ALESSANDRO VIETTI

Approcci quantitativi all'analisi della variazione linguistica: il caso di GOLDVARB 2001

The article provides an up-to-date view of quantitative sociolinguistics, paying more attention to variation analysis as carried out in American research during the last decades. The emerging directions concern on one side the feasible inclusion of probabilistic analysis of variation into a model of grammatical competence (probabilistic sociolinguistics), on the other side the combination of ethnographic methods and quantitative analysis within the same research design (variation ethnography). Within this picture the basic features of Varbrul's latest PC version (Goldvarb 2001) will be illustrated, emphasizing advantages and drawbacks.

1. *Introduzione*

L'immagine che si ricava della sociolinguistica variazionista americana e britannica, in particolare dalla prospettiva della sociolinguistica italiana, è quella di un paradigma apparentemente in fase di contrazione, i cui concetti chiave, come le regole variabili, appaiono oggi piuttosto in disuso.[1] Se, come osserva Berruto (1995: 31), l'approccio correlazionale è stato dominante nella ricerca americana ed europea negli anni Settanta, dal decennio successivo si è assistito infatti a un riorientamento degli interessi in direzione di metodi più qualitativi e incentrati sull'agire comunicativo del parlante.

Per la verità, va anche aggiunto che, nel quadro della ricerca sociolinguistica italiana, a fronte di una "rapida penetrazione delle istanze sociolinguistiche" (Mioni 1992: 508) si accompagna una relativa esiguità di studi empirici condotti con la strumentazione tecnica della sociolinguistica quantitativa.

Anche intendendo con quantitativo il semplice uso di distribuzioni di

[1] In recenti trattazioni manualistiche come Milroy / Gordon (2003), Chambers / Trudgill / Schilling-Estes (2002) o Coulmas (1996) la nozione di regola variabile non viene menzionata o al più è indicata come una fase nel paradigma laboviano.

frequenze basate su campioni di dati sufficientemente rappresentativi, il panorama delle indagini empiriche sul repertorio dell'italiano non è molto ricco.[2] La constatazione della sostanziale mancanza di una descrizione empirica delle varietà dell'italiano saldamente ancorata ai dati è espressa tra gli altri da Berruto (1987: 18), Trumper / Maddalon (1990: 161) e Giannelli (1994: 51) e rimane purtroppo in gran parte attuale. La recente costituzione di grandi basi di dati sull'italiano come il progetto dell'Accademia della Crusca (v. Cresti 2000) o AVIP (Archivio delle Varietà di Italiano Parlato[3]), da un lato, e l'informatizzazione degli atlanti dialettali (v. Lo Piparo 1990; Pennisi 1996; D'Agostino 1997; De Masi 1995; De Masi 1998), dall'altro, può contribuire a superare questa lacuna ponendo le premesse per descrizioni quantitative di fenomeni già in larga parte analizzati con strumenti qualitativi. Infatti, questa carenza sul piano empirico-quantitativo è compensata dalla ricca produzione teorica, da un lato, e da quella empirica in senso più qualitativo, dall'altro (v. le rassegne di Mioni 1992; Berruto 2002).

Il punto di vista della ricerca italiana sulla sociolinguistica quantitativa sconta quindi un difetto di percezione o, forse, di ricezione: se la sociolinguistica laboviana ha goduto di una qualche considerazione nel recente passato, ciò è avvenuto nei termini di una ricezione condizionata che ha separato per così dire il nocciolo teorico dal complemento metodologico. Va osservato però che la sociolinguistica laboviana si presentava come disciplina eminentemente empirica nella quale la dimensione metodologica era forse preponderante su quella teorica; si trattava infatti di un modo innovativo di indagare un legame, quello tra lingua e società, senz'altro non nuovo per la tradizione linguistica e dialettologica italiana.

Per queste ragioni, al sociolinguista italiano l'approccio variazionista apparirà desueto e il ragionare sull'attualità di strumenti come VARBRUL sembrerà una sorta di inutile attaccamento a oggetti di "modernariato" scientifico.

[2] La monografia su Timau di Francescato / Solari Francescato (1994) è esemplare non soltanto per la solida base empirica ma anche perché esprime un atteggiamento di cauta sfiducia nei confronti di metodi statistici più avanzati che "non sembrano però aprire la via a risultati pienamente convincenti per il linguista e quanto più diventano sofisticati tanto più si rivolgono a un piccolo numero di specialisti" (Id.: 63).

[3] Per informazioni visitare il sito < http://www.cirass.unina.it/ricerca/studi%20parlato/raccolta%20corpora/raccoltacorpora.htm >.

Tuttavia le cose non stanno proprio così e, sebbene non vi siano cambiamenti eclatanti nella strumentazione statistica della sociolinguistica laboviana, vanno almeno registrati alcuni importanti segnali che denotano come l'opera di costante accumulazione delle conoscenze empiriche stia saturando il paradigma, provocando un curioso corto circuito con sottodiscipline diverse accomunate dall'orientamento empirista (v. il par. 2.).

Tra gli indizi di vitalità più rilevanti vanno senz'altro segnalate le recenti uscite, da un lato, di GOLDVARB 2001, ovvero una versione del programma informatico VARBRUL per MS Windows, ad opera di Robinson, Lawrence, Tagliamonte (2001) e, dall'altro, di una monografia molto dettagliata di Paolillo (2002) che chiarisce i principi statistici implicati in questo tipo di analisi quantitativa.

Lo scopo di questo articolo è quello di collocare l'analisi multivariata condotta con VARBRUL[4] all'interno del più ampio quadro dei metodi di analisi statistica impiegati nella sociolinguistica quantitativa americana (v. il par. 4.1.). In secondo luogo, si cercherà di illustrare il funzionamento di GOLDVARB 2001, alla luce delle riflessioni sviluppate in Paolillo (2002), cercando di mettere il ricercatore nelle condizioni di svolgere una semplice analisi e interpretarne i risultati.

2. *Tendenze emergenti nella sociolinguistica quantitativa*

Poiché mi è apparso recentemente chiaro che l'acronimo VARBRUL (VARiaBle RULes), molto noto tra i sociolinguisti, dice poco o nulla al linguista generale, è necessario fornire almeno alcuni accenni sulla storia di questo programma informatico.

VARBRUL nasce alla fine degli anni Settanta come strumento informatico-statistico di analisi multivariata in grado di stimare il peso relativo che le diverse variabili indipendenti hanno nel determinare i diversi valori della variabile linguistica dipendente, di norma dicotomica. Il modello teorico di riferimento è quello delle regole variabili, avanzato per la prima volta da Labov (1972a) come integrazione delle *context*

[4] Con VARBRUL si intenderà una classe di strumenti di analisi, un iperonimo che comprende i vari programmi come GOLDVARB 2001.

sensitive rules di Chomsky / Halle (1968), ed elaborato sul piano statistico e informatico dal matematico e linguista canadese David Sankoff (per una sintesi molto chiara v. Sankoff 1988) in successive versioni di VARBRUL.

Mentre inizialmente Labov (1972a) ipotizzava una relazione di tipo lineare tra le variabili – con l'inconveniente, nel caso di variabili dipendenti dicotomiche, di stimare probabilità superiori a 1 o inferiori a 0 –, Cedergren / Sankoff (1974) propongono una versione di VARBRUL con un modello di regressione non-lineare di tipo moltiplicativo che ovvia a questi problemi, introducendone purtroppo di nuovi.[5] L'ultimo approdo dal punto di vista strettamente statistico è la regressione logistica in grado di tradurre il modello moltiplicativo in uno additivo utilizzando i logaritmi naturali, la funzione logit e la funzione logistica. Questo modello inoltre ben si adatta al tipo di variabili discrete che caratterizzano lo studio della variazione sociolinguistica.

Il nuovo programma informatico della "famiglia" di VARBRUL dunque rappresenta una novità interessante perché rende finalmente disponibile, anche ai non utenti di Macintosh, uno strumento più amichevole del precedente VARBRUL per ambienti DOS elaborato da Pintzuk (1988). Il precedente programma infatti non possedeva una interfaccia grafica e richiedeva quindi all'utente di inserire manualmente, in ambiente DOS, i vari comandi per attivare i diversi programmi, e presupponeva inoltre una certa conoscenza del linguaggio di programmazione LISP[6] per la compilazione del *condition file*.[7] GOLDVARB 2001 è quindi una versione per Windows del più noto GoldVarb – ora giunto alla versione 2.1 – di Rand / Sankoff (1990) e a esso, in larga parte, è simile nelle funzionalità e nell'aspetto grafico.[8]

[5] Senza addentrarmi in terreni scivolosi e fuori dalla mia competenza, si può osservare che il modello moltiplicativo è asimmetrico rispetto alla scelta del valore di applicazione della variabile dipendente: in un modello simmetrico l'analisi che si svolge per uno dei due valori della variabile dicotomica dipendente deve essere in qualche modo inversa a quella svolta per l'altro valore della variabile (per maggiori dettagli v. Paolillo 2002: 159).

[6] LISt Programming, un linguaggio di programmazione molto usato nel campo dell'intelligenza artificiale.

[7] La terminologia informatica è evidentemente oscura per il lettore a questo punto ma più avanti verrà chiarito il significato di queste espressioni e la funzione dei rispettivi referenti.

[8] A differenza di VARBRUL per DOS, GoldVarb 2.1 e GOLDVARB 2001 non possono però svolgere un'analisi di variabili dipendenti politomiche, ovvero con più di due valori.

Per quanto riguarda invece il manuale di Paolillo (2002), gli intenti sembrano essere quelli di raccogliere in un unico contesto le conoscenze statistiche impiegate dai variazionisti laboviani per presentarle non soltanto a studenti e aspiranti sociolinguisti, ma anche al pubblico più vasto dei linguisti empiristi. Questo volume sembra rispondere alle sollecitazioni indirette che, da alcuni anni, vengono dagli ambiti della linguistica che fanno maggior uso di metodi e modelli statistici, come la linguistica computazionale o anche i recenti indirizzi di quello che potremmo definire un funzionalismo in chiave cognitivista orientato all'uso della lingua (v. Bybee 2001; Bybee / Hopper 2001).

L'ondata informatica si è abbattuta sulla linguistica sotto forma di maggiori potenzialità nel trattamento di grandi quantità di dati; in questo modo la statistica assume un ruolo non soltanto come strumento di descrizione e rappresentazione dei dati, ma anche di modellizzazione dei fenomeni linguistici nei termini di grammatica emergente e probabilistica nella quale la competenza del parlante si va definendo attraverso generalizzazioni di tipo induttivo a partire da *token* linguistici.

In questo senso, la sociolinguistica variazionista può vantare senz'altro un apparato metodologico ben consolidato e un solido patrimonio di conoscenze empiriche, almeno per quanto riguarda le realtà anglofone, che può essere fecondamente interrogato su un piano teorico da un orientamento funzionalista in grado di intrecciare i vincoli linguistici alla variazione con quelli cognitivi e sociali.[9]

Da queste contaminazioni vengono coniate due etichette sintagmatiche apparentemente stravaganti o futuribili come la sociolinguistica probabilistica (Mendoza-Denton / Hay / Jannedy 2003) e l'etnografia variazionista (Bayley 2002).

La prima nasce proprio dall'esigenza di elaborare una teoria della competenza basata sulla *performance*. La forte matrice empirica della disciplina, la grande attenzione per la dimensione metodologica e il sostanziale accantonamento delle regole variabili come proposta teorica fanno sì che, a tutt'oggi, la componente teorica sia senz'altro l'aspetto più debole della disciplina. A differenza del periodo in cui Labov pro-

[9] Non mancano tra l'altro consistenti interscambi anche tra variazionismo e modelli formali come l'*Optimality Theory* (v. Kiparsky 1993; Paolillo 2002: cap. 10), anche se appaiono meno promettenti.

pose le regole variabili (Labov 1972a) oggi non mancano invece teorie cosiddette *usage-based* in grado di confrontarsi più facilmente della grammatica generativa – interlocutore originario – o di altre teorie formali con il problema della variazione, in modo particolare per quel che riguarda la fonetica e la fonologia (v. Bybee 2001; Pierrehumbert 2001).

Questo significa che le attuali direzioni di indagine in linguistica consentirebbero di riabilitare almeno l'idea – se non la sua rappresentazione formale – di regola variabile (ovviamente da aggiornare), non tanto come strumento di rappresentazione di variabili sociolinguistiche,[10] ma soprattutto come modello della competenza del parlante.

Il rifiuto delle regole variabili come parte della grammatica del parlante era legato proprio al fatto che non si pensasse alla competenza grammaticale come a un fatto probabilistico, né fosse possibile intendere le categorie grammaticali come prototipi costruiti e aggiornati in base a delle generalizzazioni empiriche operate sulle frequenze di occorrenza dei *token* linguistici.

Come sostiene Pierrehumbert (2001: 195):

> This line of research [*quella probabilistica n.d.a.*] has established that the cognitive representation of sound structure is probabilistic, with frequencies playing a crucial role in the acquisition of phonological and phonetic competence, in speech production and perception, and in long-term mental representation.

Rendendo cognitivamente più plausibile l'esistenza di una competenza formata da nuvole di esempi, alcuni dei quali più prototipici o centrali di altri, si apre la strada a una qualche forma di incorporazione dei pesi probabilistici delle regole variabili nella competenza del parlante.

Come affermano in modo ancora piuttosto ipotetico Mendoza-Denton / Hay / Jannedy (2003: 136):

> In the exemplar-theoretic view [...] social information that is interpretable by the listener is automatically stored with the exemplar, made more robust with repetition, and crucially linked to the actual instances of use of a particular variant.

[10] Fasold (1989: 18; citato da Figueroa 1994: 104) le definisce "a display device", mentre Berruto (1995: 181) sostiene che le regole variabili possono "essere proficuamente utilizzate per esprimere in maniera economica e formalizzata variabili sociolinguistiche presenti nelle situazioni indagate".

Questo modo di pensare alla competenza del parlante rappresenta un banco di prova promettente per la sociolinguistica che ha così l'opportunità di connettere le proprie istanze con quelle dei modelli *usage-* o *frequency-based*, a patto di tenersi lontana dalla tentazione di attribuire un eccessivo realismo cognitivo alla rappresentazione probabilistica.

La seconda etichetta di etnografia variazionista (Bayley 2002: 134-136) si riferisce invece a un recente indirizzo di indagine che tenta di conciliare un approccio quantitativo e correlazionale con i metodi etnografici.

Anche se l'accostamento di metodi etnografici e analisi quantitativa di tipo laboviano sembra a prima vista contraddittorio, non è affatto impossibile condurre un'analisi quantitativa di dati sociolinguistici acquisiti con metodi etnografici. Studi come quello di Eckert (2000; v. anche la rassegna su lingua e identità di Mendoza-Denton 2002) sfruttano il lavoro di osservazione sul campo anche per elaborare categorie interpretative più vicine alla realtà osservata. In questo modo, si arricchisce il paradigma correlazionale di informazioni sociologiche che consentono una comprensione più approfondita dei meccanismi sociali che presiedono alla variazione linguistica (per l'uso di scale di implicazione per ricavare categorie sociali v. Vietti, in stampa).

Questa tendenza, in realtà, non è affatto nuova e ha precedenti espliciti almeno a partire dallo studio di Labov (1963) a Martha's Vineyard (ora tradotto in italiano in Giannini / Scaglione 2003) e impliciti, ma nemmeno troppo, nell'uso della nozione di *social network* introdotta nello studio di Leslie Milroy su Belfast (v. Milroy 1980).

Le macro-categorie sociali come classe, genere ed età sono variabili estremamente potenti nella descrizione della variazione ma, allo stesso tempo, concettualmente povere. Al loro interno agiscono infatti altri elementi concettuali o altre articolazioni di significati sociali che sono in grado di "spiegare" in modo più pregnante la variazione osservata.

L'indirizzo etnografico, così come quello della rete sociale, hanno lo scopo di rendere maggiormente operative le categorie sociali e demografiche fornendo così delle ipotesi precise sul comportamento linguistico dei parlanti.

In conclusione, si può osservare come sociolinguistica probabilistica ed etnografia variazionista vadano intesi come due tentativi di procedere, con strumenti quantitativi, verso una teoria della competenza attraverso l'uso, nella quale la variazione possa trovare spazio come nucleo costitutivo.

3. *Vantaggi e svantaggi di VARBRUL*

I possibili futuri legami tra la sociolinguistica e modelli di grammatica *usage-based* più che una maggiore diffusione di VARBRUL, strumento informatico piuttosto esoterico nella ricerca linguistica, sembrano prefigurare in realtà un suo superamento a favore di programmi statistici di più ampia fruizione e, soprattutto, caratterizzati da un'alta flessibilità di impieghi.

I segnali di vitalità sopra menzionati potrebbero al contrario essere indizi di un imminente tramonto di VARBRUL come strumento informatico poiché, se si vuole porre lo studio della variazione all'interno dell'alveo della linguistica empirista, deve essere possibile la comunicazione dei risultati anche al di fuori del proprio ambito disciplinare tramite i *software* commerciali impiegati comunemente nelle scienze umane, come SPSS o SAS (per una breve guida al *software* statistico v. Bohrnstedt, Knoke 1998: 451-457).

Tuttavia, a oggi, VARBRUL resta il *software* per analisi multivariata di più largo impiego in sociolinguistica. La sua nascita come programma specificamente dedicato all'analisi della variazione sociolinguistica ha rappresentato, per almeno due decenni, un punto di forza rendendolo uno strumento potente e relativamente facile da usare. Questo aspetto però si sta rivelando oggi un limite da due punti di vista, uno sociologico e uno più metodologico.

Nel primo caso, come abbiamo già accennato, VARBRUL non è uno strumento condiviso al di fuori della sociolinguistica. Ciò significa, non solo, che il formato della rappresentazione non è immediatamente comprensibile al pubblico dei linguisti che pure hanno conoscenze di statistica, ma che, soprattutto, il programma informatico non è conosciuto dagli statistici che, di conseguenza, non lo sanno adoperare. In questo modo si preclude la possibilità di un dialogo interdisciplinare con consulenti esperti di statistica.

In realtà, la regressione logistica, modello statistico su cui si basa VARBRUL, non è certo un'invenzione dei sociolinguisti, pertanto non ci sono motivazioni intrinseche che impediscano la comprensione e la comunicazione scientifica al di fuori dei confini della sociolinguistica di un'analisi condotta con VARBRUL.

È a questo punto che il problema da sociologico diviene metodologi-

co. Infatti, rispetto al periodo in cui è stato sviluppato VARBRUL, oggi si dispone di una grande varietà di *software* statistici in grado di svolgere sia le analisi statistiche di base che quelle più avanzate, come le analisi bivariate e multivariate. I *software* statistici in commercio dunque non svolgono una regressione logistica "migliore" di VARBRUL, ma semplicemente sanno fare quello e altro, offrendo dunque più strumenti di analisi al ricercatore. Il difetto di VARBRUL dunque consiste nell'essere uno strumento molto specifico e non modificabile: anche all'interno della regressione logistica, VARBRUL consente di condurre analisi solamente con variabili indipendenti discrete, requisito non intrinseco di questo tipo di analisi multivariata.

Tuttavia, sul piano specifico della regressione logistica tra variabili discrete, VARBRUL si dimostra uno strumento molto potente e affinato in decenni di applicazione in analisi sociolinguistiche e linguistiche *tout court*.

I vantaggi indubbi di VARBRUL, oltre al fatto di essere scaricabile gratuitamente da Internet,[11] risiedono, in primo luogo, nella funzione di *recoding* che consiste nella possibilità di manipolare piuttosto agilmente le relazioni tra le variabili dipendenti e indipendenti, costruendo così modelli di analisi che meglio si adattano ai dati osservati. Una seconda preziosa caratteristica di VARBRUL è quella che viene denominata *stepwise regression* o *step up/step down analysis*[12] e che consiste, detto in termini molto semplici, nel poter verificare la bontà di un modello aggiungendo o sottraendo una variabile indipendente alla volta con lo scopo di valutarne, per così dire, il ruolo svolto nella variazione della variabile dipendente. Attraverso queste due operazioni si può modificare il modello di partenza per ottenere un modello semplice ed economico – cioè formato dal minimo numero di variabili che siano in grado di spiegare in massimo grado la variazione – che si adatti bene ai dati e che risulti significativo.

[11] GOLDVARB 2001 può essere scaricato dal sito dell'Università di York < http://www.york.ac.uk/depts/lang/webstuff/goldvarb/ > (goldvarb.zip, 385 KB); GoldVarb 2.1 (la versione per Macintosh) si trova presso il sito < http://www.crm.umontreal.ca/~sankoff/GoldVarb_Eng.html >; VARBRUL 3M, la versione per PC, è scaricabile da < http://www.crm.umontreal.ca/~sankoff/Varbrul_MS-DOS.html >.

[12] Il lettore mi scuserà per la fastidiosa presenza di anglicismi la cui abbondanza è legata unicamente alla natura tecnica dei sottocodici informatico e statistico ai quali appartengono.

4. *Come usare VARBRUL*

Prima di illustrare il funzionamento del programma informatico è importante chiarire quali sono gli obiettivi e i limiti di questa breve introduzione. L'obiettivo è quello di presentare, in modo non eccessivamente tecnico, i passi necessari per condurre un'analisi multivariata con GOLDVARB 2001: dalla preparazione dei dati al loro trattamento, fino all'interpretazione dei risultati, utilizzando come esempio i dati della ben nota ricerca di Labov sulla presenza di /r/ svolta in tre grandi magazzini newyorkesi (Labov 1972b).[13]

L'esposizione si terrà alla giusta distanza dai due poli dell'informatica e della statistica, illustrando la dimensione funzionale dello strumento informatico e del modello statistico e rimandando il lettore interessato agli aspetti più tecnici, nel primo caso, ai cosiddetti *users' manuals* (per GOLDVARB 2001 v. Robinson / Lawrence / Tagliamonte 2001; per GoldVarb 2.1 Rand / Sankoff 1990; per VARBRUL 3M Pintzuk, 1988) e, nel secondo caso, a manuali di statistica generali (cfr. Bohrnstedt / Knoke 1998) o specificamente rivolti a un pubblico di linguisti (un manuale di base è Woods / Fletcher / Hughes, 1986; contiene tecniche avanzate di analisi Rietveld / van Hout 1993).

4.1. *Regressione logistica e modelli statistici di analisi multivariata*

La scelta del modello statistico da adottare procede di pari passo con la definizione delle ipotesi da indagare e delle variabili da prendere in considerazione e con la determinazione delle tecniche di campionamento. Dunque, la valutazione delle svariate opzioni di modellizzazione quantitativa dei dati dovrebbe far parte della fase di progettazione della struttura della ricerca. A seconda, infatti, della complessità dell'ipotesi e della natura metodologica esplorativa o sperimentale, varieranno i tipi di dati, le dimensioni e la natura del campione. Ipotesi molto articolate che prevedono di osservare l'interazione tra molte variabili con molti

[13] I dati codificati, più tecnicamente il *token file*, si trovano all'indirizzo < http://ella.slis.indiana.edu/~paolillo/projects/varbrul/data/ >. Presso la *homepage* di John Paolillo < http://ella.slis.indiana.edu/~paolillo/projects/varbrul/ > si trovano anche numerose informazioni circa i progetti di aggiornamento informatico di VARBRUL.

valori necessiteranno di campioni numerosi e ben bilanciati, ovvero con un numero minimo di elementi per cella; inoltre il tipo di variabile può dare luogo a misurazioni continue o discrete. Questi due aspetti in particolare rappresentano dei discrimini forti nella selezione della attrezzatura adeguata messa a disposizione dalla statistica. Per fare un esempio, la regressione log-lineare richiede che tutte le variabili, dipendenti e indipendenti, siano di tipo discreto o categoriale, mentre la regressione lineare è applicabile anche a dati di tipo continuo.

Compito di questo breve paragrafo è pertanto fornire una mappa ingenua dei principali metodi di analisi multivariata all'interno della quale collocare la regressione logistica, il modello di analisi applicato nelle più recenti versioni di VARBRUL.

La regressione logistica appartiene a un'ampia famiglia di strumenti di analisi multivariata che si basano sul principio della regressione lineare e per questo vengono detti modelli lineari generalizzati, più comunemente indicati secondo l'acronimo GLM, *Generalized Linear Models* (sulla regressione lineare e i GLM v. Woods / Fletcher / Hughes 1986: 224-247; su tecniche avanzate come regressione multipla, log-lineare e logistica cfr. Rietveld / van Hout 1993: capp. 3, 8 e 9; una chiara introduzione ai metodi di analisi multivariata rivolta agli scienziati sociali si trova in Bohrnstedt / Knoke 1998: capp. 9-10; su GLM v. anche Paolillo 2002: 175-189).

Un'altra importante tipologia di modelli – troppo poco usata in sociolinguistica – è quella che ha lo scopo di ridurre le dimensioni di analisi, ovvero di individuare se le variabili prese in considerazione abbiano qualcosa in comune o siano istanze di una unica macro-variabile soggiacente. All'interno di questa classe troviamo tecniche come l'analisi delle componenti principali (un buon esempio di uso di questa tecnica si trova in Horvath 1985), la *cluster analysis* e l'analisi fattoriale (v. sempre Woods / Fletcher / Hughes 1986: capp. 14-15; Rietveld / van Hout 1993: capp. 6-7). L'analisi fattoriale, assieme a una rigorosa applicazione della statistica inferenziale, è impiegata in Trumper / Maddalon (1990), uno dei rari studi di sociolinguistica italiana condotti con strumenti di analisi multivariata.

La famiglia delle tecniche di regressione, di cui ci si occuperà di seguito, condivide una proprietà comune, ovvero la forma base dell'equazione che stima una relazione lineare tra le variabili:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \beta_k X_{ki} + \varepsilon$$

dove:
- Y_i rappresenta il valore della variabile dipendente per una data osservazione i,
- X_{1i}, X_{2i} ecc. rappresentano i valori delle variabili indipendenti o esplicative per la data osservazione i;
- α è la costante, ovvero il valore che assumerebbe Y se le variabili indipendenti fossero uguali a zero;
- β_1, β_2 ecc. sono i parametri del modello o coefficienti di regressione che esprimono l'effetto esercitato dalla variabile indipendente X su quella dipendente Y;
- ε rappresenta l'errore, cioè quella parte del valore assunto da Y che non è spiegabile con il modello di regressione lineare.

La relazione si dice lineare perché a un incremento nel valore di una variabile indipendente X corrisponderà sempre un incremento costante nella variabile dipendente Y pari a β volte X. Il significato del termine regressione è invece meno trasparente e la sua scelta è piuttosto infelice poiché si è estesa all'intera classe di procedimenti la denominazione di un caso specifico, ovvero la regressione verso la media.[14] Tuttavia, in statistica oggi si usa "regredire" con il significato di ricavare i valori di Y sulla base di quelli di X, "regrediamo" cioè Y rispetto a X.

Questi modelli sono applicati normalmente a variabili, indipendenti e dipendenti, continue e il loro impiego in linguistica è pertanto legato soprattutto a quegli ambiti che prevedono dati con misurazioni quantitative quali la fonetica o la linguistica acquisizionale come nel caso della misurazione del tempo (come per esempio la relazione tra età e lunghezza media degli enunciati, MLU - *mean length of utterance*, nell'acquisizione della prima lingua).

I modelli appartenenti a questa famiglia dunque condividono la forma base lineare, questa può però essere trasformata attraverso una fun-

[14] L'introduzione del termine "regressione" si deve allo statistico Karl Pearson che, studiando la relazione tra le altezze di un campione di padri e figli, osservò l'esistenza di una particolare relazione lineare: i padri più bassi della media avevano figli alti più di loro, mentre i padri più alti della media avevano figli più bassi di loro. In ogni caso i figli erano più vicini dei rispettivi padri all'altezza media. Per esprimere questa relazione Pearson adottò l'espressione "regressione verso la media", i figli regredivano rispetto ai padri verso la media.

zione matematica consentendo così di concepire le relazioni tra variabili in modo non lineare. Una sottofamiglia di funzioni è quella logaritmica che dà luogo ai modelli di regressione log-lineare e logistica.

Il primo modello nasce come sviluppo dell'analisi di tabelle di contingenza multivariate o multidimensionali (v. par. 4.2.2. tab. 2) e poiché si applica esclusivamente a variabili di tipo discreto ha goduto di grande successo nel campo delle scienze umane. La regressione log-lineare viene denominata anche analisi delle frequenze poiché non assume a priori una distinzione tra variabili dipendenti o indipendenti, ma stima quale sia l'effetto delle singole variabili e delle combinazioni tra di esse sulla frequenza di un dato fenomeno.

Questo modello non è stato molto applicato in sociolinguistica perché, di norma, in questa disciplina la variazione è concepita, anche se non sempre esplicitamente, come una scelta tra due varianti da parte dei parlanti. Tuttavia la variazione può anche essere intesa come l'effetto di un insieme di variabili sulla frequenza relativa di un fenomeno. Per fare un esempio in ambito italiano si potrebbe analizzare la variazione del pronome clitico *ci*, dativo di III persona singolare (*a lui ci piace la cioccolata*), non in relazione alle sue varianti più standard, ma osservando quali variabili – per esempio età, strato sociale o grado di istruzione – influenzano la sua maggiore o minore frequenza. Un esempio di impiego della regressione log-lineare è De Masi (1995) che utilizzando i dati dell'Atlante NADIR propone un'analisi esplorativa dell'interazione tra competenza del dialetto, scelta di codice (italiano e dialetto), età e istruzione.

La regressione logistica è invece un modello molto usato in sociolinguistica proprio grazie alla trasposizione informatica ad opera di David Sankoff e collaboratori (Cedergren / Sankoff 1974; Rousseau / Sankoff 1978; Sankoff 1988).

Il modello utilizzato in VARBRUL corrisponde a un caso particolare di regressione logistica nel quale le variabili indipendenti devono essere categoriche o discrete. In realtà, questo vincolo non è posto dalla regressione logistica come tale, ma riflette le esigenze tipiche dell'indagine variazionista nella quale i fattori che influenzano la variazione sono normalmente categorie socio-demografiche oppure categorie linguistiche discrete (es. ± contesto seguente consonantico). Inoltre, non tutte le versioni di VARBRUL – e tra queste purtroppo figura, per ora, anche il nuovo GOLDVARB 2001 – consentono di condurre analisi su variabili

dipendenti politomiche, cioè con più di due valori, possibile in teoria con la regressione logistica.[15]

In sintesi, la regressione logistica impiegata in VARBRUL stabilisce, in termini di probabilità, quale sia l'effetto relativo di una serie di fattori sociali e linguistici sulla variazione linguistica, solitamente intesa come scelta tra due valori (sulla regressione logistica impiegata in VARBRUL v. Paolillo 2002: 153-173). Nel caso dello studio di Labov (1972b) sulla cancellazione o meno di [r] in posizione finale di parola, la scelta può essere influenzata da fattori come lo strato sociale, il genere, l'appartenenza etnica, il contesto fonetico precedente e seguente e così via. Da un lato, dunque, avremo la variabile dipendente dicotomica, presenza o assenza di [r], dall'altro, le variabili indipendenti dicotomiche o politomiche come le categorie socio-demografiche e i fattori linguistici.

VARBRUL calcola per ogni cella, cioè ogni combinazione unica delle variabili indipendenti, la proporzione di cancellazioni di [r][16] sulle occorrenze totali. Poi trasforma queste proporzioni in valori numerici che variano lungo tutto lo spettro dei numeri reali da più infinito a meno infinito attraverso la funzione logit. Questi valori numerici vengono poi utilizzati per stimare i vari parametri dell'equazione di regressione attraverso una procedura iterativa denominata stima di massima verosimiglianza (in inglese MLE, *Maximum Likelihood Estimation*) che ha lo scopo appunto di determinare l'equazione che più si avvicina ai dati osservati, ovvero è in grado di "spiegarli" meglio di altre (su MLE cfr. Paolillo 2002: 170-173; Bohrnstedt / Knoke 1998: 301-303).

I parametri dell'equazione espressi in termini di logit vengono poi ritrasformati in probabilità, in valori numerici compresi cioè tra 0 e 1, utilizzando la funzione inversa alla logit ovvero la funzione logistica. I parametri espressi in forma di probabilità costituiscono il risultato finale dell'intero processo e sono i cosiddetti "pesi" che i diversi valori delle variabili hanno nel determinare la scelta tra le varianti.

Una caratteristica matematico-statistica dalla quale discende un notevole vantaggio metodologico è la relativa adattabilità dell'analisi anche

[15] L'analisi con variabili dipendenti politomiche, denominata nel lessico "varbruliano" *Multinomial Analysis*, può essere condotta con la versione di VARBRUL per DOS tramite il programma MVARB. Questo limite di GOLDVARB 2001 è parzialmente superabile concependo i valori della variabile dipendente come una successione gerarchica di analisi dicotomiche.

[16] O di realizzazione, questo dipende dalle scelte del ricercatore.

a campioni non perfettamente bilanciati. Cosa significa questo? A differenza di altre tecniche statistiche, il perfetto bilanciamento del campione – ovvero un uguale numero di occorrenze per combinazione di valori delle variabili – non è un requisito matematico necessario al funzionamento della regressione logistica.

Questo consente dunque un certo grado di sbilanciamento nel campione, tipico di dati raccolti in ricerche sul campo e non in contesti controllati come quelli sperimentali. Poiché la maggior parte delle indagini sociolinguistiche è di tipo esplorativo, l'acquisizione di dati sul campo è, per necessità, non del tutto controllabile da parte del ricercatore. Anche per questa ragione la regressione logistica ben si adatta alle esigenze della ricerca variazionista.

Nel caso dell'indagine di Labov (1972b), Paolillo (2002: 43) tenta di ricostruire il campione in un'unica tabella – non presente nell'articolo originale di Labov – che sintetizzi le distribuzioni di frequenza per le varie categorie sociali prese in considerazione.

Social categories	Sacks	Macy's	Klein's
African American	3	14	25
White Female	49	65	27
Other	16	46	19
Floorwalkers	–	13	–
Sales personnel	–	105	–
Stock clerks	–	7	–
Older	9	38	20
Middle	31	54	36
Younger	23	26	22
Total	68	125	71

Tabella 1 - Distribuzioni di frequenza del campione di Labov (1972b) in Paolillo (2002: 43)

Come appare evidente dalla tabella 1, il campionamento è fortemen-

te disomogeneo e lacunoso rispetto alle categorie individuate e pertanto non è possibile generalizzare sull'uso della variabile dipendente tra gli impiegati afro-americani del grande magazzino Sacks, che in questa indagine rappresenta un indicatore dello strato sociale (alto), o prendere in considerazione fattori come l'appartenenza etnica o il tipo di qualifica (*Floorwalkers* ecc.). Tuttavia anche se le caselle non sono riempite con un uguale numero di occorrenze è possibile includere i fattori età e strato sociale tra le variabili indipendenti.

È necessario quindi distinguere tra requisiti matematici, necessari per poter eseguire l'analisi, e quelli di statistica inferenziale che consentono di generalizzare i risultati con diversi gradi di attendibilità. Il fatto che i requisiti matematici non escludano l'esistenza di caselle vuote o non impongano il riempimento di ogni casella con un uguale numero di occorrenze non autorizza il ricercatore ad affrontare con leggerezza la fase di campionamento. In poche parole, se il campione non è stato approntato secondo criteri di rappresentatività, avendo cura di evitare celle vuote o con pochi elementi, l'analisi che verrà effettuata, sebbene possibile, non sarà attendibile e si incorrerà in interpretazioni errate del fenomeno osservato (Paolillo 2002: 44).[17]

4.2. *Mini-guida all'uso di GOLDVARB 2001 per Windows*

Illustrati i principi generali di VARBRUL da un punto di vista statistico si possono ora indicare i passi basilari per procedere a un'analisi informatica dei dati sociolinguistici, in questo caso rappresentati dall'indagine di Labov (1972b).

Come già osservato, oltre ad avere riflessi sul tipo di campionamento, la regressione logistica svolta con VARBRUL pone forti limitazioni anche nella fase di progettazione dell'indagine. Sebbene l'obiettivo di questa e delle seguenti sezioni sia quello di introdurre il funzionamento di GOLDVARB 2001, è necessario includere come primo passo

[17] I molti problemi statistici e metodologici legati alla fase di campionamento non possono essere trattati qui per evidenti ragioni di spazio; per approfondimenti sul piano metodologico v. Milroy (1987: 18-38) e il recente Milroy / Gordon (2003: 23-48), sugli aspetti del campionamento relativi a VARBRUL v. Paolillo (2002: 40-46), mentre per quanto riguarda le basi statistiche del campionamento e la statistica inferenziale si può cfr. Woods / Fletcher / Hughes (1986: capp. 4, 6, 7).

nell'analisi almeno la fase di generazione delle ipotesi. La realtà empirica osservata deve infatti passare per il collo di bottiglia rappresentato dai requisiti dell'analisi quantitativa, così come le ipotesi che intendiamo verificare devono essere codificate secondo i parametri richiesti dal programma.

Pertanto, il primo passo per condurre un'analisi quantitativa con VARBRUL è quello di concepire il fenomeno osservato in termini di variabile dipendente e di più variabili indipendenti. Se poi la relazione tra questi due tipi di variabile sia da intendere come un rapporto di causazione multipla o come una mera correlazione tra più fattori dipenderà dall'interpretazione del ricercatore e non dall'analisi statistica in sé.

L'oggetto di indagine, cioè l'alterna presenza di varianti di un fenomeno linguistico variabile, deve essere costretto preferibilmente in due valori (per l'individuazione e l'analisi della variabile dipendente v. Wolfram 1993). Nel caso di Labov (1972b) la variabile dipendente sarà /r/ in posizione postvocalica (*car*, *fourth* ecc.) nei parlanti inglese americano a New York e i due valori saranno alternativamente la sua presenza o assenza.

Il problema della generazione di ipotesi riguarda principalmente quante e quali variabili indipendenti includere, ovvero, nella terminologia di VARBRUL, quali e quanti *factor groups* codificare e prendere in considerazione. Questo è evidentemente un problema che il ricercatore deve risolvere in relazione alle opzione teoriche e metodologiche poste dalla singola indagine.

In questo contesto vanno però messe in rilievo le possibili conseguenze che alcune scelte possono avere sull'analisi. Un problema interessante è rappresentato dalla quantità di *factor groups* da inserire, ovvero di quante e quali ipotesi di correlazione verificare. Young / Bayley (1996: 257-258) suggeriscono al ricercatore di adottare un atteggiamento piuttosto *liberal* in questa fase, moltiplicando le ipotesi da testare, poiché è più semplice eliminare ipotesi non significative che non inserirne di nuove una volta codificati i dati.

Il principale rischio insito in un uso poco accorto di questa strategia "liberale" è quello della sovracodifica (*overcoding*) che, richiedendo un campione esponenzialmente proporzionale al numero di variabili e valori per variabile, espone fatalmente il modello a numerose caselle vuote per carenze di campionamento (*sampling zero*) o perché semplicemente

alcune intersezioni di variabili non si danno nella realtà empirica (*structural zero*). Paolillo (2002: 135) cita proprio uno studio di Young (1991) nel quale la combinazione dei 34 valori delle 10 variabili indipendenti dà 46.080 possibili celle, delle quali nei dati ne sono attestate solo 799 in un campione di 1.600 *token*. Se pensassimo che almeno 5 occorrenze per ogni cella siano sufficientemente rappresentative, dovremmo costruire un campione di almeno 230.400 occorrenze. Questo significa che, o il numero delle ipotesi e dei fattori da prendere in considerazione è stabilito prima di procedere al campionamento, oppure le ipotesi che si possono generare deve essere commisurato al campione che è stato raccolto. In caso contrario, la regressione logistica può essere senz'altro effettuata, ma si corre il rischio di voler estrarre molte più informazioni dal modello di quanta evidenza empirica ci sia in realtà a giustificarle. Il modello poggerebbe dunque su basi molto incerte perché la stima dei parametri dell'equazione è molto poco attendibile.

Se è dunque consigliabile, su un piano metodologico, non spingersi troppo oltre nella moltiplicazione dei fattori, è forse opportuno anche per motivi di ordine teorico non costruire modelli eccessivamente complessi e, in ultima analisi, di difficile interpretazione.

Applicando queste riflessioni al caso dell'indagine di Labov (1972b) possiamo concludere che il campione non consente di testare delle ipotesi relative all'appartenenza etnica, all'impiego, al sesso e all'età, almeno utilizzando una tecnica come la regressione logistica. Si possono senz'altro ricavare molte informazioni di tipo qualitativo anche per queste variabili e Labov (1972b: 229-239) stesso ne propone alcune come ulteriore prova a supporto della sua ipotesi principale. Le distribuzioni di frequenza sembrano infatti confermare l'ipotesi generale della stratificazione sociale della variabile sociolinguistica /r/ (Labov 1972b: 222): "if any two subgroups of New York City speakers are ranked in a scale of social stratification, then they will be ranked in the same order by their differential use of (r)".

Le variabili indipendenti dei dati newyorkesi che utilizzeremo nell'analisi con VARBRUL saranno dunque quelle linguistiche relative ai due contesti fonetici elicitati (nelle parole *fourth* e *floor*) e allo stile di enunciazione normale ed enfatico, e una sociologica relativa allo strato sociale del parlante rappresentato dai tre grandi magazzini *Sacks*, *Macy's* e *Klein*. Labov e i suoi collaboratori chiedevano ai vari tipi di

commessi di questi tre grandi magazzini l'indicazione per qualche merce che si trovasse al quarto piano e, fingendo, di non aver compreso correttamente sollecitavano una ripetizione. Ognuna di queste richieste anonime di indicazioni veniva poi rapidamente registrata su un taccuino, raccogliendo per ogni indicazione quattro occorrenze della variabile.

4.2.1. *Codificare i dati*

L'input di partenza, la prima cosa che si "dà in pasto" al programma, è un *token file*, intuitivamente costituito dall'insieme delle occorrenze del fenomeno la cui variazione si vuole comprendere.

Questo insieme di occorrenze va ovviamente costruito codificando i dati audio e video che il ricercatore ha trascritto seguendo le proprie idiosincratiche convenzioni (di norma rappresentate da un documento di testo) oppure alcuni degli standard più comunemente usati negli studi linguistici come TEI (*Text Encoding Initiative*), XML (*Extensible Markup Language*) oppure CHILDES (*Child Language Data Exchange System*, v. MacWhinney 1997). Va osservato che in sociolinguistica non esiste in realtà uno standard, anche in conseguenza del fatto che la maggioranza degli studi sono di tipo esplorativo e il *corpus* di dati raccolto e trascritto dal singolo ricercatore non è di dimensioni tali da rendere opportuna una codifica e un trattamento quantitativo.

Per condurre un'analisi con VARBRUL non è necessario che i dati siano codificati in un particolare tipo di linguaggio informatico, ma può rivelarsi un utile risparmio di tempo se i dati sono interrogabili anche solo attraverso una lista di concordanze che individuino il fenomeno interessato e un intorno linguistico, in caso contrario al ricercatore non resterà che compulsare manualmente le trascrizioni alla ricerca di occorrenze della variabile dipendente da codificare.

L'operazione di codifica dei dati per renderli leggibili da VARBRUL dunque non è altro che un'osservazione delle trascrizioni alla ricerca delle occorrenze del fenomeno indagato alle quali assegnare dei simboli. Ogni *token* codificato rappresenta quindi una sintesi delle informazioni rilevanti per l'analisi e si presenta come una stringa di numeri o lettere.

Nell'esempio dell'indagine di Labov (1972b) assegneremo ai due

valori della variabile dipendente /r/ i simboli *0* per l'assenza e *1* per la presenza; ai tre valori della variabile strato sociale i simboli *S* per *Sacks*, *M* per *Macy's* e *K* per *Klein*; ai due valori della variabile stile *n* per l'enunciazione *normale* ed *e* per quella enfatica; infine ai due valori della variabile contesto fonetico *4* per *fourth* e *F* per *floor*.

Così un "brano" di un *token file* si presenterà come segue:

(1Sn4 commesso)
(1Sn4 ricercatore presente)
(1SnF giornata piovosa)
(1Se4 John riga 124)
(1Mn4 ...)
(0Sn4 ...)
(0Se4 ...)

La sintassi di un *token file* è importante perché il programma interpreterà le informazioni a seconda della posizione occupata: dopo la parentesi aperta il valore in prima posizione deve essere quello della variabile dipendente; i tre simboli seguenti rappresentano i valori delle tre variabili indipendenti per un dato *token*; l'informazione che è seguita da uno spazio è invece un commento che viene accluso alla codifica del *token*, in genere le indicazioni di riga in un documento di testo, il nome dell'intervistato, caratteristiche contestuali rilevanti e così via. Il risultato finale è il *token file*, cioè un file di testo che contiene una lista di tante righe quante saranno le occorrenze della variabile da studiare.[18]

Una volta approntato il *token file* non resta che aprirlo all'interno di GOLDVARB 2001 andando al menu *View* della finestra principale e scegliendo il comando *Token*. Questa operazione aprirà la finestra *Tokens* dove sarà sufficiente scegliere *Open* dal menu *File* e procedere all'apertura del *token file*.[19]

Prima di procedere nell'analisi, è possibile verificare la correttezza della codifica, istruendo il programma con i dati relativi ai valori dei diversi *factor groups*. Nella finestra *Tokens* si dovrà scegliere *Action/Generate factor spec's...*, rispondere *OK* alla domanda nella finestra di dia-

[18] Per facilitare l'operazione di codifica si possono utilizzare dei programmi come fogli di calcolo che rendono in parte automatica la compilazione della stringa.

[19] È anche possibile creare un nuovo *token file* direttamente all'interno di questa finestra.

logo e il programma genererà i valori da assegnare ai *factor groups* passando in rassegna i *tokens*; all'utente non resterà che controllare nella finestra *Groups* che i valori rilevati corrispondano con quelli stabiliti, quindi scegliere *Save to Token file* e poi *OK*. In alternativa, si può aprire la finestra *Groups* scegliendola all'interno del menu *View* della finestra principale. Qui si clicca su *New Group*, si seleziona la casella *Factors* (in modo che sia circondata da una linea punteggiata), quindi nel campo *New Factor* va inserito il primo valore del primo *factor group*, in questo caso *0*, e quindi si preme *Add*. Una volta inseriti i valori del *factor group* è necessario aggiungere un valore di *default* (tra quelli del *factor group*) nella relativa casella, poi si continua con le altre variabili premendo ogni volta *New Group*. Al termine scegliere *Save to token file* e quindi *OK*.

A questo punto nella finestra *Tokens* scegliere *Action* e poi *Check tokens* per controllare che non vi siano errori di cattiva codifica. Se non vi sono, nella finestra principale apparirà la scritta *Checking of tokens completed. 729 tokens in 729 lines.*, in caso contrario il programma darà per esempio il seguente messaggio *Error in group #2 in the token "1sn4" Tkn:11.* con l'indicazione del *factor group* e del *token* mal codificato. In questo caso sarà sufficiente scorrere il *token file* fino alla riga 11 e modificare il valore errato.

4.2.2. *Esplorare i dati*

Dopo aver creato e verificato il *token file* tramite la definizione dei valori dei *factor groups*, il passo successivo è quello di creare un *condition file*. Quest'ultimo contiene, in sintesi, il modello che vogliamo testare, cioè le istruzioni su come devono essere intesi i rapporti tra le variabili contenute nelle varie stringhe del *token file*: quale deve essere la variabile dipendente e quali quelle indipendenti, o se alcune variabili o valori di variabili devono essere fuse in sovracategorie, o ancora se alcuni valori o intere variabili debbano sparire perché non significative (v. *infra* 4.2.4.).

Inizialmente è opportuno eseguire un'analisi, per così dire, semplice, con tutte le variabili codificate: scegliendo dal menu *Action* nella finestra *Tokens* il comando *No recode* verrà creato un *condition file* di de-

fault, cioè senza particolari condizioni imposte all'analisi successiva.

L'oggetto successivo di cui GOLDVARB 2001 ha bisogno per svolgere l'analisi è il cosiddetto *cell file*, un file che contiene, in un linguaggio non molto trasparente all'occhio umano, l'equivalente di una tabella di contigenza multidimensionale, cioè l'incrocio dei diversi valori delle variabili rappresentati come nella tabella 2.

	S				M				K			
	n		e		n		e		n		e	
	4	F	4	F	4	F	4	F	4	F	4	F
0	16	31	16	21	33	48	13	31	3	5	6	7
1	39	18	24	12	81	62	48	20	63	59	40	33
Tot	55	49	40	33	114	110	61	51	66	64	46	40

Tabella 2 - Esempio di tabella di contingenza multidimensionale

Per creare un *cell file* l'utente dovrà aprire la finestra *Results* dal menu *View* della finestra principale, qui dal menu *Action* selezionare *Load cells to memory*; a questo punto apparirà una finestra di dialogo che chiede di scegliere l'*application value*, cioè il valore di applicazione della supposta regola variabile. Se in questo caso assegnamo il valore 1, i risultati finali andranno interpretati come il contributo di ciascun fattore alla pronuncia di [r], ovvero la variante ritenuta prestigiosa.

Nella finestra *Results* apparirà dunque il *results file* che rappresenta una tabella sintetica con le frequenze assolute e percentuali dei valori di applicazione e non applicazione per ogni gruppo di fattori. È opportuno salvare il *results file* e, nelle rispettive finestre, anche il *condition file* e il *cell file* assegnando loro delle etichette per quanto possibile trasparenti. Poiché questi rappresentano i file di un'analisi semplice, usando cioè la funzione *No recode*, si possono nominare per esempio *dsNoRec.res*, *dsNoRec.cnd* e *dsNoRec.cel*.

Altro utile aiuto per il ricercatore è quello di tenere in un documento di testo a parte una sorta di resoconto delle varie analisi che si compiono, in particolare quando si ripetono molte analisi prima di giungere al modello ottimale.

Come accennato, il *results file* costituisce una sorta di prima sintesi dei dati. Una sua osservazione approfondita da parte del ricercatore può già mettere in evidenza alcuni potenziali ostacoli all'analisi successiva. Un esempio di *results file* di Labov (1972b) è offerto di seguito nella tabella 3.

```
CELL CREATION
=============
Name of token file: C:\ds.tkn
Name of condition file: dsNoRec.cnd
(
(1)
(2)
(3)
(4)
)
      Number of cells:       12
      Application value(s):   1
      Total no. of factors:   7
```

Group		Apps	Non-apps	Total	%
1 (2)					
S	N	84	93	177	24
	%	47	52		
M	N	125	211	336	46
	%	37	62		
K	N	21	195	216	29
	%	9	90		
Total	N	230	499	729	
	%	31	68		
2 (3)					
n	N	136	322	458	62
	%	29	70		
e	N	94	177	271	37
	%	34	65		
Total	N	230	499	729	
	%	31	68		

```
 3 (4)
    4  N    87   295   382    52
       %    22    77

    F  N   143   204   347    47
       %    41    58

Total  N   230   499   729
       %    31    68
──────────────────────────────────
Total  N   230   499   729
       %    31    68

Name of new cell file: dsNoRec.cel
```

Tabella 3 - Esempio di *results file*

In alto appaiono le indicazioni del *token file* e del *condition file*, una copia del *condition file* e alcune informazioni riassuntive come il numero delle celle, il valore di applicazione scelto e il numero dei fattori. Di seguito è riportata invece la tabella del *results file* con le distribuzioni assolute e percentuali dei valori di applicazione per i vari fattori. È da notare che i gruppi di fattori sono rinumerati in questa fase: mentre la variabile dei grandi magazzini è la seconda a comparire nelle stringhe dei *token* e nella specificazione dei *factor groups*, qui diventa il gruppo 1.

In questa fase, il ricercatore deve andare alla ricerca di quelli che vengono definiti *knockout factors* e *singleton groups*, che nel caso dei dati sui grandi magazzini in realtà non compaiono. I primi sono quei fattori che presentano tutte occorrenze di applicazione o di non-applicazione, sono fattori quindi dove di fatto non c'è variazione ma appartenenza categorica a un valore della variabile dipendente. Si avrebbe un *knockout factor* se, per esempio, le 177 occorrenze di Sacks della tabella 3 fossero tutte o applicazioni ([r]) o non applicazioni ([Ø]). In questi casi l'analisi non può funzionare e quindi bisogna modificare le condizioni di analisi nel *condition file*.

I secondi invece sono quei gruppi di fattori nei quali tutte le occorrenze ricadono all'interno di un unico fattore di fatto quindi senza una variazione interna al gruppo di fattori; questi gruppi di fattori o vengono eliminati dall'analisi oppure, se possibile, possono essere scomposti in altri fattori. Il *singleton group* si avrebbe nel caso in cui tutte le occor-

renze della variabile grandi magazzini ricadessero all'interno del valore Sacks.

In ogni caso, il programma segnala la presenza di questi potenziali problemi con un avvertimento a fianco del fattore o gruppo di fattori nel *results file*.

Oltre all'individuazione degli aspetti problematici, l'osservazione di questa tabella fornisce già qualche indicazione circa le tendenze nella variazione della variabile. In questo caso, per esempio, è evidente come la presenza della [r], variante di prestigio, diminuisca man mano che da Sacks – magazzino della classe alta – ci si sposti verso Klein.

Il nucleo della fase esplorativa dei dati alla ricerca di regolarità nella variazione è rappresentato, ancora più che dal *results file*, dalle tabelle incrociate (*cross-tabulation*), ovvero quelle tabelle che intersecano due variabili indipendenti alla volta.

Per creare queste tabelle è necessario scegliere il comando *Cross tabulation* dal menu *Action* della finestra *Results*. Per costruire una tabella incrociata delle variabili strato sociale e stile l'utente dovrà immettere nei campi *Recoded Group 1* e *Recoded Group 2* rispettivamente i valori 1 e 2 (corrispondenti alla nuova numerazione). L'opzione *Text* o *Grid* nell'area *Output format* riguarda il formato della tabella incrociata (testo o tabella) che appare come segue nella tabella 4.

Group	1	S	S	M	M	K	K	Total	Total
2	App Value	Count	%	Count	%	Count	%	Count	%
n	1	47	45	81	36	8	6	136	30
n		57	55	143	64	122	94	322	70
n	Total	104		224		130		458	
e	1	37	51	44	39	13	15	94	35
e		36	49	68	61	73	85	177	65
e	Total	73		112		86		271	
Total	1	84	47	125	37	21	10	230	32
Total		93	53	211	63	195	90	499	68
Total	Total	177		336		216		729	

Tabella 4 - Esempio di tabella di contigenza multidimensionale: strato sociale e stile

In questo caso la tabella incrociata è utile per confermare la tendenza già osservata di una diminuzione della presenza di [r] da Sacks (45%) a Macy's (36%) e a Klein (6%) nel discorso normale, ma anche per evidenziare il diverso effetto della variabile stile sulla variabile dipendente suddivisa per strato sociale.

Analizzare le tabelle incrociate serve anche per individuare eventuali interazioni tra variabili indipendenti in grado di falsare l'analisi multivariata dei dati. Nel caso seguente, la tabella 5 che incrocia contesto fonetico e stile di elicitazione mette in evidenza un'interazione tra le due variabili.

Group	3	4	4	F	F	Total	Total
2	App Value	Count	%	Count	%	Count	%
n	1	52	22	84	38	136	30
n		183	78	139	62	322	70
n	Total	235		223		458	
e	1	35	24	59	48	94	35
e		112	76	65	52	177	65
e	Total	147		124		271	
Total	1	87	23	143	41	230	32
Total		295	77	204	59	499	68
Total	Total	382		347		729	

Tabella 5 - Contesto fonetico e stile

Per la parola *fourth* la differenza tra stile normale (22%) ed enfatico (24%) è solo del 2%, mentre nella parola *floor* l'effetto dello stile sembra più evidente con una differenza del 10%. Sembra dunque esserci un'interazione tra le due variabili la cui origine non è a prima vista chiara e che dunque meriterebbe di essere verificata statisticamente prima di avanzare delle ipotesi.

4.2.3. *Analisi multivariata*

Dopo aver creato il *condition file*, il *cell file* e aver ispezionato il *results file* e le tabelle di contingenza alla ricerca di potenziali ostacoli o tendenze emergenti, è possibile eseguire una regressione logistica dei dati.

Nella finestra *Results* scegliere dal menu *Action* il comando *Binomial, 1 level* per eseguire un'analisi, per così dire, elementare dei dati. Prima di visualizzare la finestra chiamata *Binomial Varbrul*, appare una finestra contenente un diagramma di dispersione che fornisce, in forma grafica, informazioni circa la bontà di adattamento dei dati al modello di regressione. La linea diagonale rappresenta l'ottimale adattamento dei dati al modello costruito perché è costituita dalla coincidenza di valori osservati e attesi sulla base del modello.

La finestra *Binomial Varbrul* contiene invece i risultati dell'analisi vera e propria come riportato di seguito nella tabella 6.

```
Binomial Varbrul, 1 step
========================
Name of cell file: dsNoRec.cel

Using fast, less accurate method.
Averaging by weighting factors.

- One-level analysis only:One-level binomial analysis:

Run # 1, 12 cells:
Convergence at Iteration 6
Input 0,274

  Group  Factor   Weight   App/Total   Input&Weight

    1:       S      0,704       0,47          0,47
             M      0,605       0,37          0,37
             K      0,202       0,10          0,09

    2:       n      0,470       0,30          0,25
             e      0,551       0,35          0,32
    3:       4      0,383       0,23          0,19
             F      0,628       0,41          0,39
```

Cell	Total	App'ns	Expected	Error
SnF	49	31	28,070	0,716
Sn4	55	16	18,183	0,392
SeF	33	21	21,462	0,028
Se4	40	16	16,262	0,007
MnF	110	48	50,986	0,326
Mn4	114	33	27,516	1,441
MeF	51	31	27,801	0,809
Me4	61	13	18,677	2,487
KnF	64	5	8,000	1,286
Kn4	66	3	3,299	0,028
KeF	40	7	6,615	0,027
Ke4	46	6	3,128	2,828

```
Total Chi-square  = 10,3757
Chi-square/cell   = 0,8646
Log likelihood    = -394,812
```

Tabella 6 - I risultati della regressione logistica

Le prime righe contengono informazioni sul file di input e il metodo di analisi,[20] mentre di seguito si trovano delle indicazioni circa il numero di celle analizzate, il numero di iterazioni dell'algoritmo che stima i parametri e il valore dell'input.

Quest'ultimo può essere pensato come una specie di media dei valori delle variabili e rappresenta la probabilità che il fenomeno in questione ha di presentarsi indipendentemente dai vari fattori in gioco. L'*input value* o l'*input probability* di un modello è dunque una misura cruciale perché costituisce una sorta di punto di riferimento in relazione al quale valutare gli effetti positivi e negativi dei diversi fattori.

La prima tabella rappresenta il cuore dell'analisi perché presenta i pesi dei fattori in riferimento alla variabile dipendente. Le prime due colonne indicano i gruppi di fattori e i fattori mentre nella terza colonna *Weight* si trovano i pesi dei diversi fattori. La quantificazione del ruolo di ciascun fattore nella pronuncia o meno della /r/ è indicato in termini

[20] "Using fast, less accurate method" si riferisce alla precisione di calcolo, mentre "averaging by weighting factors" si riferisce al metodo usato per stabilire il valore di input che altro non è che il valore di α dell'equazione di regressione.

di probabilità, quindi lungo una scala numerica che varia tra 0 e 1. I valori dei fattori vanno interpretati diversamente a seconda che siano maggiori di 0,5 o minori di 0,5: nel primo caso (> 0,5) hanno un effetto positivo sulle probabilità di verificarsi del fenomeno indagato (la presenza di [r] in posizione postvocalica), nel secondo caso (< 0,5) non hanno un effetto positivo più attenuato, bensì un effetto negativo. Così, mentre *S* e *M* (strati alto e medio) favoriscono l'uso di [r], *K* (strato basso) lo sfavorisce. I fattori *n* e *e* attestandosi entrambi attorno allo 0,5 indicano che l'intero gruppo di fattori (la variabile) forse non è significativa nella variazione di /r/. Per la terza variabile, *F* sembra favorire la presenza di [r], mentre *4* eserciterebbe un effetto negativo.

La quarta colonna riporta i valori osservati nei dati raccolti, mentre la quinta colonna rappresenta la probabilità attesa di occorrenza di [r] calcolata combinando l'effetto del peso del fattore con il valore dell'input.

Osservando questi risultati dunque possiamo concludere che l'ipotesi iniziale di Labov (1972b) è verificata, poiché la presenza di [r] è condizionata principalmente dallo strato sociale e successivamente dall'ambiente fonologico.

La seconda tabella introduce invece una serie di informazioni che riguardano principalmente l'adeguatezza del modello, la sua capacità di adattamento ai dati osservati. Nella prima colonna sono infatti elencate le varie celle analizzate, nella seconda il numero delle occorrenze, nella terza il numero di applicazioni (ovvero di produzioni di [r]), nella quarta il numero delle applicazioni attese, nella quinta colonna si trova invece una misurazione dell'errore, un numero che si ricava combinando valore osservato e atteso.

L'adattamento del modello si può valutare confrontando per ogni cella la distanza tra valori osservati e attesi nonché il rispettivo errore. Per la cella *Me4* il modello prevede che nel campione vi siano 18,677 [r], mentre nei dati ne osserviamo solamente 13 con una differenza di 5,677 e un errore di 2,487. Valori dell'errore per cella al di sopra di 3,84[21] devono richiamare l'attenzione del ricercatore che dovrà osservare i dati alla ricerca di possibili interazioni tra i fattori oppure riconfigurare il *condition file*. La somma dei valori della colonna degli errori dà

[21] Questo valore rappresenta una probabilità di 0,05 nella distribuzione del chi quadrato con 1 grado di libertà.

una misura totale del chi quadrato che serve a valutare l'adattamento complessivo del modello ai dati.

L'ultimo dato riguarda il *log-likelihood*, una misura anche questa di buon adattamento al modello molto simile, nel principio e nella distribuzione, al chi-quadrato. Questa misura è in grado di dirci quale parte della varianza presente nei dati è spiegata dal modello e qual è invece la proporzione di quella residua.

Da un mero punto di vista pratico ci è utile sapere che il *log-likelihood* è sempre un numero negativo con un massimo teorico pari a zero che rappresenterebbe la piena verosimiglianza del modello ai dati. In poche parole più vicino allo zero è il *log-likelihood*, migliore sarà il modello elaborato (sulle modalità di calcolo del *log-likelihood* v. Paolillo 2002: 137-139). Questa misura è molto importante in senso comparativo quando si tratta di confrontare la bontà di modelli diversi attraverso una misura chiamata *likelihood ratio* o rapporto di verosimiglianza.

4.2.4. *Analisi* step-up / step-down

Il rapporto di verosimiglianza costituisce anche una misura di significatività che permette di valutare quali fattori o gruppi di fattori non siano significativi e vadano quindi eliminati dal modello (su questa misura si v. Paolillo 2002: 140-142; Rietveld / van Hout 1993: 334-339; Bohrnstedt / Knoke 1998: 301-303).

Lo scopo dell'analisi statistica infatti è quello di ottenere il modello più semplice ed economico che risulti maggiormente significativo. Questo obiettivo può essere raggiunto con GOLDVARB 2001 effettuando quella che viene chiamata *Step-up/Step-down analysis* e che consiste molto semplicemente nella comparazione, tramite il rapporto di verosimiglianza, di modelli che differiscono l'uno dall'altro per la presenza o assenza di una variabile o gruppo di fattori. Il processo viene eseguito in salita o per aggiunta di variabili (*step-up*) confrontando il modello del livello 0, cioè con nessuna variabile indipendente, con quelli di livello 1, cioè con una variabile indipendente, e, successivamente, quelli del livello 1 con quelli del livello 2, cioè modelli con due variabili indipendenti, fino a costruire il modello con tutte le variabili indipendenti codificate. Di seguito il programma effettua un'analisi per sottrazione, in di-

scesa (*step-down*) per così dire, confrontando il modello con tutte le variabili indipendenti con i modelli del livello immediatamente inferiore con, a turno, una variabile in meno fino ad arrivare al livello uno.

Ogni modello è confrontato con il migliore del livello precedente attraverso appunto un test di significatività rappresentato dal rapporto di verosimiglianza che, avendo una distribuzione analoga a quella del chi quadrato, può essere calcolato in modo simile[22].

Al termine dell'analisi il programma individua automaticamente i due modelli migliori in salita e in discesa che, se non vi sono ostacoli legati a un cattivo campionamento o a una interazione tra variabili indipendenti, dovrebbero coincidere.

Per eseguire questo processo nella finestra *Results*, nel menu *Action* invece di scegliere il comando *Binomial, 1 level* si clicca sull'opzione *Binomial Up & Down*, a questo punto si aprirà nuovamente la finestra *Binomial Varbrul* con i risultati.

Un esempio completo dell'output di questo tipo di analisi non può essere riportato per ragioni di spazio; quanto segue rappresenta l'analisi in salita, il modello migliore di quella in discesa e i risultati finali.

```
Binomial Varbrul
================
Name of cell file: dsNoRec.cel

Using fast, less accurate method.
Averaging by weighting factors.
Threshold, step-up/down: 0,050001

# Stepping up:                                  analisi per aggiunta (step-up)
# Stepping up:

──────── Level # 0 ────────                     livello zero:
                                                modello con nessuna variabile

Run # 1, 1 cells:
Convergence at Iteration 2
Input 0,316
Log likelihood = -454,481
```

[22] Il rapporto di verosimiglianza, denominato anche G^2, si basa sul rapporto tra *log-likelihood* del modello più complesso e quello meno complesso. Il numero dei gradi di libertà è rappresentato dalla differenza tra gradi dei due modelli, mentre il numero dei gradi di libertà di un singolo modello si ricava sottraendo dal numero dei fattori il numero dei gruppi di fattori e aggiungendo una unità.

```
──────── Level # 1 ────────
```
livello 1: modello a 1 variabile

```
Run # 2, 3 cells:
Convergence at Iteration 5
Input 0,284
Group # 1 — S: 0,695, M: 0,599, K: 0,214
Log likelihood = -413,116 Significance = 0,000
```
strato sociale

```
Run # 3, 2 cells:
Convergence at Iteration 4
Input 0,315
Group # 2 — n: 0,479, e: 0,536
Log likelihood = -453,506 Significance = 0,171
```
stile

```
Run # 4, 2 cells:
Convergence at Iteration 4
Input 0,308
Group # 3 — 4: 0,399, F: 0,611
Log likelihood = -440,092 Significance = 0,000
```
contesto fonetico

```
Add Group # 1 with factors SMK
```
risultato livello 1: aggiungere strato sociale al modello

```
──────── Level # 2 ────────
```
livello 2: modello a 2 variabili

```
Run # 5, 6 cells:
Convergence at Iteration 5
Input 0,283
Group # 1 — S: 0,693, M: 0,602, K: 0,212
Group # 2 — n: 0,474, e: 0,544
Log likelihood = -411,793 Significance = 0,105
```
strato sociale
stile

```
Run # 6, 6 cells:
Convergence at Iteration 6
Input 0,275
Group # 1 — S: 0,706, M: 0,602, K: 0,204
Group # 3 — 4: 0,385, F: 0,626
Log likelihood = -396,501 Significance = 0,000
```
strato sociale
contesto fonetico

```
Add Group # 3 with factors 4F
```
risultato livello 2: aggiungere contesto fonetico al modello

```
──────── Level # 3 ────────
```
livello 3: modello a 3 variabili

```
Run # 7, 12 cells:
```

```
Convergence at Iteration 6
Input 0,274
Group # 1 – S: 0,704, M: 0,605, K: 0,202      strato sociale
Group # 2 – n: 0,470, e: 0,551                stile
Group # 3 – 4: 0,383, F: 0,628                contesto fonetico
Log likelihood = -394,812 Significance = 0,070

No remaining groups significant               risultato livello 3:
                                              non aggiungere stile

Groups selected while stepping up: 1 3        risultato analisi per aggiunta:
                                              aggiungi strato sociale (1)
                                              e contesto fonetico (3)

Best stepping up run: #6
_____________________

...

Run # 10, 6 cells:                            miglior modello con analisi
                                              per sottrazione
Convergence at Iteration 6
Input 0,275
Group # 1 – S: 0,706, M: 0,602, K: 0,204      strato sociale
Group # 3 – 4: 0,385, F: 0,626                contesto fonetico
Log likelihood = -396,501 Significance = 0,070

...

Groups eliminated while stepping down: 2      risultato analisi per sottrazione:
                                              elimina stile (2)

Best stepping up run: #6                      risultato finale delle due analisi
                                              (step-up e step-down)
Best stepping down run: #10
```

Dopo alcune informazioni generali, seguono i risultati del processo di analisi in salita. Ogni livello contiene l'indicazione del numero di celle analizzate, del numero di iterazioni dell'algoritmo prima di raggiungere la convergenza sui parametri dell'equazione, il valore dell'input, il gruppo di fattori preso in esame e il peso relativo per ogni fattore. L'ultima riga è dedicata alla misura del *log-likelihood* del modello e del livello di significatività del rapporto di verosimiglianza che, come già osservato, ci dice se, aggiungendo una data variabile, il modello sia in gra-

do di spiegare meglio i dati osservati. Per esempio, in *Run # 3* l'apporto della variabile 'stile' è valutato come un contributo poco significativo con p = 0,17, cioè con un'alta probabilità che non sia significativo[23].

Al termine dell'analisi in salita il programma individua le variabili 'strato sociale' e 'contesto fonetico' come significative e scarta invece la variabile 'stile'. Anche nell'analisi in discesa si ottiene il medesimo risultato, confermato nel riassunto finale dove il miglior modello in salita e in discesa coincidono.

A questo punto il ricercatore ha di fronte due possibilità: accogliere le indicazioni dell'analisi informatica ed eleggere il modello a due variabili indipendenti come il migliore rispetto ai dati raccolti, oppure ritornare circolarmente all'analisi, cercando di capire per quale motivo la variabile 'stile' sia stata eliminata.

Seguire questa seconda opzione porterebbe questa mini-guida ben oltre la sua natura, poiché si tratta di operazioni, per così dire, avanzate per le quali si rimanda a Paolillo (2002: 57-70) per una trattazione generale, e a Robinson / Lawrence / Tagliamonte (2001) per le indicazioni più pratiche.

Questo processo è di norma denominato *recoding* e consiste nella manipolazione del *condition file* allo scopo di vagliare nuove ipotesi attraverso la selezione del comando *Recode setup* nel menu *Action* della finestra *Tokens*. Tramite questa funzione si possono costruire modelli differenti per esempio eliminando la variabile 'stile', oppure scomponendo i dati in due sottocampioni – uno per 4, *fourth*, e uno per F, *floor* – per indagare l'effetto differente che la variabile stile esercita sulla frequenza del valore di applicazione della variabile dipendente.

Dunque, l'analisi circolare dei dati porta non soltanto alla verifica o falsificazione dell'ipotesi di partenza ma, molto spesso, alla generazione di nuove ipotesi da testare in successive indagini.

5. *Conclusioni*

La natura puramente illustrativa di mini-guida all'uso di VARBRUL non implica in realtà che venga tracciata nessuna particolare conclusio-

[23] Stabilendo come criterio di riferimento $p \leq 0{,}05$.

ne e, inoltre, la mancanza di una tradizione di uso di questo strumento nella sociolinguistica italiana non permette nemmeno un bilancio della ricerca svolta o l'indicazione di inedite direzioni di ricerca.

La domanda che ci si potrebbe porre a questo punto è: quale può essere l'utilità di VARBRUL? Se si vuole perseguire la strada dell'analisi statistica multivariata ovviamente VARBRUL non è l'unica soluzione. Oggi, esistono infatti strumenti informatici a pagamento estremamente versatili come SPSS o SAS con strutture modulari in grado di soddisfare sia le esigenze di base che quelle più avanzate. Inoltre se la sociolinguistica quantitativa aspira a essere comprensibile all'intera comunità dei linguisti di area funzional-empirista, la scelta di VARBRUL potrebbe risultare particolarmente marcata.

Tuttavia, dal punto di vista del sociolinguista italiano, GOLDVARB 2001 può costituire un ottimo laboratorio per incominciare a cimentarsi con l'analisi multivariata attraverso uno strumento mirato e, soprattutto, avendo a disposizione un consistente e ormai quasi trentennale *corpus* di acquisizioni scientifiche nella sociolinguistica anglofona con il quale confrontarsi.

Non è da escludere, inoltre, la possibilità che la conoscenza di VARBRUL al di fuori del mondo della sociolinguistica possa farne uno strumento di più ampio utilizzo, come è già accaduto per la ricerca in linguistica acquisizionale (v. in particolare Bayley / Preston 1996; ma anche Adamson 1988; Tarone 1988; Preston 1989). Ciò che dovrebbe essere condivisa tra i ricercatori è una comune conoscenza dell'analisi statistica e non l'impiego dei medesimi strumenti informatici o di standard di presentazione grafica dei risultati, quest'ultima senz'altro utile sebbene non indispensabile.

Bibliografia

Adamson, Hugh D., 1988, *Variation and Second Language Acquisition*, Washington D.C., Georgetown University Press.

Bayley, Robert, 2002, "The quantitative paradigm". In: Chambers / Trudgill / Schilling-Estes (eds.): 117-141.

Bayley, Robert / Preston, Dennis (eds.), 1996, *Second Language Acquisition and Linguistic Variation*, Philadelphia, Benjamins.

Berruto, Gaetano, 1987, *Sociolinguistica dell'italiano contemporaneo*, Roma, La Nuova Italia Scientifica.

Berruto, Gaetano, 1995, *Fondamenti di sociolinguistica*, Roma / Bari, Laterza.

Berruto, Gaetano, 2002, "Sociolinguistica". In: Lavinio, Cristina (a c. di), *La linguistica italiana alle soglie del 2000 (1987-1997 e oltre)*, Roma, Bulzoni: 471-503.

Bohrnstedt, George W. / Knoke, David, 1998, *Statistica per le scienze sociali*, Bologna, Il Mulino [*Statistics for Social Data Analysis*, Itasca, Peacock, 1994].

Bybee, Joan, 2001, *Phonology and Language Use*, Cambridge, Cambridge University Press.

Bybee, Joan / Hopper, Paul (eds.), 2001, *Frequency and the Emergence of Linguistic Structure*, Amsterdam / Philadelphia, Benjamins.

Cedergren, Henrietta J. / Sankoff, David, 1974, "Variable rules: Performance as a statistical reflection of competence". *Language* 50/2: 333-355.

Chambers, Jack K. / Trudgill, Peter / Schilling-Estes, Natalie (eds.), 2002, *The Handbook of Language Variation and Change*, Oxford, Blackwell.

Chomsky, Noam / Halle, Morris, 1968, *The Sound Pattern of English*, New York, Harper and Row.

Coulmas, Florian, 1996, *The Handbook of Sociolinguistics*, Oxford, Blackwell.

Cresti, Emanuela, 2000, *Corpus di italiano parlato. Introduzione*, vol. I, Firenze, Accademia della Crusca.

D'Agostino, Mari (a c. di), 1997, *Aspetti della variabilità. Ricerche linguistiche siciliane*, Palermo, Centro di studi filologici e linguistici siciliani.

De Masi, Salvatore, 1995, "Un modello di analisi quantitativa per il NADIR-Salento". In: Romanello, Maria Teresa / Tempesta, Immacolata (a c. di), *Dialetti e lingue nazionali*, Atti del XXVII Congresso della Società di Linguistica Italiana (Lecce, 28-30 ottobre 1993), Roma, Bulzoni: 135-154.

De Masi, Salvatore, 1998, "Le solidarietà della lingua. Alcuni strumenti di analisi quantitativa". In: D'Onofrio, Salvatore / Gualdo, Riccardo (a c. di), *Le solidarietà. La cultura materiale in linguistica e antropologia*, Atti del seminario di Lecce (novembre-dicembre 1996), Galatina, Congedo: 161-181.

Eckert, Penelope, 2000, *Linguistic Variation as Social Practice: The Linguistic Construction of Identity in Belten High*, Oxford, Blackwell.

Fasold, Ralph, 1989, "The quiet demise of the variable rules". Paper presented at NWAVE 18, North Carolina, Duke University.

Figueroa, Esther, 1994, *Sociolinguistic Metatheory*, Pergamon, Elmsford.

Francescato, Giuseppe / Solari Francescato, Paola, 1994, *Timau. Tre lingue per un paese*, Galatina, Congedo.

Giannelli, Luciano (a c. di), 1994, *Una teoria e un modello per l'analisi quantificata dell'italiano substandard*, Padova, Unipress.

Giannini, Stefania / Scaglione, Stefania (a c. di), 2003, *Introduzione alla sociolinguistica*, Roma, Carocci.

Horvath, Barbara, 1985, *Variation in Australian English*, Cambridge, Cambridge University Press.

Kiparsky, Paul, 1993, "An optimality-theoretic perspective on variable rules". Paper presented at NWAVE 23, Stanford, Stanford University.

Labov, William, 1963, "The social motivation of a sound change". *Word* 19: 273-307.

Labov, William, 1972a, "Contraction, deletion, and inherent variability of the English copula". In: Labov, William, *Language in the Inner City. Studies in the Black English Vernacular*, Philadelphia, University of Pennsylvania Press: 65-129.

Labov, William, 1972b, "The social stratification of (r) in New York City department stores". In: Linn, Michael D. (ed.), 1998, *Handbook of Dialects and Language Variation*, San Diego, Academic Press: 221-244 [originariamente in Labov, William, 1972, *Sociolinguistic Patterns*, Philadelphia, University of Pennsylvania Press].

Lo Piparo, Franco (a c. di), 1990, *La Sicilia linguistica oggi*, Palermo, Centro di studi filologici e linguistici siciliani.

MacWhinney, Brian, 1997, *Il Progetto CHILDES. Strumenti per l'analisi del linguaggio parlato*, Tirrenia, Edizioni del Cerro.

Mendoza-Denton, Norma, 2002, "Language and identity". In: Chambers / Trudgill / Schilling-Estes (eds.): 475-495.

Mendoza-Denton, Norma / Hay, Jennifer / Jannedy, Stefanie, 2003, "Probabilistic sociolinguistics: Beyond variable rules". In: Bod, Rens / Hay, Jennifer / Jannedy, Stefanie (eds.), *Probabilistic Linguistics*, Cambridge Mass., MIT Press: 97-138.

Milroy, Lesley, 1980, *Language and Social Networks*, Oxford, Blackwell.

Milroy, Lesley, 1987, *Observing and Analysing Natural Language. A Critical Account of Sociolinguistic Method*, Oxford, Blackwell.

Milroy, Lesley / Gordon, Matthew, 2003, *Sociolinguistics. Method and Interpretation*, Oxford, Blackwell.

Mioni, Alberto M., 1992, "Sociolinguistica". In: Mioni, Alberto M. / Cortelazzo Michele A. (a c. di), *La linguistica italiana degli anni 1976-1986*, Roma, Bulzoni: 507-536.

Paolillo, John C., 2002, *Analyzing Linguistic Variation. Statistical Models and Methods*, Stanford, CSLI Publications.

Pennisi, Antonino, 1996, "Si può informatizzare la variabilità linguistica? Esperienze dell'ALS e dell'OLS". In: Thun, Harold / Radtke, Edgar (Hrsg.), *Neue Wege der Romanischen Geolinguistik*, Kiel, Westensee-Verlag: 389-429.

Pintzuk, Susan, 1988, *VARBRUL Programs for MS-DOS*, Philadelphia, University of Pennsylvania Department of linguistics, < http://www.crm.umontreal.ca/~sankoff/Varbrul_MS-DOS.html >.

Pierrehumbert, Janet B., 2001, "Stochastic Phonology". *Glot International* 5/6: 195-207.

Preston, Dennis R., 1989, *Sociolinguistics and Second Language Acquisition*, Oxford, Blackwell.

Rand, David / Sankoff, David, 1990, *GoldVarb. A variable rule application for Macintosh* (version 2.1), Montreal, Centre de recherches mathématiques, Université de Montréal, < http://www.crm.umontreal.ca/~sankoff/GoldVarbManual.Dir/ >.

Rietveld, Toni / van Hout, Roeland, 1993, *Statistical Techniques for the Study of Language and Language Behaviour*, Berlin / New York, de Gruyter.

Robinson, John / Lawrence, Helen / Tagliamonte, Sali, 2001, *GOLDVARB 2001. A Multivariate Analysis Application for Windows*, < http://www.york.ac.uk/depts/lang/webstuff/goldvarb/manual/manualOct2001.html >.

Rousseau, Pascale / Sankoff, David, 1978, "Advances in variable rule methodology". In: Sankoff, David (ed.), *Linguistic Variation: Models and Methods*, New York, Academic Press: 57-69.

Sankoff, David, 1988, "Variable rules". In: Ammon, Ulrich / Dittmar, Norbert / Mattheier, Klaus J. (eds.), *Sociolinguistics: An International Handbook of the Science of Language and Society*, Vol. 2, Berlin, de Gruyter: 984-997.

Tarone, Elaine, 1988, *Variation in Interlanguage*, London, Arnold.

Trumper, John / Maddalon, Marta, 1990, "Il problema delle varietà: l'italiano parlato nel Veneto". In: Cortelazzo, Michele A. / Mioni, Alberto M. (a c. di), *L'italiano regionale*, Atti del XVIII Congresso della Società di Linguistica Italiana (Padova-Vicenza, 14-16 settembre 1984), Roma, Bulzoni: 159-191.

Vietti, Alessandro, in stampa, *Come gli immigrati cambiano l'italiano. L'italiano di peruviane come varietà etnica*, Franco Angeli, Milano.

Wolfram, Walt, 1993, "Identifying and interpreting variables". In: Linn, Michael D. (ed.), *Handbook of Dialects and Language Variation*, San Diego, Academic Press: 291-306 [apparso in Preston, Dennis (ed.), 1993, *American Dialect Research*, Philadelphia, Benjamins: 193-221].

Woods, Anthony / Fletcher, Paul / Hughes, Arthur, 1986, *Statistics in Language Studies*, Cambridge, Cambridge University Press.

Young, Richard, 1991, *Variation in Interlanguage Morphology*, New York, Lang.

Young, Richard / Bayley, Robert, 1996, "VARBRUL Analysis for Second Language Acquisition Research". In: Bayley / Preston (eds.): 253-606.