, merely the fact of sharing bundles of derived (or apomorphic) characters is the sign of a close relationship. But our approach here is closer to typological sociolinguistics than to genetic linguistics. We adopt a strategy enabling us to integrate linguistic hypotheses before making inferences on the evolution of linguistic traits and languages, and potentially to refute them. To test the heuristic value of this methodology, we apply cladistics to dialectal data from different sources, through an original coding of philological derivations [12], [13] and [14].

On the one hand, the results obtained in our Salentinian study lend themselves to a network sociolinguistic analysis (according to the variability theory of Labov [15]). On the other hand, they provide a quantitative taxonomy based on concepts other than “cumulative distance, dissonance, and isogloss”. These qualitative terms are replaced, in sociolinguistic cladistics, by a more explicit geometric representation of sociolinguistic norms in the social space, as shown by the taxinomic trees our cladistic software generate. Nevertheless, our survey is still at an experimental stage and should not be considered definitive.

2. Methods

The starting point of our study was a questionnaire constructed following the conventional criteria of Southern Italo-Romance dialectology and sociolinguistics [16] and [17] taking into account several phonological consonantal variables which are specific of the southern Salentinian dialects [18]. The survey was undertaken by the author, a native Southern Salentinian dialect speaker, with the collaboration of town councils. The Southern Salento social background is characterized by a network of small and densely populated towns of around 12,000 inhabitants each [19]. The towns surveyed are the following: Ugento (UGE), Ruffano (RUF), Acquarica del Capo (ACQ), Morciano di Leuca (MOR), Tiggiano (TIG), Cavallino (CAV). Cavallino was chosen as a northern point for the purposes of geolinguistic comparison.

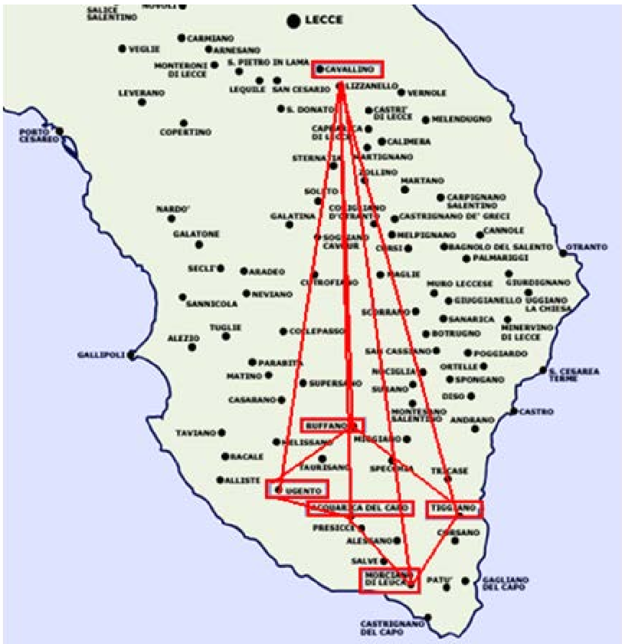

Figure 1: *Map of Salento, sub-region of Puglia, southern Italy. The selected points form a pyramid with the top point (CAV) which is geographically closer to Lecce, capital of Salento.*

2.1. The dialectological survey

2.1.1. *Speakers*

In all, 64 native speakers were surveyed: 6 men and 6 women in each town, categorized according to age (<30, 31-50 and >50), sex, profession, and education level. All informants were bilingual speakers of Italian and Salentinian. The stimuli were presented in Italian, but interactions with the interviewer were mostly in Salentinian.

2.1.2. *Corpus*

The questionnaire consisted of 314 phonological items including 95 consonantal variables from which we selected 35 for cladistic processing, as follows:
A. = Stops [α voiced]
{gatto, grande, fegato, litigare {dente, ditale, cado, credo}}
B. = Sibilants /s/ retraction + /str-/ reduction > Σ
{maestra, finestra, minestra {vostra, mostrare}}
C. = Laterals: retroflection, gemination and rhotacization
{gallo, quello, bello {capelli, cavalli, quelli}}
D. = Palatal laterals (palatalization of -lj-) [- lateral]
{figlio, famiglia, moglie, voglia}
E. = Palatal affricates
{giovedì, fuggire, gelo {ceci, cenere}}
F. = Voiced labial stops, with [O] [+continuant] [+tense] spirantization, gemination)
{bocca, braccio, basso, febbre, tavola {bere, battere}}
G. = [+ tense] [+ palatal] [- continuant, labial] [α voiced]
{vedo, vieni {vento, vomito}} (e.g., Lat. *vomitare* > Sal. [ˈommiku]/[ˈvomitu]/[ˈvommuku]/[vummaˈkare]).

2.1.3. *Procedure*

Data was collected in the following way. Each item on the question list was read in Italian once or twice to the informant, who was asked to translate and repeat the expected form in the Salentinian dialect, twice in isolation and once in a spontaneous sentence. The interviews were recorded with a SONY ECM-MS907 microphone and a SONY MZ-N710 minidisc recorder. The phonetic transcription was performed using SoundForge 7 and Praat to check systematically the auditory impressions of the transcriber.

2.2. Diachronic analysis

This dialectological data allowed us to draw 38 diachronic trees based on each word’s Latin etymon. These trees were constructed following principles of general linguistic. The procedure recalls the descriptive method used by G.I. Ascoli [20], based on the systematic comparison of vowels and consonants of Latin with those of the Romance languages. Such a method has much in common with the genealogical method introduced by Schleicher [21]; it could be applied to one or many languages from to the same geographic area and has the advantage of characterizing a dialect or a whole area through the presence or absence of distinctive features.
As a result, we obtained 38 trees based on three principles:
Pr. 1. *Principle of areologic continuity*: implies a gradual theory of linguistic change whose stages can be reconstructed on the basis of areal indications.
Pr. 2. *Principle of CVCV* [22]: clusters in the diachronic trees are based on a CVCV analysis. Contoïds and vocoïds are embedded in a hierarchy of phonological processes within onset-nucleus templates. Taxonomic clustering relies therefore on constituent interactions rather than on bundles of isoglosses.
Pr. 3. *Principle of parsimony*: claims that the diasystem (in the sense of [23]) evolves in a relatively small number of steps, rather than relying on complex chains of phonetic laws. No more than two or three branches for the explanation of a variable’s evolution.

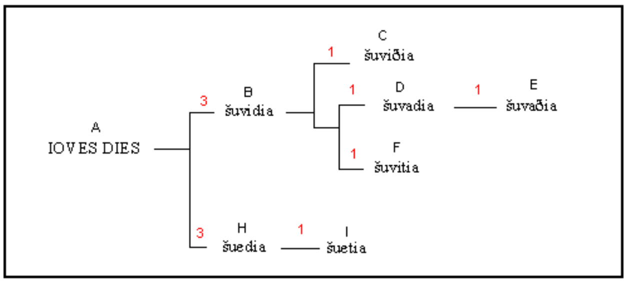

Figure 2: *Diachronic tree of *IOVES DIES.*

In the tree represented in Figure 2, the capital letters represent the different evolutionary states of the Latin variable, while the numbers in red indicate the weights applied according to conditions of phonological markedness [24] in terms of frequency and articulatory difficulty (Index Weighting range is fixed from 1 to 5 points). Examples are given in Table 1.

PHONOLOGICAL PROCESS	INDEX WEIGHTING
Palatal delateralization (*mulierem* > [muǵǵere])	5
Retroflexion + unvoicing (*gallum* > [kaɖɖu])	4
Retroflection (*caballi* > [kavaɖɖi])	3
Other process (lenition, gemination, palatalization..)	2 and 1

Table 1: *Index Weighting for Salentinian variables.*

In this way the data was indexed, weighted and directed to be processed with PAUP 4.0 (Phylogenetic Analysis Using Parsimony [25]) in order to generate cladograms such as the one shown in Figure 4.

2.3. Cladistic analysis

The main steps required for cladistic analysis are the following:
(i) Construction of diachronic trees
(ii) Coding of evolutionary states
(iii) Weighting states
(iv) Factorization
(v) Construction of encoded digital matrices
(vi) Phylogenetic analysis with PAUP 4.0.
Indeed, to complete a cladistic analysis of the data, one must use an encoded expression of diachronic trees, integrating a range of linguistic assumptions (e.g., the three principles enumerated in § 1.2). Encoding is performed a) by associating each variant with a letter indicating its position in the derivation tree, and b) by presenting them in a factored form in which for each encoding letter, each variant takes a binary value of 0 or 1, depending on its place in the derivation tree (software Factor [26]).

W	3111131
A	0000000
B	1000000
C	1100000
D	1010000
E	1011000
F	1000100
H	0000010
I	0000011

Figure 3: *A factorized matrix.*

Thus, for the tree IOVES DIES ‘Thursday’, It. ‘giovedì’ (Figure 2), the Latin form is encoded <A:>, the relationship between <A> and <B> clusters as <A:B>, and the relation between <A> and <C> clusters as <A:C> etc. In its factored form, the pattern <A> is encoded by the vector [00000000000000000], the pattern <B> by the vector [10000000000000000], 1 in first position representing the transformation of <A> into <B>. The pattern <C> is represented by the vector [01000000000000000], where 1 represents the transformation of A into C, and so forth.

The weighting vector W (corresponding to the red numbers in Figure 2) is eventually incorporated into the factorized matrix (Figure 3).

The files are created and edited with the software Factor, then imported into PAUP 4.0 in order to find the most parsimonious tree.

2.4. Results

2.4.1. The cladogram

The automized dialectal data treatment yielded the results represented in Figure 4.

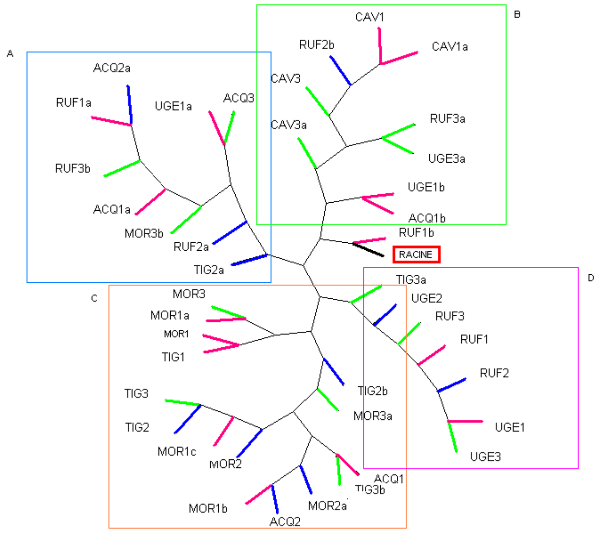

Figure 4: *Salentinian cladogram. Colors and numbers indicate the three age groups (pink-1 for the old-, green-3 for the young, blue-2 for the middle-aged). Lowercase letters indicate idiolectal duplicates.*

2.4.2. Explanation of the cladogram

The cladogram in Figure 4 shows the sociolinguistic shape of the region of Salento according to our cladistic analysis. Colors at the end of the branches indicate different generations (1, 2, 3); lowercase letters indicate idiolectal variants of each variety. For the word *giovedì* for example, in variety MOR and generation 1 (the oldest) we have 4 idiolectal variants: MOR1, MOR1a, MOR1b, MOR1c, corresponding to the heterogeneous answers given by the informants. The sociolinguistic interpretation of the cladogram must be made according to the presence/absence of sinapomorphy (stemming characters labeled by acronyms) in each micro-diasystem. We have a clad (D) of the type Ugento-Ruffanese that groups all main variants (UGE 1-2-3 and RUF 1-2-3). This branch consists of the three generations of varieties geographically in the center of the studied area. It is adjacent to the branch of the central-northern type represented by B, the only type that includes the varieties of Cavallino accompanied by an idiolectal synapomorphy (RUF3a, UGE3a, UGE1b, ACQ1b, RUF2b). In the diasystem A sinapomorphy is perfectly balanced across generations (3 for each generation), with a dialectal characterization of type Acquarica-Ruffanese. Finally, the clad C is a little richer in sinapomorphy, with 15 variants, and represents a more important heterogeneity in comparison with the three other clads. Here we see a sort of *crescendo* of sinapomorphy: 4 for the young generation, 5 for the middle generation, and 6 for the older one. We also can see that sinapomorphy in green (3) belongs to the varieties MOR and TIG (that is to say Capo di Leuca, the most southern area); they are linked to the group in blue (2) of the same varieties, as well as ACQ and still to the same elements of group 1. As we see it, this clad belongs entirely to the extreme-southern type (MOR-TIG-ACQ) which is generationally and geolinguistically homogenous. The three generations of this cluster use the dialect in the same way, with no differentiation of age, sex, or socio-cultural level.

By proceeding in this manner, and working with a larger corpus, it would be possible to generate cladograms and improve the granularity of descriptions of the areal configurations of any dialect.

3. Conclusions

The Salento region’s sociolinguistic variation does not seem to depend on age range as the three generational pools are often clustered. This can be explained by a high degree of cross-generational interactivity within the network as a trend to dialect norm synchronization, contrary to official findings which declare a diminution of the dialect’s exclusive use and claim that the use of dialect is proportional to age (ISTAT data 2006 and 2007). This assumption is also disconfirmed by our surveys made in Salento [27] where 77, 7% of speakers of every generation declared to use dialect every day. In other words, the generational distribution of sociolinguistic markers shows up as entirely homogenous and symmetrical in the Salento cladogram. The use of dialect, therefore, is not linked to age, a conclusion that agrees with the well known vitality of Southern Italian (and Salentinian) dialects. At the same time, these results imply that socio-cultural differences between speakers (differentiated by educational level, profession, age and sex, see § 2.1.1.) tend to be reduced by regular use of (and strong proficiency in) the local dialect. The Italianization of speakers’ mother tongue has probably not yet played a very important role in this part of Salento, as the dialect remains the language of daily interaction and sociability. It is currently used within the family and between friends, as is typical for many conditions of dialect bilingualism in Southern Italo-

Romance areas, characterized by a mild diglossy (less conflictual than in cases of overt centralization and national assimilation, such as in France). Moreover, our variationist analysis through cladistics confirms the geographical variation of local norms (north, central, and south) in distinction to the sociologically unified character of these dialects.

Cladistics does not capture the global resemblance between languages (unlike dialectometry), and does not establish regular correspondences between dialects. Rather, cladistics rather aims at pointing out the degree of relationships or the structural convergence between several languages or dialects. In the case of Salento, cladistics allows us to analyse the granularity of a small sociolectal network by crossing different sociolinguistic factors. We shall complete this first attempt with more sociolinguistic data. Nevertheless, our preliminary study provides insights into the application of cladistics and will be extended to other dialectological and typological surveys of phonological variation in linguistic families and in dialect networks.

4. References

[1] Hennig W. 1953. Kritische Bemerkungen zum phylogenetischen System der Insekten. *Entomologie* 3. 1-85.

[2] Hennig W. 1988. *Systématique cladistique.* Paris: Société Française de Systématique. Darlu P. and P. Tassy. 1993. *La reconstruction phylogénétique: concepts et méthodes.* Paris: Masson.

[3] Ben Hamed M. 2005. Neighbour-nets portray the Chinese dialect continuum and the linguistic legacy of China's demic history. *Proceedings of the Royal Society of London: Biological Sciences* 272. 1015–1022.

[4] Goebl H. 1981. Éléments d'analyse dialectométrique (avec application à l'AIS). *Revue de Linguistique Romane* 45. 349-420.

[5] Goebl H. 1987. Points chauds de l'analyse dialectométrique: pondération et visualisation. *Revue de Linguistique Romane* 51. 63-118.

[6] Ben Hamed M., P. Darlu and N. Vallée. 2005. On cladistic reconstruction of linguistic trees through vocalic data. *Journal of Quantitative Linguistics* 12 (1). 79-109.

[7] Hoenigswald M.H. and F.L. Wiener. 1987. *Biological Metaphor and Cladistic Classification: An Interdisciplinary Perspective.* Philadelphia: University of Pennsylvania Press.

[8] Holden C.J. 2001. Bantu language trees reflect the spread of farming across sub-saharian Africa: a maximum parsimony analysis. *Proceeding of the Royal Society of London: Biological Sciences* 269. 793-799.

[9] Nakhleh L., D. Ringe and T. Warnow. 2005. Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages. *Language* 11 (2). 382-420.

[10] Wang W. S. Y. 1987. Representing languages relationships. In Hoenigswald, M.H. and F.L. Wiener (eds.), *Biological Metaphor and Cladistic Classification: an Interdisciplinary Perspective.* Philadelphia: University of Pennsylvania Press.

[11] Dell'Aquila V., A. Gaillard-Corvaglia and J. L. Léonard. 2011. Le mazatec (langue otomangue, Mx SE) : cartographie, géolinguistique et typologie linguistique. *Pays et Paysages Linguistiques* (January 28-29, 2011). Paris: France.

[12] Gaillard-Corvaglia A., J. L. Léonard and P. Darlu. 2007. Testing cladistics on dialect networks and phyla (Gallo-Romance and southern Italo-Romance). *Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology* (June 23–30, 2007). Prague: Czech Republic.

[13] Gaillard-Corvaglia A., J. L. Léonard and P. Darlu. 2008. Analyse cladistique des vocalismes d'Oïl à partir des données de l'ALF. *Bollettino dell'Atlante Linguistico Italiano* 32 (III serie). 55-96.

[14] Labov W. 2001. *Principles of linguistic change. Vol.1.* Oxford: Blackwell.

[15] Grassi C., A. Sobrero and T. Telmon. 1999. L'uso del dialetto in Italia: aspetti sociali e pragmatici. In (id.), *Fondamenti di dialettologia italiana*. Bari: Laterza. 161-269

[16] Sobrero A. and I. Tempesta. 2002. Puglia. *Profili linguistici delle regioni*. Bari: Laterza.

[17] Grassi C., A. Sobrero and T. Telmon. 1999. I dialetti in Italia. In (id.), *Fondamenti di dialettologia italiana.* Bari: Laterza. 91-118.

[18] Ruzzo M. (ed). 2003. *Puglia in cifre 2002*. Bari: Progredit.

[19] Ascoli G. I. 1882-1885. L'Italia Dialettale. *Archivio Glottologico Italiano* 8. 98-128.

[20] Schleicher W.A. 1871. *Compendium der Vergleichenden Grammatik der Indogermanischen Sprachen*. Weimar: Hermann Böhlau.

[21] Lowenstamm J. 1996. CV as the only syllable type. In Durand J. and B. Laks (eds.), *Current trends in phonology: models and methods. Vol. 2.* Salford: ESRI, University of Salford. 419-441.

[22] Weinreich U. 1954. Is a structural dialectology possible? *Word* 10. 388-400.

[23] Calabrese A. 1995. A constraint-based theory of phonological markedness and simplification procedures. *Linguistic Inquiry* 26 (3). 373-463.

[24] Swofford D.L. 2002. *PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods).* Version 4.0. Sinauer Associates. Sunderland. Massachussets.

[25] Felsenstein J. 2004. *PHYLIP. Phylogeny Inference Package*. Version 3.6b. Department of Genome Sciences, University of Washington. Seattle. Washington.

[26] Gaillard-Corvaglia A. 2006. Lo spazio dialettale salentino: uno studio sociolinguistico. In Pettorino M., A. Giannini, M. Vallone and R. Savy (eds). *Atti del convegno internazionale "La Comunicazione Parlata"* (February 23-25, 2006). Napoli: Liguori. 1169-1178.