1 La metodologia impiegata

1.0 Introduzione

In questo capitolo si illustrerà la metodologia impiegata per stabilire le distanze strutturali reciproche tra i dialetti del sardo e individuare quindi le varietà della lingua. Per questa ricerca abbiamo scelto di determinare la distanza strutturale tra 77 dialetti sardi. La nostra ricerca si limita ad investigare le conseguenze fonologiche, morfologiche e lessicali delle differenze tra queste lingue, visto che queste sono le uniche differenze che si possono misurare sulla base della distanza fonetica. La determinazione quantitativa delle distanze tra strutture sintattiche è ancora al di fuori della portata delle attuali tecniche di ricerca linguistica.

I processi diacronici che hanno portato alla differenziazione dei dialetti sardi costringono a porsi delle domande come le seguenti: "Quante parole di una lingua sono coinvolte dei processi di mutamento linguistico? Quanto è cambiata la struttura di queste parole? Quanto è grande la distanza strutturale tra le varietà di questa lingua che risultano da questi processi?" Per poter rispondere a queste domande si dovrebbero, in linea di principio, analizzare tutte le parole in tutte le varietà in questione, e per di più con tutte le loro possibili derivazioni morfologiche. Necessariamente, deve aver luogo una selezione preventiva dei dati. Una simile selezione, però, costringe a determinare *a priori* e in modo soggettivo ciò che è rilevante per la ricerca, a meno di non impiegare un metodo statisticamente giustificato (Kessler 2001).

Un'altra fonte di soggettività da evitare è costituita dal fatto che il calcolo della distanza strutturale tra due lingue (o due varietà della stessa lingua) è praticamente impossibile da effettuare senza ricorrere a tecniche computazionali (Nerbonne & Heeringa 1998). Ogni parola selezionata in ogni lingua (o fase della lingua) deve essere confrontata con ogni parola corrispondente nell'altra lingua (o fase) per poter stabilire le rispettive distanze fonologiche. Senza un approccio computazionale, l'esecuzione di questi calcoli costituisce un'impresa irrealizzabile. Attraverso la tecnica stastistica della campionatura randomizzata dei dati (*random sampling*) si può però effettuare una selezione oggettiva dei dati. Ricorrendo alle tecniche sviluppate all'interno della *Dialettologia Computazionale* (Kessler 1995; Nerbonne & Heeringa 1998), si può fra l'altro determinare quantitativamente il mutamento di una lingua, come conseguenza di evoluzioni fonologiche e/o morfologiche, o del contatto con altre lingue.

1.1 La selezione delle parole e delle varietà linguistiche

La distanza strutturale fra le varietà che si sono comparate è stata determinata sulla base della *Distanza Levenshtein* (si veda § 1.4) di 200 parole selezionate *at random* e tradotte nelle diverse varietà linguistiche. Le parole in questione provengono da un *corpus* di 257.000 parole che compongono una serie di testi scritti in diverse varietà del sardo contemporaneo. I testi consistono di romanzi, traduzioni, articoli di giornali, presentazioni su Internet, i quali erano disponibili in formato elettronico. Questi testi si possono considerare rappresentativi del sardo scritto moderno. Le 200 parole selezionate sono anche indirettamente rappresentative della frequenza delle parole nel sardo scritto, dato che le parole che più spesso sono presenti in un testo hanno anche maggiori probabilità di essere selezionate.

Mediante l'uso di uno specifico programma informatico, è stata effettuata la selezione randomizzata di 400 parole. Questa prima selezione è stata successivamente ridotta alle prime 200 parole che rispondevano ai seguenti requisiti:

(i) la presenza nel dizionario più comprensivo della lingua sarda (Puddu 2000);

(ii) non costituire una variante grafica o dialettale di una parola già selezionata in precedenza.

Le 200 parole selezionate che costituiscono il campione da comparare sono presentate nell'Appendice 1. Tutte le parole sono state tradotte nelle diverse varietà linguistiche contemporanee, cosa questa che in taluni casi ha comportato l'uso di costruzioni perifrastiche.

In alcuni casi, le parole selezionate sono probabilmente soltanto delle parole grafiche che consistono di un verbo e dei pronomi enclitici che li seguono (per es. *apporrindeli* 'porgendogli/ porgendole'; *apprettandelos* 'assillandoli'). Il loro *status* di parole fonologiche e morfologiche cambia da lingua a lingua, ciononostante abbiamo scelto di considerare queste strutture come delle unità in quanto in ciascuna delle lingue in questione si usano le stesse convenzioni grafiche per la rappresentazione di queste costruzioni. Abbiamo anche scelto di mantenere nel nostro campione forme diverse degli stessi verbi (per es. *tenner/tenes/tentu* 'avere-INF/3 SG/PP'), in modo da far contare nelle misurazioni delle distanze strutturali anche la morfologia verbale.[1]

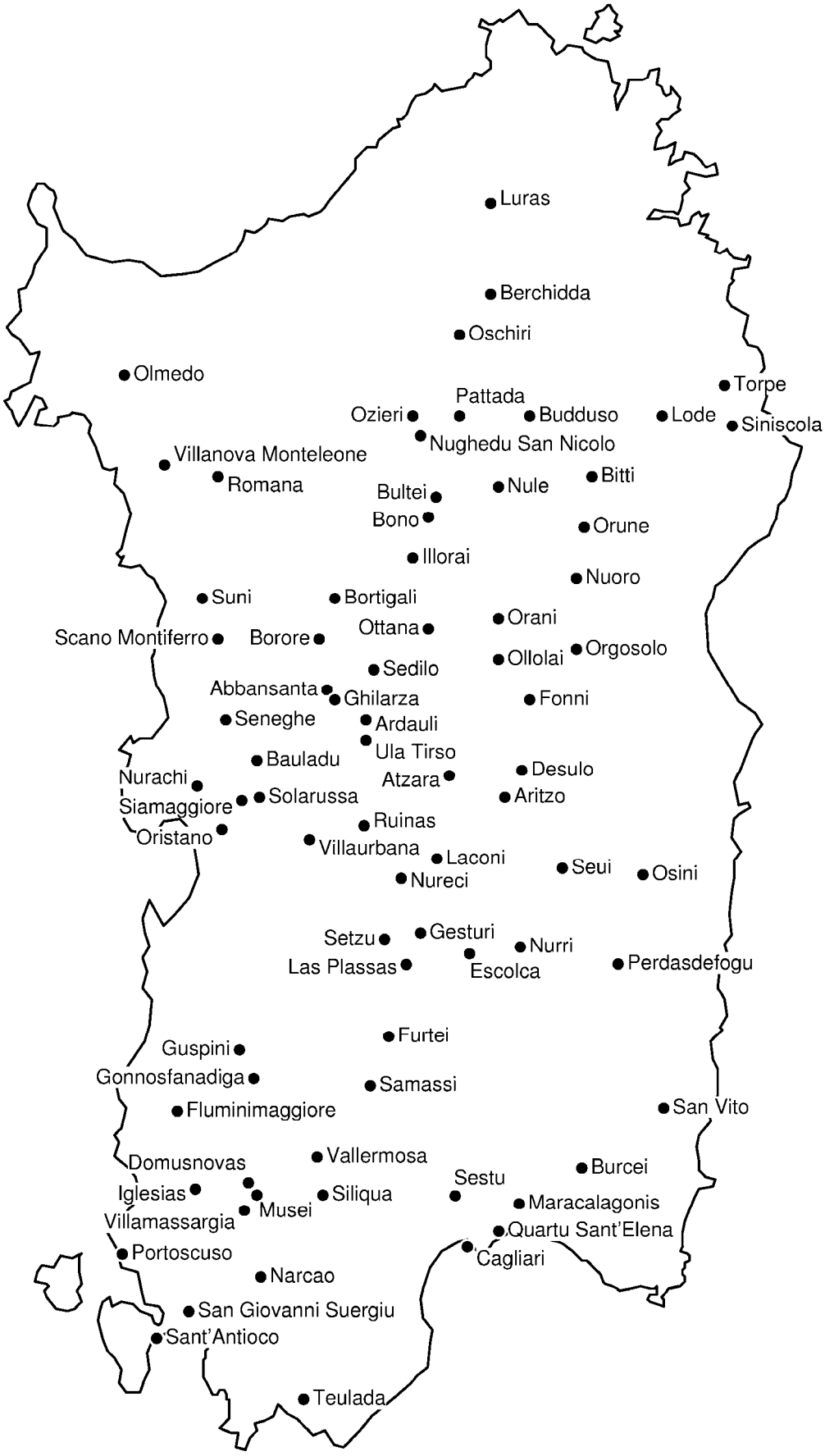
1.2 I dialetti sardi

Le 200 parole selezionate sono state proposte a parlanti di 77 dialetti sardi. Agli informatori è stato chiesto di tradurre e pronunciare le suddette parole nel loro dialetto. Le pronunce attestate sono state trascritte nell'alfabeto fonetico X-Sampa (un sistema in cui i simboli API vengono sostituiti da caratteri ASCII). Eventuali differenze nelle convenzioni grafiche tra le diverse varietà non hanno quindi giocato alcun ruolo.

La scelta dei dialetti è stata in parte dettata dalla necessità di rappresentare le principali varietà del sardo. In parte, invece, è stata la disponibilità di parlanti a determinare la scelta di un dato dialetto locale, anziché un altro. Le diverse varietà sono state individuate sulla base dell'Atlante Dialettologico della Sardegna (Contini 1987), sulla base della descrizione dei dialetti meridionali contenuta in Bolognesi (1998) e sull'analisi quantitativa della variazione dialettale in Sardegna contenuta in Bolognesi & Heeringa (2005). I dialetti della Sardegna centrale sono abbondantemente rappresentati per poter esaminare nel miglior modo possibile la transizione dai dialetti meridionali a quelli centrosettentrionali. La lista dei dialetti esaminati non è certo esaustiva, nemmeno per quanto riguarda le sub-varietà, ma è sufficientemente estesa per il tipo di ricerca condotto in questa sede. Le caratteristiche delle diverse varietà sono ben rappresentate, come anche i contrasti fra esse, e la loro distribuzione geografica. I 77 dialetti vengono presentati nella seguente cartina (Cartina 1), in cui è indicata anche la loro distribuzione geografica

.[1] Questa scelta ha comportato ovviamente una certa ridondanza del nostro campione. I seguenti verbi sono rappresentati diverse volte in forme diverse: *andai* 'andare (3 volte), *ponni* 'mettere' (3 volte); *tenni* 'avere (3 volte), *domandai* 'chiedere, domandare' (2 volte), *ai* 'avere-AUX' (2 volte), *fai* 'fare' (2 volte), *perdi* 'perdere' (2 volte), *essi* 'essere' (2 volte), *biri* 'vedere' (2 volte). 7 dei 9 verbi sono però irregolari e mostrano una grande variabilità che porta a risultati molto diversi nelle varie lingue.

Cartina 1


Luras
Berchidda
Oschiri
Olmedo
Torpe
Pattada
Ozieri
Budduso
Lode
Siniscola
Nughedu San Nicolo
Villanova Monteleone
Romana
Bultei
Nule
Bitti
Bono
Orune
Illorai
Nuoro
Suni
Bortigali
Orani
Ottana
Scano Montiferro
Borore
Ollolai
Orgosolo
Sedilo
Abbansanta
Ghilarza
Fonni
Seneghe
Ardauli
Ula Tirso
Bauladu
Desulo
Nurachi
Atzara
Aritzo
Siamaggiore
Solarussa
Ruinas
Oristano
Villaurbana
Laconi
Seui
Osini
Nureci
Setzu
Gesturi
Nurri
Las Plassas
Escolca
Perdasdefogu
Furtei
Guspini
Gonnosfanadiga
Samassi
Fluminimaggiore
San Vito
Vallermosa
Domusnovas
Burcei
Sestu
Iglesias
Siliqua
Maracalagonis
Musei
Villamassargia
Quartu Sant'Elena
Portoscuso
Cagliari
Narcao
San Giovanni Suergiu
Sant'Antioco
Teulada

Agli informatori è stato chiesto di fornire per ciascuna parola tutte le forme della cui presenza e uso nella loro comunità dialettale essi fossero a conoscenza. Se in un dialetto era presente più di una forma di una parola della lista (per es. la parola sarda originaria e il corrispondente prestito dall'italiano), oppure se una data parola corrisponde a più parole o a più significati, sono state indicate tutte le forme presenti.

Il confronto fra più forme della stessa parola ha luogo nel modo seguente. Supponiamo che si voglia determinare per una certa parola la distanza fra due dialetti, e che in uno o in entrambi di questi dialetti siano state fornite due o più forme. In questo caso vengono appaiate le forme dell'uno e dell'altro dialetto in modo tale che la media delle distanze fra le coppie di parole sia minima. Perciò se in un dialetto si trovano le forme A e B di una parola, e in un altro dialetto la forma C, la distanza tra i due dialetti sarà data dalla media delle distanze tra A e C e B e C.

1.3 La misurazione delle distanze fonologiche tra lingue

Esistono diversi algoritmi per misurare le distanze tra dialetti sulla base di trascrizioni. In Hoppenbrowers & Hoppenbrowers (2001) viene presentato il Metodo della Frequenza dei Tratti (MFT), e la sua applicazione ai dialetti olandesi. Per ogni dialetto si determinano le frequenze dei tratti fonologici distintivi presenti in un una determinata trascrizione. L adistanza tra due dialetti viene determinata tramite il confronto tra le frequenza dei tratti.

Un approccio diverso è stato applicato ai dialetti irlandesi da Kessler (1995). I dialetti vengono paragonati tra di loro misurando la distanza tra parole corrispondenti tramite l'algoritmo di Levenshtein. Una descrizione delle applicazioni di questo algoritmo sui dialetti olandesi si trova in Nerbonne & Heeringa (1998), mentre l'applicazione a vari dialetti sardi è presentata in Bolognesi & Heeringa (2005).

Per questa ricerca si è fatto ricorso all'approccio di Levenshtein. Questo approccio presenta due vantaggi sul primo, per la precisione, il fatto che una parola viene trattata come un'unità linguistica, e l'altro fatto che l'algoritmo tine conto dell'ordine lineare dei segmenti che compongono una parola. Qui sotto viene descritto l'uso della Distanza-Levenshtein (§ 1.4), e i miglioramenti apportati all'algoritmo tramite l'introduzione delle distanza graduali tra segmenti (§ 1.5).

1.4 La *Distanza-Levenshtein*

Attraverso la *Distanza-Levenshtein*, le lingue vengono comparate mediante la comparazione di una parola di una lingua con la parola corrispondente in un'altra lingua. La comparazione si effettua trovando il modo più semplice per trasformare una data parola in un'altra attraverso l'inserzione di suoni, la loro cancellazione o la loro sostituzione. Nella forma più semplice dell'algoritmo tutte le operazione menzionate hanno lo stesso costo, per esempio 1.

Supponiamo che la parola *usare/impiegare* in un dialetto sardo sia pronunciata *impr↔ar↔*, mentre in un altro dialetto sia pronunciata *imp↔rai*. Il passaggio da una variante alle altre si effettua nel modo seguente:

[impr↔ar↔]	cancella [r]	1	
[impr↔a↔]	sostituisci [↔] con [ɪ]		1
[impr↔ai]	cancella [r]	1	
[imp↔ai]	inserisci [r]	1	
[imp↔rai]			
		4	

Per determinare questa distanza attraverso l'algoritmo di Levenshtein, le parole vengono allineate una sotto l'altra, in modo da poter stabilire quali segmenti di una parola corrispondono ai segmenti di un'altra. Il risultato viene chiamato *Allineamento*. La forza dell'algoritmo di Levenshtein consiste nel fatto che questo trova sempre quella specifica distanza che è calcolata sulla base di un allineamento in cui la corrispondenza tra segmenti è scelta in modo tale che il costo dell'operazione risulta minimo. Nel nostro esempio l'allineamento si presenta nel modo seguente:

i	m	p	ϕ	r	↔	a	r	↔
i	m	p	↔	r	ϕ	a	ϕ	i

0 + 0 + 0 + 1 + 0 + 1 0 + 1 + 1 = 4

Confrontando in questo modo due parole, la distanza tra parole più lunghe sarà mediamente maggiore di quella tra parole più brevi. Più lunga è la parola, maggiore è la probabilità che esistano differenze rispetto alla parola corrispondente in un altro dialetto. Poiché questo contrasta con l'idea che le parole costituiscano delle unità linguistiche, indipendentemente dal numero di elementi che le compongono, la *Distanza Levenshtein* viene divisa per la lunghezza dell'allineamento (la lunghezza elaborata delle parole). Come si vede la lunghezza dell'allineamento è uguale a 9 unità. La distanza strutturale fra le parole è adesso perciò uguale a 4/9 = 0.44. Spesso sono possibili più allineamenti che, oltre a comportare le stesse lunghezze, comportano anche un costo uguale per le operazioni impiegate. In tal caso si divide la distanza per l'allineamento più lungo, dato che questo comporta sempre il maggior numero di abbinamenti. Si parte anche dal presupposto che l'allineamento più lungo costituisca la miglior approssimazione del modo in cui gli umani percepiscono la differenza tra due parole.

Una volta stabilita la lunghezza dell'allineamento più lungo, diventa anche possibile esprimere la distanza tra due parole in termini percentuali. In tal caso la somma dei costi delle operazioni eseguite va divisa per il prodotto della lunghezza dell'allineamento più lungo moltiplicato per il costo più alto possibile, moltiplicando poi il quoziente che ne risulta per 100. In questo esempio, tutti i costi hanno un valore uguale a 1. Espressa in percentuale, la distanza è adesso uguale a [4/ (9*1)]*100 = 44%.

Poiché il confronto fra varietà linguistiche diverse avviene sulla base di 200 parole, dai confronti fra due lingue si ottengono 200 *Distanze Levenshtein* espresse in percentuali. La distanza espressa in percentuale tra due varietà è quindi uguale alla media delle 200 *Distanze-Levenshtein* espresse in percentuale, e si calcola dividendo la somma delle 200 *Distanze-Levenshtein* espresse in percentuale per 200. Si può vedere che applicando la *Distanza-Levenshtein* non solo si tiene conto dei confini di parola, ma si prende in considerazione

anche l'ordine lineare dei suoni di una parola. Questo approccio è stato utilizzato in tutto il resto del lavoro

Visto che si confrontano 200 coppie di parole corrispondenti tra di loro in tutte le coppie che si possono formare dalla 60 varietà linguistiche, in totale si calcolano [((60*60)-60)/2] * 200 =354.000 distanze tra parole. È chiaro che effettuare a mano tutti questi calcoli richiederebbe dei tempi enormi. Un approccio quantitativo alla linguistica implica perciò necessariamente l'uso del computer, e per questo viene anche definito approccio *computazionale*. Dato che uno degli scopi di quest'articolo è quello di presentare i vantaggi dell'introduzione della *Distanza Levenshtein* nello studio del contatto linguistico, è necessario anche essere espliciti rispetto ai suoi limiti.

Innanzitutto, il sistema misura le distanze tra parole sulla base delle rappresentazioni segmentali della loro pronuncia. Caratteristiche suprasegmentali come l'intonazione e l'accento vengono sistematicamente tralasciate. Il nostro 'appello' a favore della *Distanza Levenshtein* non va però assolutamente preso come un invito a trascurare quelle differenze linguistiche che non possono essere analizzate in modo soddisfacente sulla base di questo metodo. Per questo tipo di analisi occorre utilizzare altri metodi.

Un secondo limite è costituito dal fatto che occorrono le trascrizioni fonetiche delle pronunce delle stesse parole in molte località diverse. Il fatto che il sistema possa elaborare una gran mole di dati costituisce naturalmente un grosso vantaggio, ma una gran mole di dati è anche necessaria per poter raggiungere dei buoni risultati.

1.5 Distanze graduali tra suoni

Quando si confrontano le lingue sulla base di trascrizioni effettuate mediante simboli fonetici non si tiene conto del fatto che certi suoni sono molto simili e altri molto diversi tra di loro. Per esempio i suoni che compongono la coppia [b,p] sono molto più simili di quelli che compongono la coppia [a,p]. Inoltre, nei confronti basati sui simboli fonetici non si tiene conto dei segni diacritici. Confrontando per esempio una [a] con una [a~], diventa molto difficile stabilire *quanto* i due suoni differiscano. In questi casi occorre operare una scelta drastica: considerare i due suoni come completamente uguali, oppure considerarli come completamente diversi. Dato che le similitudini tra suoni che sono distinti solo da segni diacritici sono sempre maggiori delle dissimilitudini, in precedenti lavori si era scelto di ignorare queste ultime. Una [a] e una [a~] venivano quindi considerate identiche. Per di più, con questo sistema è impossibile esprimere il fatto che, per esempio, se un'epentesi consiste dell'inserzione di una vocale bassa, questa debba pesare molto di più che non l'inserzione di un quasi inaudibile colpo di glottide.

Tali problemi si possono risolvere rappresentando ciascun suono come una serie di caratteristiche distintive e sostituendo il simbolo fonetico con una matrice (*feature matrix*) che contiene le varie caratteristiche distintive. Ciascun tratto distintivo si può considerare come una caratteristica fonetica (generalmente articolatoria) che può fungere da elemento distintivo e/o classificatorio per tutto il fonema. Una matrice contiene per ciascuna caratteristica distintiva un valore che indica la misura in cui questa proprietà la caratterizza. Rappresentando i suoni per mezzo di tali matrici si può tenere conto anche dei segni diacritici, rappresentando anche a questi per mezzo di caratteristiche distintive e attribuendo ad esse corrispettivi valori. Per esempio, la caratteristica *lunghezza* ha come *default* il valore 0. Se però un suono viene specificato come semilungo, allora gli viene attribuito il valore 1, mentre se il suono è indicato come lungo il valore della *lunghezza* e 2. La distanza può essere

calcolata come la radice quadrata della somma dei quadrati delle differenze fra matrici corrispondenti (*Distanza Euclidea*).

Per poter stabilire anche il costo graduale delle inserzioni e delle cancellazioni di un suono, è necessario definire anche il 'silenzio' in termini di caratteristiche distintive. Dato però che il 'silenzio' consiste appunto dell'assenza di qualunque caratteristica distintiva, la sua introduzione all'interno di questo quadro teorico ne impone una definizione artificiosa.

Inoltre, anche se l'approccio basato delle caratteristiche distintive può condurre a dei risultati soddisfacenti nella misurazione delle distanze strutturali tra lingue, i sistemi di caratteristiche distintive non sono basati su delle misurazioni reali. Le differenze qualitative tra caratteristiche distintive rimangono in fondo intrinsecamente impossibili da misurare.

Questa problemi, ma in particolare quello della definizione del 'silenzio' si possono risolvere ricorrendo al confronto tra gli spettrogrammi dei suoni. Il 'silenzio' si può perciò definire come assenza dell'intensità per tutte le frequenze di tutti gli spettri di un suono.

Durante il processo di acquisizione del linguaggio, i bambini non hanno bisogno di apprendere di apprendere esplicitamente le caratteristiche articolatorie dei suoni che gradualmente imparano a produrre. Il segnale acustico del parlato contiene tutte le informazioni necessarie ai bambini per imparare a padroneggiare il sistema fonologico della lingua alla quale sono esposti. Il segnale acustico contiene perciò anche informazioni sufficienti sulle caratteristiche articolatorie usate normalmente per descrivere i suoni del parlato nella letteratura fonetica e fonologica.

Uno spettrogramma costituisce la rappresentazione visiva del segnale acustico di un suono. Così come il segnale acustico è sufficiente a distinguere un dato suono da qualunque altro suono prodotto in circostanze simili, lo spettrogramma di un suono costituisce una rappresentazione unica e non confondibile con quelle di altri suoni. Le differenze visive tra spettrogrammi rispecchiano le distanze acustiche tra suoni.

In questa ricerca si è fatto uso dei suoni registrati da John Wells e Jill House nella cassetta *The Sounds of the International Phonetic Alphabet*, pubblicata nel 1995.

In questa registrazione le consonanti sono talvolta precedute e sempre seguite da una [a]. Queste vocali sono sempre state eliminate dagli spettrogrammi. Successivamente, per entrambi i parlanti, è stata stabilita l'altezza media del tono per mezzo del programma *Praat*.[2] L'altezza media del tono è stata stabilita sulla base di un campione contenente 28 vocali concatenate. L'altezza media del tono della voce di John Wells è apparsa uguale a 127.9929 Hertz, mentre quella della voce di Jill House è apparsa uguale a 191.5735 Hertz. Sono stati quindi monotonizzati tutti i campioni di John Wells e Jill House sulle loro rispettive altezze medie di tono.

Successivamente, utilizzando il programma *Praat*, è stato prodotto lo spettrogramma di ciascuno dei suoni pronunciati da entrambi i parlanti. Abbinato a *Praat*, si è scelto anche di filtrare gli spettrogrammi con il *Bark-filter*, il quale costituisce un modello plausibile della percezione umana per via delle seguenti proprietà:

(i) Si fa uso di una scala di frequenza più o meno logaritmica. Di conseguenza si tiene conto del fatto che la distanza fra toni bassi viene percepita come maggiore di quella fra toni alti. Per stabilire la scala di frequenza, in Traunmüller (1990) viene presentata la seguente formula: *Bark* = [(26,81 * *Hertz*)/ (1960 + *Hertz*)] – 0.53.

[2] Questo programma può essere scaricato gratis all'indirizzo: http://www.fon.hum.uva.nl/praat/.

(ii) Nel caso delle ampiezze (le intensità delle frequenze) si utilizzano i loro valori logaritmici. Di conseguenza si tiene conto del fatto che i toni bassi non vengono percepiti come più intensi, malgrado in realtà essi lo siano.

Le altre caratteristiche distintive che è stato possibile introdurre nelle misurazioni grazie all'adozione degli spettrogrammi sono quelle rappresentate dai segni diacritici della nasalità vocalica (per es. [ã]) e dell'apicalità delle fricative [s⇒] e [z⇒]. Non essendo disponibili i campioni relativi, né nella cassetta di John Wells e Jill House, né altrimenti, per poter introdurre queste caratteristiche si è proceduto nel modo seguente: (i) la distanza prodotta dalla nasalità tra una vocale non nasale 1 e un'altra vocale nasale 2 è stata calcolata come media della distanza fra la vocale non nasale 1 e la versione non nasale della vocale 2, e la distanza tra la vocale 1 e la consonante nasale [n]; (ii) la distanza prodotta dall'apicalità bei confronti delle altre consonanti è stata calcolata come media della distanza tra una data consonante e le fricative non apicali (a) sorda ([s]) e (b) sonora ([z]), e tra la stessa consonante e le fricative alveo-palatali (a) sorda ([♣]) e (b) sonora ([ ]).

Per poter esprimere la distanza tra parole in termini percentuali occorre stabilire il valore del costo massimo che risulta dal passaggio da una forma all'altra di una parola (si veda il § 7.3.1). La distanza massima è quella attestata tra lo spettrogramma della vocale [a] e quello del 'silenzio'. Nei calcoli, perciò, si considera la differenza tra [a] e il 'silenzio' come uguale al 100%, per cui le distanza tra tutti gli altri suoni saranno inferiori. Dai risultati raggiunti si è visto che le liquide e le nasali sono molto simili alle vocali. Per poter tenere conto delle combinazioni tra suoni che si verificano all'interno della struttura sillabica è stato necessaria una piccola revisione dell'algoritmo di Levenshtein. L'algoritmo è stato modificato in modo da allineare, in due forme diverse di una parola, le vocali esclusivamente con le vocali e le consonanti esclusivamente con le consonanti. Date le loro caratteristiche intermedie, l'algoritmo tratta però le vocali [i], [u] e schwa sia come vocali che come consonanti, mentre le semivocali [j] en de [w] vengono trattate sia come consonanti che come vocali.

Sono stati integrati nella *Distanza-Levenshtein* anche i seguenti tratti suprasegmentali: extrabreve, semilungo e lungo. Questi valori della lunghezza sono stati integrati adattando le trascrizioni prima delle misurazioni. Nelle trascrizioni i segmenti privi di indicazioni sulla lunghezza vengono raddoppiati, i segmenti semilunghi vengono triplicati e quelli lunghi quadruplicati.

Il confronto tra i risultati ottenuti usando i tratti distintivi e quelli ottenuti con gli spettrogrammi ha mostrato che questi ultimi concordano maggiormente con ciò che è lecito attendersi in base alla distribuzione geografica dei dialetti, da un lato, e dai risultati della dialettologia tradizionali, dall'altro. La scelta di basare le misurazioni sugli spettrogrammi è quindi non solo basata sulla necessità di una metodologia più accurata, ma anche su risultati empiricamente più soddisfacenti.

1.6 Classificazione delle varietà linguistiche

Quando si comparano fra di loro 77 varietà linguistiche, le *Distanze-Levenshtein* possono essere ordinate gerarchicamente in una matrice che consiste di 77 righe e 77 colonne. La tabella è paragonabile a una tabella delle distanze in chilometri tra città. In questo modo si possono mettere in evidenza strutture che altrimenti rimarrebbero nascoste. Si è fatto uso di due diversi metodi di classificazione che si integrano a vicenda: l'analisi mediante *clustering* (§ 1.7; § 2.9) e la *scalatura multidimensionale* (§ 1.8; § 2.8). Il risultato dell'analisi mediante

clustering comporta una suddivisione netta delle varietà linguistiche in gruppi, mentre il risultato della *scalatura multidimensionale* mette bene in evidenza il rapporto tra le diverse varietà, anche quando queste appartengono a gruppi diversi.

1.7 Analisi gerarchica tramite *Clustering*

Il *clustering* è una tecnica di uso corrente nelle discipline storiche, ma viene applicata anche alla psicolinguistica. Lo scopo del *clustering* è quello di identificare raggruppamenti rilevanti all'interno di strutture complesse. Quando un *cluster* (o 'agglomerato') fa parte di un *supercluster* (e questo a sua volta di un *supersupercluster*) si può osservare che esiste un rapporto gerarchico fra *cluster*, e si parla di analisi gerarchica tramite *Clustering*. L'algoritmo si può spiegare più agevolmente usando un esempio. Supponiamo che si abbia la matrice seguente

	Lodé	**Luras**	**Gesturi**	**Iglesias**	**Portoscuso**
Lodé	0	17,997	32,958	33,624	34,311
Luras		0	30,095	31,101	32,963
Gesturi			0	11,843	12,692
Iglesias				0	4,876
Portoscuso					0

In questa matrice le cifre indicano le distanze reciproche tra cinque varietà diverse. Il valore di ciascuna cella (*i,j)* è naturalmente uguale a 0 (la distanza di una lingua da se stessa). Poiché la matrice è simmetrica non occorre rappresentare nuovamente i dati della metà in basso a sinistra della matrice.

Il *clustering* costituisce un processo iterativo. In ogni passaggio del processo si individua la distanza più piccola nella matrice e le lingue tra cui esiste questa distanza vengono riunite in un *cluster*. Successivamente si determina la distanza tra il *cluster* formato e le altre lingue. Ai fini di questa ricerca, l'algoritmo che ha fornito i risultati più soddisfacenti (cioè, più logici) si è rivelato quello che prende in considerazione la media delle distanze. La distanza di *k* da un nuovo *cluster* [*ij*] è costituita dalla media delle distanze tra *i* e *k* e tra *j* e *k*. Per ogni *k* si effettua quindi il seguente calcolo:

$$\mathbf{d_{k\,(ij)}} = \frac{\mathbf{d_{ki} + d_{kj}}}{\mathbf{2}}$$

Nella matrice delle distanza presentata qui sopra, la distanza tra Iglesias e Siniscola si rivela la più piccola. Dopo aver raggruppato le due località in un *cluster*, si calcolano le distanze tra il nuovo *cluster* e gli elementi rimasti. Per esempio, la distanza tra il dialetto di Lodé e quelli ci Iglesias e Portuscuso si calcola nel modo seguente:

$$\mathbf{Lodé}_{\text{(Iglesias,Portoscuso)}} = \frac{\mathbf{d}_{\text{Lodé,Iglesias}} + \mathbf{d}_{\text{Lodé,Portoscuso}}}{\mathbf{2}}$$

$$= \frac{33{,}6 + 34{,}3}{2}$$

Dopo aver calcolato la distanza tra il dialetto di Lodé e la media di quelli di Iglesias-Portoscuso, Luras e Iglesias-Portoscuso e Iglesias-Portoscuso e Gesturi si ottiene la matrice seguente matrice (i nuovi valori sono rappresentati in grassetto, mentre quelli introdotti in precedenza sono rappresentati con caratteri normali):

	Lodé	**Luras**	**Gesturi**	**Iglesias & Portoscuso**
Lodé	0	17,9971	32,9584	**33,95**
Luras		0	30,095	**32,03**
Gesturi			0	**12,26**
Iglesias & Portoscuso				0

Il processo in cui ad ogni iterazione si effettua la riduzione di due lingue a un *cluster* si ripete fino a quando non è più possibile formare un nuovo *cluster*. Il risultato finale costituisce un raggruppamento gerarchico completo delle varietà linguistiche, che può essere visualizzato sotto forma di un dendrogramma: un albero in cui le foglie corrispondono alle singole varietà e la lunghezza dei rami rappresenta le distanze fonetiche. Il dendrogramma che risulta dal *clustering* di tutte le 77 varietà prese in esame è presentato in § 2.9.

1.8 Scalatura multidimensionale

Le distanze reciproche tra una serie di località si possono determinare sulla base delle loro coordinate. È anche possibile effettuare il procedimento contrario: a partire dalle distanze reciproche è possibile stabilire un sistema ottimale di coordinate che contiene quelle delle località in questione. Questo procedimento è reso possibile da una tecnica matematica conosciuta come *scalatura multidimensionale*. La scalatura multidimensionale è una tecnica matematica paragonabile all'analisi fattoriale (Kruskal & Wish 1984). Sulla trama di una scalatura multidimensionale, le lingue fortemente correlate vengono collocate le une vicine alle altre, mentre le lingue dissimili vengono distanziate.

Nei nostri esperimenti si è fatto uso delle *Multidimensional Scaling-routines* nel modulo statistico R, versione 1.3.0 (per informazioni e download: http://www. r-project. org/), il quale è stato applicato alla tabella che contiene le distanze tra i 77 dialetti sardi. Il modulo offre tre forme di scalatura multidimensionale, per la precisione: *Classical Multidimensional Scaling*, *Sammon's Non-Linear Mapping*, e *Kruskal's Non-metric Multidimensional Scaling*. La correlazione maggiore fra le *Distanze-Levenshtein* originarie nella tabella e le distanze euclidee misurate tra i punti della scalatura multidimensionale (0.99) è stata trovata facendo uso del *Kruskal's Non-metric Multidimensional Scaling*. Il risultato della scalatura multidimensionale ottenuto sulla base dei 77 dialetti sardi si può trovare al § 2.8.

Per poter valutare meglio il significato di entrambe le dimensioni della scalatura, si sono determinate separatamente le distanze euclidiche di entrambe le dimensioni tra le varietà, come riportato nella scalatura. Queste distanze sono correlate separatamente alle distanze-Levenshtein di ciascuna delle duecento parole. Da ciò risulta che la prima dimensione (la coordinata *y* nella scalatura) è quella maggiormente correlata con le distanze-Levenstein della parola *bandai* 'andare': *r*=0.93. Le varianti più importanti sono [a╜a™↔] (dialetti settentrionali), [anda™↔] (dialetti centrali) e [andai] (dialetti meridionali. La seconda dimensione (la coordinata *x* della scalatura) è maggiormente correlata con le distanze-Levenstein della parola *cosa* 'idem'. Le varianti più importanti sono [k←┘za] (dialetti settentrionali, occidentali e meridionali) e [k←┘z⇒a] (dialetti centro-orientali).