L'estadística en l'anàlisi de la variació fonètica: una aplicació del programa Goldvarb

per Josefina Carrera i Sabaté

Resum

Amb aquest article, l'autora pretén descriure el funcionament del programa de càlcul Goldvarb, tenint en compte les aplicacions estadístiques d'aquest en l'anàlisi sociolingüística i, en concret, de la variació fonètica.

1. Aspectes generals

La sociolingüística variacionista ha esmerçat molts esforços en el perfeccionament de les tècniques quantitatives d'anàlisi que es proposen demostrar la importància dels contextos lingüístics i socials en la variació. Així, «probability theory to the data allows us to extract higher-order regularities that govern variation in the community» (Labov, 1994:25). Amb aquesta finalitat, «el método variacionista busca el cálculo de la probabilidad de que aparezca un rasgo lingüístico determinado en unas circunstancias lingüísticas, sociológicas y contextuales determinadas. A partir de los datos de frecuencia recogidos en un grupo de hablantes, se crea un modelo teórico formado por las probabilidades de que se dé un fenómeno cuando concurren diversas circunstancias. La estadística se encarga de precisar hasta qué punto las probabilidades calculadas son verosímiles y cuáles son las circunstancias que, al darse simultáneamente, pueden explicar mejor un hecho lingüístico». (Moreno, 1994:95)

La sociolingüística treballa amb dos tipus d'estadística: a) *l'estadística descriptiva*, que compta i ordena quantitativament les dades extretes de la realitat; i b) *l'estadística d'inferències*, que aplica els resultats de l'estadística descriptiva i els adapta a les realitats d'una comunitat lingüística no estudiades.

En l'ús de *l'estadística d'inferències* l'objecte principal de l'estudi és allò que s'anomena la *variable dependent*. Aquesta variable dependent es veu determinada per les *variables independents* o *explicatives*, que en estudis de llengua són els elements lingüístics i els elements sociosituacionals. Per poder establir la *probabilitat* que un fenomen variable es manifesti d'unes determinades maneres cal saber, en primer lloc, quantes vegades s'ha aplicat en relació amb tots els casos possibles; això s'aconsegueix amb el recompte de les freqüències d'aparició d'un determinat tret en cadascuna de les condicions previstes i en els discursos recollits d'una mostra de parlants. Després, un cop comptats els casos particulars en què es manifesta un factor, es busca quina és la *freqüència* amb què es dóna el fenomen quan coincideixen diversos factors explicatius.

L'anàlisi probabilística permet saber: 1) en quin grau diferents grups de *factors explicatius* determinen la variació d'un element quan tots aquests factors explicatius actuen conjuntament; i 2) quin és el comportament general d'una comunitat, encara que només s'hagin recollit les dades d'una mostra estratificada de parlants. Amb les probabilitats s'elabora un model de la competència sociolingüística dels parlants que prediu tendències de futur.

El primer model probabilístic aplicat a l'anàlisi lingüística parteix d'un model additiu; després, s'arriba a un model logisticomultiplicatiu[1] basat en la fórmula que segueix:

$$\frac{p}{1-p} = \frac{p_o}{1-p_o} \times \frac{p_i}{1-p_i} \times \frac{p_j}{1-p_j} \times \ldots$$

Els avenços matemàtics de la sociolingüística han aparegut entre 1969 -a partir del treball de Labov (1969)- i 1978, i han estat complementats amb programes informàtics que utilitzen el càlcul estadístic. Els noms de dos programes que segueixen aquesta línia d'investigació són: Varbrul per a Pc i Vax, que realitza anàlisis multinomials,[2] i Goldvarb, per a Macintosh, que calcula la probabilitat amb una variable dependent binomial.

[1] Vegeu una presentació dels models a Cedergren i Sankoff (1974) i (1988), López Morales (1989) i Moreno (1994). Per a una explicació detallada de les limitacions dels models additiu i multiplicatiu vegeu Kay i McDaniels (1979).
[2] L'anàlisi multinomial es basa en l'estudi d'una variable dependent que pot tenir més de dos valors diferents; en la binomial, els valors de la variable dependent són només dos.

2. Objectius

L'objectiu d'aquest article és explicar de manera breu el funcionament del programa de càlcul Goldvarb a partir de l'estudi d'un procés de canvi en curs analitzat a la població segrianenca d'Alguaire i valorar l'aplicació del programa en el camp de la sociolingüística variacionista.

El procés de variació fònica que analitzo s'inclou en el vocalisme àton del lleidatà i té a veure amb *e-* inicial en posició pretònica inicial absoluta de mots com *enciam, escala* o *eriçó*. Diferents estudis dialectològics realitzats al llarg del segle XX han demostrat que, tradicionalment, aquesta vocal s'ha emès amb la solució [a]; tanmateix, s'observa un procés de variació fònica que tendeix cap a la substitució de la solució [a] per una altra que correspon a les formes de la llengua escrita: [e]. L'anàlisi se centra en informants de 3 a 80 anys del poble d'Alguaire i mostra, en definitiva, la influència que la llengua escrita té sobre el model formal del català nord-occidental, atès que les modificacions fòniques que s'hi observen depenen del coneixement de català dels parlants, del tipus d'escolarització que han rebut, i, relacionat amb tot això, de l'edat.[3]

3. Les variables de l'estudi

A l'hora de treballar amb els resultats, i per tal de poder aconseguir les mitjanes i les probabilitats d'aparició de cada segment fonètic, he hagut d'assignar uns valors a la *variable dependent* (que és la vocal pretònica) i a les *independents* (que són les que defineixen el marc d'aparició de la variable dependent, és a dir, els factors lingüístics i extralingüístics). A continuació presento les variables lingüístiques i extralingüístiques que inicialment[4] he considerat importants:

1) Les variables lingüístiques que he tingut en compte al principi de l'anàlisi han estat: el tipus de vocals pretòniques, la posició de l'accent de cada mot, el context consonàntic adjacent a la vocal pretònica, el punt i el mode d'articulació dels sons adjacents a la vocal pretònica, la vibració de les cordes vocals del so següent a la vocal, la qualitat de la síl·laba pretònica, l'etimologia de cada paraula, la síl·laba tònica de cada mot i el so corresponent a la posició pretònica del mateix mot traduït a l'espanyol.
2) Les variables extralingüístiques, que estan directament relacionades amb els emissors de cada mot, són: sexe, estatus sociocultural, coneixements de català escrit, estudis i edat.

Cadascuna d'aquestes variables conté diferents factors (per exemple, la variable *sexe* inclou els factors *dona* i *home*) i he assignat a cada factor un dígit o una lletra que el representa (d'aquesta manera, *dona* és representada per *d* i *home*, per *h*). El conjunt d'aquestes variables ha donat com a resultat una sèrie de codificacions que s'han utilitzat per qualificar cada emissió analitzada. Un dels exemples de codificació és el que segueix.

aa2cr4a2stbOch-1n9fA08836

L'explicació d'aquesta codificació és:

VARIABLE DEPENDENT
a: so pretònic emès [a]

VARIABLES INDEPENDENTS
a: tipus de vocal pretònica: inicial absoluta
2: posició de l'accent: vv'v
c: context adjacent del davant de la vocal: existent
r: punt d'articulació del davant de la vocal pretònica
4: mode d'articulació del davant de la vocal pretònica: líquid
a: punt d'articulació del darrere de la vocal: alveolar
2: mode d'articulació del darrere de la vocal: fricatiu
s: vibració de les cordes vocals del so de darrere de la vocal: sord
t: qualitat de la síl·laba pretònica: travada
b: etimologia: derivat amb la partícula *es-*
O: síl·laba tònica del mot
c: so inicial del mot corresponent en espanyol: consonant
h: sexe: home
-: context sociocultural: mitjà baix
1: nivell d'estudis: sense estudis
n: coneixements de català normatiu: no
9: edat: de 3 a 5 anys

[3] Per a un major aprofundiment sobre el tema vegeu Carrera (1999) i (2001).
[4] Presento aquesta consideració perquè després d'aplicar l'anàlisi estadística alguns factors no han estat rellevants en l'estudi i s'han descartat.

f: estil de parla: formal
A: comunitat lingüística: Alguaire
088: codi de la paraula: *estisores*
36: codi de l'informant

4. Aplicació del programa d'estadística Goldvarb[5]

El programa d'estadística Goldvarb té quatre fases indispensables, que també conté el programa Varbrul. Aquests quatre passos, que permeten obtenir l'estadística descriptiva, són:

a) Introducció dels resultats de les enquestes en el que s'anomena fitxer de *tokens* o dades.[6] Aquest fitxer és la base per realitzar totes les anàlisis i per això, òbviament, cal que tots els resultats hi estiguin introduïts correctament. Aquí es col·locaran totes les codificacions corresponents a cada emissió dels enquestats, similars a la que acabo d'explicar.
b) Un cop emplenat el fitxer de *tokens* (dades) cal configurar el fitxer de condicions on s'han d'introduir els paràmetres que han de combinar els resultats obtinguts amb les variables de l'anàlisi —dependent i independents. Segons l'estructura d'aquest fitxer es podrà explicar d'una manera o altra la variació lingüística. Val a dir, però, que aquest fitxer es veu força modificat al llarg de tota l'anàlisi, atès que cal buscar la millor combinació de factors per explicar bé el fenomen variable. Per això, es trobaran, entre altres modificacions, variables i factors no modificats, i factors i variables reagrupats.
c) Acabada la configuració del fitxer de condicions, el programa crea un fitxer d'iteració de les variables independents: és el que s'anomena *fitxer de cel·les*. En aquesta part el programa combina tots els factors independents (síl·laba tònica, edat dels informants, estudis, etc.) i els relaciona amb les dades reals de l'anàlisi. L'usuari del programa només ha de prémer unes instruccions a partir de les quals es creen les esmentades cel·les. Tanmateix, cada cop que es canvien les condicions, cal refer aquest fitxer de cel·les.
d) Al final, el programa presenta de manera automàtica els resultats de la combinació de les dades i de les condicions en quantitats reals i en percentatges.
En resum, en aquesta primera fase les dades (*tokens*) i les condicions amb tots els canvis pertinents han de ser introduïdes per l'usuari del programa. Les cel·les i els resultats, en canvi, són calculats pel programa en triar l'opció pertinent. Al final d'aquest primer estadi es compta amb l'estadística descriptiva i les incidències reals d'ús que s'han donat per a cada variable independent segons els factors assenyalats com a dependents.

Finalitzades les quatre fases de l'etapa inicial que acabo de descriure, es pot procedir a l'estadística d'inferència o anàlisi probabilística, que permet calcular: 1) l'aparició de la variable dependent en relació amb els factors de diverses variables explicatives o independents, és a dir, en el cas que m'ocupa, la probabilitat d'aparició de la variable dependent (solució [a]) quan es troba en síl·laba oberta o travada, quan la vocal tònica té un timbre o un altre, etc.; 2) l'aplicació del model teòric segons les dades obtingudes; això és, observar quina és la probabilitat general d'aparició de la vocal [a] en unes causes de varibales independents com ara la qualitat de la síl·laba pretònica i de la vocal tònica, l'edat, els estudis, etc. Tal com ja he explicat, la forma final d'aquesta part té una relació directa amb el fitxer de condicions.

El resultat de les iteracions de tots els factors individualment i en conjunt ve donat per l'*input* d'entrada de la regla o de l'aplicació del model teòric, que dóna la significació necessària en mostrar si les condicions seleccionades per explicar la variació són rellevants o no. Així, si l'*input* és més gran de 0.5 vol dir que els resultats considerats en conjunt faciliten l'aplicació de la regla; si és inferior, no. Per aconseguir observar l'*input* general i la probabilitat de manteniment o de canvi de la variable estudiada en relació amb els diferents factors independents, hi ha dues operacions complementàries:

-*Binomial Up & Down* (U&D)
-*Binomial 1 level* (1L)

U&D és un procediment d'anàlisi que té l'objectiu de cercar quines són les variables independents significatives per presentar la probabilitat d'aplicació de la regla variable. Mostra uns càlculs de regressió on s'analitzen tots els grups de factors independents, primer d'un en un, i després en combinació, fins que s'arriba a la composició de regla variable més versemblant. El resultat que dóna no és la probabilitat definitiva sinó el *pes (weight)* de cada factor independent d'acord amb els altres factors de l'anàlisi. Segons el pes de cada factor lingüístic o extralingüístic es pot saber en quina direcció es modificarà la variable dependent: si és més gran de 0.5 vol dir que influirà i, si no, significa que aquest factor no tindrà gaire a veure amb els canvis de la variable dependent.

A més, el càlcul U&D presenta la *significació de l'anàlisi* escollida com a bona en relació amb altres raonaments calculats anteriorment, ja que, abans d'arribar a l'anàlisi òptima, el programa ha realitzat diversos càlculs i proves com el logaritme de la versemblança (*log. likelihood*) i la prova de c2 (*X-square*), que mesura si són independents les variables que s'analitzen. Aquesta prova s'observa amb el resultat que dóna *p*, el qual ha de

[5] El programa que utilitzo és una versió anglesa i, per això, les anotacions que hi apareixen són en anglès.
[6] Aquests resultats estan codificats d'acord amb els paràmetres que he exposat anteriorment.

ser inferior a 0.005 perquè així es rebutja el que en estadística s'anomena la *hipòtesi nul·la*, és a dir, la hipòtesi que considera que la variació no ve donada pels factors independents escollits per explicar-la.

Un cop observada la millor combinació de factors, i tenint presents aquestes dades, es pot passar a observar la probabilitat definitiva de cada factor independent en la incidència de la regla variable (en el meu cas és de manteniment de la vocal pretònica) a partir de l'altre tipus d'anàlisi: *Binomial 1 level* (*1L*). *1L* presenta de manera relativament ràpida si el conjunt de paràmetres inicials -variables independents- és adequat per explicar la regla d'entrada. Vegem un exemple:

BINOMIAL VARBRUL, 1 step • 08/10/1•17:47 ••••••••••••••••••••••••••••••••••••
Name of cell file: 01.Cel

Using more accurate method.
Averaging by weighting factors.
One-level binomial analysis…

Run # 1, 117 cells:
Iterations: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Convergence at Iteration 15
Input 0.636

Group Factor		Weight	App/Total	Input&Weight
1:	t	0.568	0.66	0.70
	o	0.253	0.36	0.37
2:	i	0.349	0.40	0.48
	e	0.489	0.61	0.63
	a	0.565	0.65	0.69
	o	0.413	0.54	0.55
	u	0.624	0.74	0.74
3:	1	0.877	0.92	0.93
	2	0.574	0.73	0.70
	3	0.594	0.72	0.72
	4	0.292	0.37	0.42
	5	0.343	0.46	0.48
	6	0.236	0.37	0.35
4:	9	0.395	0.90	0.53
	7	0.430	0.40	0.57
	5	0.554	0.59	0.68
	2	0.565	0.80	0.69

Cell	Total	App'ns	Expected	Error
tu67	8	4	3.759	0.029
tu65	17	9	10.095	0.292
tu57	21	11	12.594	0.504 (…fins arribar a 117)

De les 20 iteracions possibles que preveu el programa, els factors inicials han trobat el seu punt òptim a la quinzena iteració. L'*Input* d'aplicació de la vocal pretònica [a] en aquests factors demostra que la regla «s'aplica» (0.636), és a dir, que, en termes generals, es manté la solució [a] ja que depassa el 0.5, xifra atribuïble a la mateixa aparició de tots dos factors dependents.

Si reprenem la fórmula del model logístic que regeix aquest programa podem observar d'on provenen els càlculs que es presenten en l'anàlisi d'*1L*. Així veurem que:

$$\frac{p}{1-p} = \frac{p_o}{1-p_o} \times \frac{p_i}{1-p_i} \times \frac{p_j}{1-p_j} \times \ldots$$

p = **Input & Weight** (és a dir, probab. de cada factor independent)

p_0 = ***Input*** la probab. de tota la regla variable. Aquí és **0.636**

p_i = ***Weight*** en el cas del factor *t* del grup *1* és **0.568**

p_j = ***Weight'*** en el cas del factor *o* del grup *1* és **0.253**

A partir d'aquesta fórmula podem trobar la probabilitat definitiva, en el meu cas, de manteniment de la solució [a], segons cada factor independent si relacionem l'*Input* i el *Weight* de cada factor de la següent manera:

$$\frac{p}{1-p} = \frac{p_o}{1-p_o} \times \frac{p_i}{1-p_i} \quad ; \; amb \quad p_o = 0.636 \; i \; p_i = 0.568$$

El resultat d'aquesta equació dóna el valor de *p* (*Input & Weight*), que és la probabilitat de manteniment de [a] segons la regla variable exposada:

En el cas del factor *t* del grup 1, p= 0.696, quantitat que, si s'arrodoneix, és la que correspon a la tercera columna de resultats que presenta el programa. És a dir, a 0.70.

D'aquesta manera, *1L* mostra, primer, l'*Input* general de manteniment de [a] (0.636), el pes o *Weight* de cada factor, el percentatge d'ús de [a] segons cada variable independent, que apareix amb el nom de *App/Total*, i, finalment, la probabilitat de manteniment de [a] en relació amb aquests factors independents, que és l'*Input & Weight*.

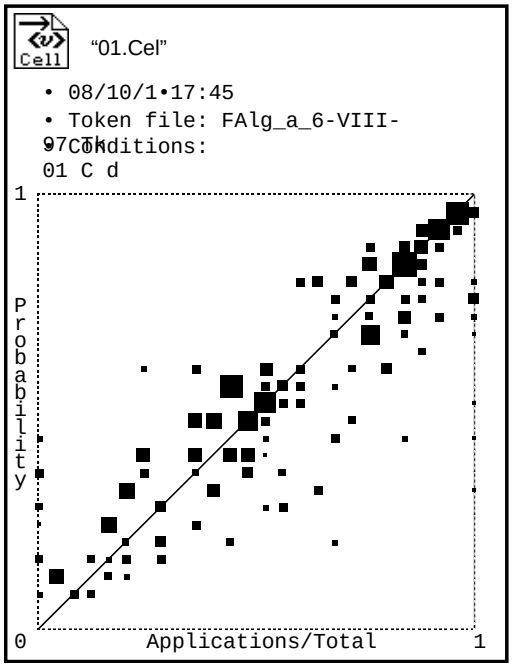
Un cop detallats els resultats, aquest programa mostra una descripció dels errors que hi ha entre la probabilitat teòrica o esperada i la mostra emprada. Cal tenir en compte que qualsevol anàlisi probabilística comporta la presència d'error; si no hi hagués error, el camp d'estudi seria la matemàtica funcional. Ara bé, la probabilitat cerca el marge més petit d'error entre allò esperat i les dades reals amb què treballa; d'aquesta manera, com més adequació hi hagi entre les columnes que presenten la freqüència (*App/Total*) i la probabilitat (*Input & Weight*), més garanties d'èxit presentarà l'anàlisi.

En aquesta part hi ha tres proves que expliquen si les condicions teòriques s'adeqüen a les dades de l'estudi:

-Logaritme de la versemblança (*Log. likelihood*)
-Prova de c2 (*X-square*)
-Diagrama de dispersió (*Scattergram*)

Vegeu l'exemple que es deriva de l'anàlisi anterior:

Total Chi-square = 150.8532
Chi-square/cell = 1.2893
Log likelihood = -1362.634
Maximum possible likelihood = -1277.040
Fit: X-square(104) = 171.187, rejected, p = 0.0000



La xifra que presenta el logaritme de la versemblança s'ha de posar en relació amb la quantitat màxima que proposa el programa amb el nom de *Maximum possible likelihood*. Si les dues xifres són similars, tenim ja una primera garantia de l'adequació teòrica i pràctica de la regla.

La segona prova és la de c2, que s'usa en moltes anàlisis estadístiques i que ja hem vist en l'*U&D*. La quantitat que dóna la pauta per saber si els resultats s'adeqüen a la teoria és la que segueix la lletra *p*. Com més propera sigui a 0, més fiable serà la regla.

Finalment, el diagrama de dispersió de les dades presenta de forma gràfica aquests resultats. Si els punts de convergència entre el model teòric i les dades reals segueixen la línia marcada del gràfic, el grau de confiança de l'anàlisi és molt elevat.

Així, a partir d'aquesta breu introducció, es pot veure que, per tal de tenir una anàlisi amb un grau de fiabilitat elevat amb el programa Goldvarb, és necessari presentar l'*Input* d'aplicació de la regla d'entrada i alguna de les proves que assegurin la versemblança del model teòric construït a partir de la iteració de variables independents. Després, es podrà procedir a interpretar les dades, la qual cosa haurà d'anar sempre en connexió amb la xifra 0.5. D'aquesta manera, es podrà determinar quan un factor afavoreix la presència de la regla d'entrada (en el cas que m'ocupa, el manteniment de la solució [a]) i quan no facilita que s'apliqui la regla inicial, tal com es pot observar en els factors *o* (de la variable 1) i *4*, *5* i *6* (de la variable 3).
Per acabar, és important tenir en compte que si bé l'anàlisi probabilística permet observar tendències de futur, no proporciona una anàlisi exhaustiva de totes les combinacions possibles de factors. Per això, moltes vegades és interessant completar l'anàlisi amb tabulacions creuades de factors, amb l'objectiu de detallar les connexions dels percentatges obtinguts en cada factor independent segons diferents creuaments de variables. Una de les aplicacions del programa que he descrit permet la creació de taules creuades de factors un cop s'han elaborat els fitxers de condicions, de cel·les i de resultats a través de la instrucció *cross tabulation*. D'aquesta manera, el programa Goldvarb permet realitzar anàlisis de diferents característiques partint de la mateixa classificació de dades inicial.

5. El programa Goldvarb en la sociolingüística variacionista

El variacionisme ha esmerçat esforços considerables per tal d'adaptar un mètode adequat a proporcionar anàlisis rigoroses i interpretacions lingüístiques fiables, i el programa Goldvarb n'és un dels seus resultats.
Aquest programa, tal com acabem de veure, permet obtenir, com d'altres, l'estadística descriptiva i arribar a presentar molt acuradament l'estadística d'inferències de fenòmens variables. Prèviament, però, cal conèixer els factors que són susceptibles de variació (que poden provenir tant dels components gramaticals foneticofonològic, morfològic, lèxic, semàntic o sintàctic) i, sobretot, saber trobar els factors lingüístics i extralingüístics que d'una manera o altra poden explicar els diferents fenòmens variables. Per tal d'arribar a aquestes intuïcions, caldran anàlisis qualitatives prèvies sobre la variació i la naturalesa dels diversos fenòmens per analitzar.
Si trobar els factors variables i explicatius de la variació és important, no és menys rellevant saber combinar-los adequadament per tal d'obtenir els resultats més representatius de la realitat variable, i això implica buscar i, en última instància, saber trobar la millor combinació de factors explicatius, que passa, molts cops, per refer les anàlisis una i una altra vegada.
En definitiva, doncs, el programa Goldvarb permet explicar qualsevol procés de variació, sempre que tingui un bon plantejament inicial que estigui acompanyat d'una sèrie de factors que en puguin donar raó.

6. Bibliografia

CARRERA-SABATÉ, J. (1999): *L'Alternança a/e al Segrià*. Tesi doctoral. UB. Barcelona.
CARRERA-SABATÉ, J. (2001): "La normativització del català modifica els hàbits fonètics dels parlants?" *Llengua i literatura*, 12: 175-199.
CEDERGREN, H. J.; SANKOFF, D. (1974): "Variable rules: Performance as a statistical reflection of competence". *Language*. 50: 333-355.
KAY, P.; McDANIEL, C. (1979): "On the logic of variable rules". *Language in Society*. 8: 151-187.
LABOV, W. (1969): "Contraction, Deletion, and Inherent Variability of the English Copula". *Language*. 45: 715-762.
LABOV, W. (1994): "*Principles of linguistic change. Internal factors"*. Blackwell. Cambridge.
LÓPEZ MORALES, H. (1989): *Sociolingüística*. Gredos. Madrid.
MORENO, F. (1994): "Status quaestionis: sociolingüística, estadística e informática". *Lingüística*. 6: 95-154.
SANKOFF, G. (1988): "Variable Rules". U. Ammon; N. Dittmar; K. J. Mattheier (ed.): *Sociolinguistics. An international handbook of the science of language and society*. Walter de Gruyter. Berlin & New York: 984-997.

Josefina Carrera-Sabaté
Universitat de Lleida
Universitat de Barcelona
jcarrera@filcat.udl.es