CAPÍTOL 11

FREQÜÈNCIA D'ÚS DE LES LLETRES

Índex Capítol

CAPÍTOL 11

FREQÜÈNCIA D'ÚS DE LES LLETRES

11.1 Introducció

No hem trobat en la literatura aplicacions en les que l'estudi de l'ús de les lletres hagi sigut útil en la determinació de l'autoria o en l'estudi de l'homogeneïtat de texts. Tot i així, les lletres tenen totes les característiques llistades en el capítol 2 de la tesi, que fan que una unitat textual sigui útil en estilometria:

> *"ser rellevants, estructurals, freqüents, fàcilment quantificables i relativament immunes al control conscient de l'escriptor."*

Estudiem l'ús de lletres en els capítols de més de 200 paraules. En primer lloc es fa una anàlisi de la relació entre els nombres i les proporcions de lletres i la llargada de capítol. A continuació s'estudia l'evolució temporal de les proporcions d'ús de les lletres mitjançant gràfics de control, i s'estima la localització del possible punt de canvi.

Hem comptat para cada capítol el nombre de vegades que apareixen cadascuna de les lletres. Ho hem fet seguint dos criteris diferents:

- considerant com a lletres les de l'alfabet. Hem considerat, per tant, 24 lletres, de les quals 5 són vocals.
- considerant com a lletres diferents les que tenen grafia diferent. Així *l*, *ll* i *l·l* són tres lletres diferents, i *e*, *é* i *è* són tres vocals diferents. Comptem, doncs, 36 lletres diferents, de les quals 14 són vocals.

Hem exclòs de l'estudi la lletra *k*, que apareix només 18 vegades en els capítols en estudi, concentrades entre els capítols 74 i 84.

11.2 Notació i forma de les dades

La taula 11.1 mostra el tros corresponent als capítols 1-44 de la taula que farem servir per a l'anàlisi de l'ús de lletres. La taula sencera té 425 files que corresponen als capítols de més de 200 paraules. La primera columna indica el número de capítol, les columnes centrals formen una taula de contingència, i cadascuna conté el nombre d'ocurrències d'una determinada lletra, la penúltima indica el nombre total de vocals i la darrera el nombre total de lletres en el capítol. A la taula podem llegir, per exemple, com en el primer capítol la lletra *a* apareix 125 vegades, la lletra *e* apareix 191 vegades i la *i* 76 vegades, i hi ha un total de 495 vocals i 1140 lletres.

Anomenem y_{ji} al nombre d'ocurrències de la lletra *j*-èssima, per $j=1,2,...l$, amb l igual al nombre total de lletres, en el capítol i-èssim. El vector $\boldsymbol{y_i}=(y_{1i},y_{2i},...,y_{li})$ es distribueix $Mult(N_i,\boldsymbol{\pi_i})$, on $\boldsymbol{\pi_i}=(\pi_{1i},\pi_{2i},...,\pi_{li})$ és el vector tal que π_{ji} és la probabilitat que una lletra presa a l'atzar en el capítol i-èssim sigui la *j*-èssima.

Capítol	a	e	i	o	u	...	Vocals	Lletres
1	125	191	76	62	41	...	495	1140
2	258	269	135	124	74	...	860	1974
3	541	707	251	384	200	...	2083	4771
4	388	363	161	184	112	...	1208	2770
5	524	625	268	324	187	...	1928	4457
6	308	339	165	167	109	...	1088	2435
7	141	196	73	76	53	...	539	1228
8	129	134	61	73	31	...	428	1004
9	116	148	70	64	48	...	446	999
10	453	517	190	223	142	...	1525	3469
11	300	317	139	164	79	...	999	2202
12	226	211	89	115	61	...	702	1624
14	572	638	221	306	168	...	1905	4404
15	201	302	106	155	68	...	832	1888
16	143	164	59	95	59	...	520	1187
17	314	417	146	213	107	...	1197	2780
18	280	325	111	153	97	...	966	2205
19	657	786	274	415	208	...	2340	5369
20	282	329	142	158	104	...	1015	2251
21	269	346	118	184	91	...	1008	2297
22	387	397	166	213	115	...	1278	2926
23	461	519	209	245	146	...	1580	3643
24	314	396	146	190	114	...	1160	2655
25	458	559	191	307	137	...	1652	3892
26	743	805	278	407	195	...	2428	5568
27	963	1148	442	593	325	...	3471	7819
28	227	265	118	110	74	...	794	1813
29	264	290	98	128	68	...	848	1995
30	138	130	58	55	32	...	413	957
31	150	169	73	85	46	...	523	1184
32	299	470	149	208	123	...	1249	2891
33	367	423	156	186	129	...	1261	2897
34	314	361	145	169	101	...	1090	2473
35	433	458	177	203	136	...	1407	3266
36	231	285	96	135	70	...	817	1940
37	147	188	85	91	49	...	560	1352
38	372	375	148	183	108	...	1186	2674
39	471	619	234	281	155	...	1760	4074
40	128	150	60	70	45	...	453	1033
41	426	557	215	256	151	...	1605	3720
42	335	443	149	187	94	...	1208	2790
43	165	185	63	76	32	...	521	1196
44	428	467	156	188	77	...	1316	3099

Taula 11.1: Part corresponent als capítols 1-44 de la taula de contingència per l'ús de lletres. Cada fila correspon a un capítol, la primera columna dóna el número de capítol, les columnes centrals contenen el nombre d'ocurrències per a les cinc vocals, la penúltima el nombre total de vocals i la darrera el nombre total de lletres en el capítol. Els capítols que no apareixen a la taula tenen una llargada inferior a les 200 paraules.

11.3 Relació de la proporció de lletres amb Ni

Anomenem $\hat{\pi}_{ji} = y_{ji}/N_i$ a la proporció d'ocurrències de la lletra j-èssima, per $j=1,2,...l$, i π_{ji} al seu valor esperat, de manera que:

$$E\left(\hat{\pi}_{ji} = \frac{y_{ji}}{N_i}\right) = \pi_{ji}.$$

Sota la hipòtesi de independència, y_{ji} té distribució Binomial(N_i,π_{ji}). Per tant:

$$E(y_{ji}) = N_i \pi_{ji},$$

$$Var(y_{ji}) = N_i \pi_{ji}(1-\pi_{ji}),$$

$$Var\left(\frac{y_{ji}}{N_i}\right) = \frac{\pi_{ji}(1-\pi_{ji})}{N_i}.$$

A nivell empíric, la dependència de $E(y_{ji})$ i de $Var(y_{ji})$ amb N_i s'observa en la figura 11.1: en el gràfic de dalt a l'esquerra es mostra com, per la lletra *a,* hi ha una relació lineal molt forta entre el nombre d'ocurrències, y_{ji}, i la llargada del capítol, N_i, i el gràfic de la dalt a la dreta mostra una relació lineal molt forta entre y_{ji} i el nombre de lletres en el capítol. Els gràfics de baix mostren la relació entre y_{ji}/N_i i N_i i de $y_{ji}/N_{Lletres,i}$ amb $N_{Lletres}$, el nombre de lletres en el capítol, per la mateixa lletra. En el gràfic s'observa com la variança de y_{ji}/N_i creix en disminuir N_i. Gràfics semblants s'obtenen amb les altres lletres.

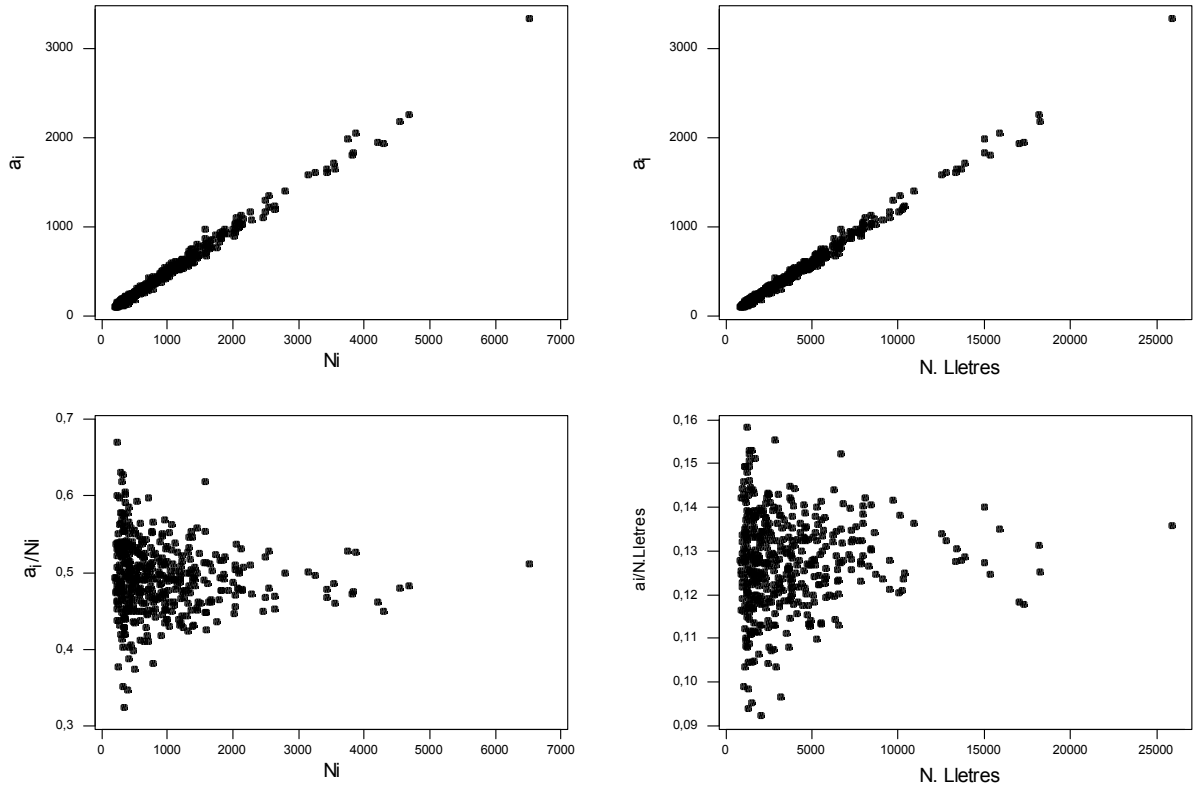

Figura 11.1: Relació del nombre d'ocurrències de la lletra *a*, a_i, amb la llargada del capítol, N_i (dreta) i amb el nombre de lletres en el capítol (esquerra) a dalt, i de la proporció a_i/N_i amb N_i (dreta) i de $a_i/N_{Lletres,i}$, amb $N_{Lletres}$, el nombre de lletres en el capítol, (esquerra), per la mateixa lletra a baix.

11.4 Evolució temporal de les seqüències multinomials

Es pot representar una seqüència de dades multinomials en un gràfic de control Chi-quadrat, tal i com s'ha explicat en el capítol 5 i com s'ha fet per la llargada de paraula en el capítol 8. En la figura 11.2 s'han representat els gràfics Chi-quadrat per a les seqüències multinomials formades per:

- les 24 lletres que formen l'alfabet, exclosa la *k*, a dalt a l'esquerra
- les 5 vocals, a dalt a la dreta
- les 36 lletres amb grafia diferent, exclosa la *k*, a baix a l'esquerra
- les 14 grafies diferents de les vocals, a baix a la dreta.

Sota la hipòtesi d'homogeneïtat, $\pi_{ji}=\pi_j$ per $i=1,2,...,n$, i s'ha pres com a estimació de π_j, $\hat{\pi}_j$, la mitjana de les proporcions de la lletra j-èssima per a tots els capítols de més de 200 paraules:

$$\hat{\pi}_j = \frac{1}{n}\sum_{i=1}^{n}\frac{y_{ji}}{N_i},$$

i s'ha monitoritzat l'estadístic:

$$X_i^2 = \sum_{j=1}^{l}\frac{(y_{ji} - N_i\hat{\pi}_j)^2}{N_i\hat{\pi}_j}$$

amb $l=24,\ 5,\ 36$ i 14. De l'estudi dels gràfics no observem cap tret que pugui indicar manca de homogeneïtat en el text.

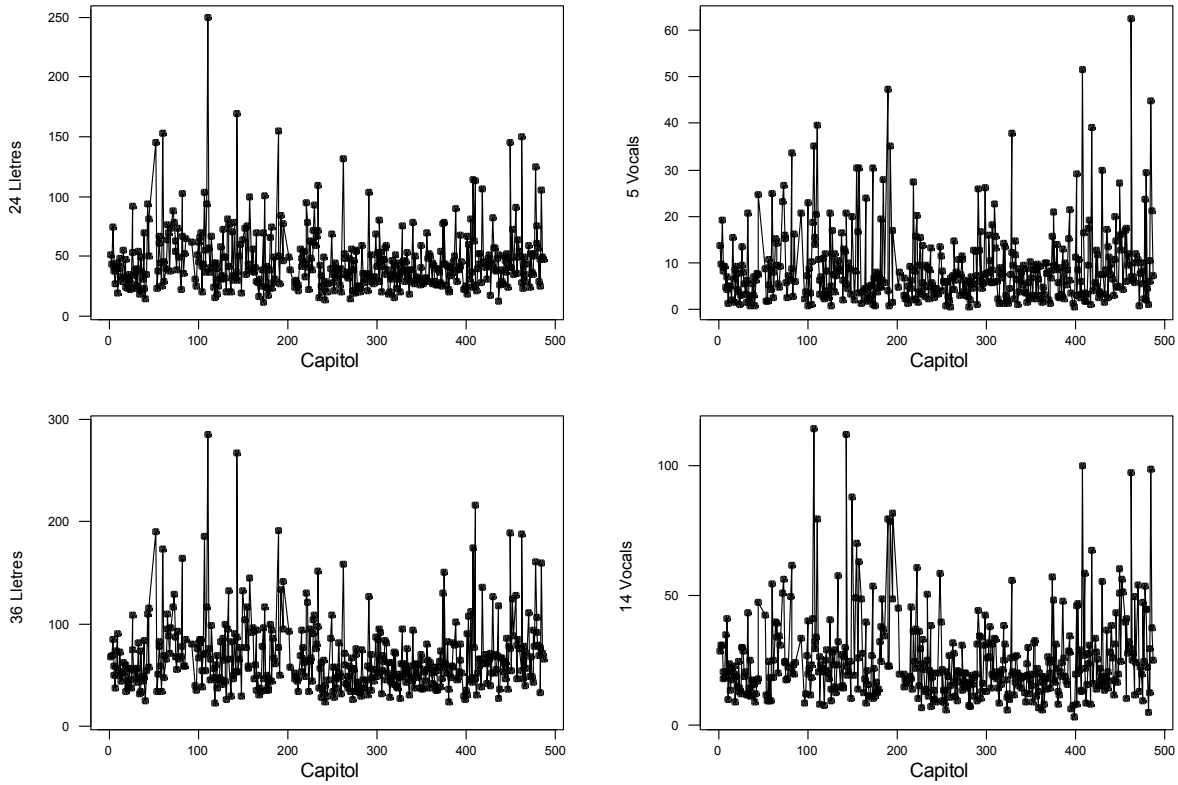

Figura 11.2: Gràfic Chi-quadrat per les seqüències multinomials formades per les 24 lletres (esquerra) i 5 vocals (dreta) de l'alfabet a dalt i 36 lletres (esquerra) i 14 vocals amb grafia diferent (dreta) a baix.

Una segona possibilitat per estudiar l'evolució temporal de la seqüència multinomial consisteix en fer l'anàlisi de correspondències i representar el valor de la primera component en funció del capítol. La figura 11.3 conté aquests gràfics per a les quatre seqüències multinomials abans esmentades. En ells tampoc no s'observa cap canvi en el nivell o en la variabilitat que faci pensar en un canvi en l'estil.

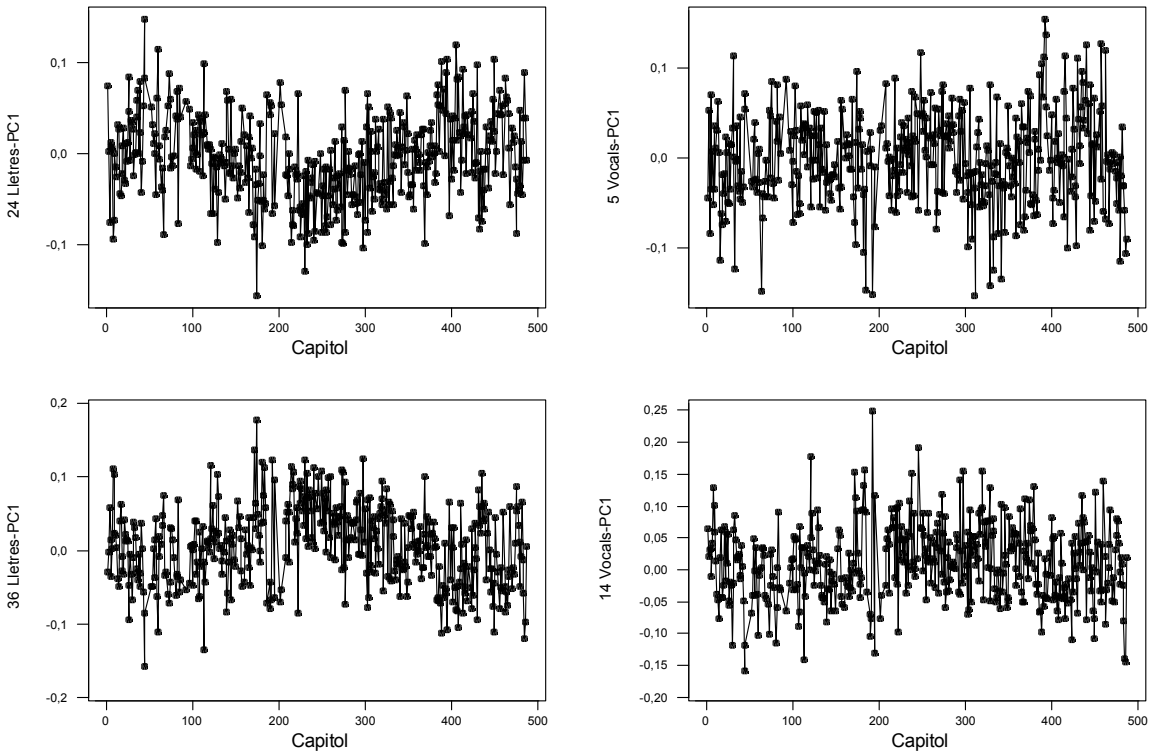

Figura 11.3: Gràfic en seqüència temporal per a la primera component de l'anàlisi de correspondències per les seqüències multinomials composades per les 24 lletres (esquerra) i 5 vocals (dreta) de l'alfabet a dalt i 36 lletres (esquerra) i 14 vocals (dreta) amb grafia diferent a baix.

11.5 Estimació del punt de canvi en seqüències multinomials

Emprenem l'estudi d'un possible punt de canvi, tot i que l'estudi gràfic no reveli de forma clara la seva existència.

Per a l'estimació del punt de canvi en primer lloc es fa servir el mètode proposat per seqüències de Normals, i s'aplica a les seqüències dels valors de la primera component de l'anàlisi de correspondències en el quatre casos estudiats a 11.4:

- les 24 lletres que formen l'alfabet, exclosa la *k*,
- les 5 vocals,
- les 36 lletres amb grafia diferent, exclosa la *k*,
- les 14 grafies diferents de les vocals.

A continuació, s'estima el punt de canvi en la seqüència binomial de proporcions vocals. Per últim, s'estima el punt de canvi en les quatre seqüències multinomials esmentades fent servir el model politòmic.

11.5.1 Punt de canvi en seqüències Normals

S'ha estimat el punt de canvi per la primera component dels quatre anàlisis de correspondències fent servir les tècniques proposades per a dades Normals a la secció 6.2.

Sota les hipòtesi de independència i Normalitat, el valor de la primera component, pc_i, es distribueix:

$$pc_i \sim N(\mu = \beta_0^{(r)} + \beta_1^{(r)} Ind_{1i}^{(r)}, \sigma_i^2 = \frac{\sigma^2}{N_i}),$$

per *r=1,2,..., 424*, on $Ind_{1i}^{(r)}$ és una variable indicadora que pren valor *0* per *i=1,2,..,r* i pren valor *1* per *i=r+1,...,n=425*. S'ajusten els *n-1* models, per *r=1,...,n-1*, i estimem el punt de canvi com**:**

$$\hat{r}_N = \max_{1 \le r \le n-1} F_r$$

on F_r és el valor de l'estadístic *F* de la taula ANOVA pel model lineal amb pesos amb variable indicadora $Ind_{1i}^{(r)}$, ajustat fent servir el criteri dels mínims quadrats ponderats amb $w_i=N_i$.

El gràfic de la figura 11.4 mostra l'evolució de F_r en funció de *r* per a les quatre seqüències. S'obtenen estimacions diferents en els quatre casos:

- per a la seqüència de les 24 lletres que formen l'alfabet, $\hat{r}_N$=119, tot i que existeix un segon màxim per *r*=382.
- per a la seqüència de les 5 vocals, el màxim es troba a $\hat{r}_N$=71.

- per a la seqüència de les 36 lletres amb grafia diferent, $\hat{r}_N$=382, i existeix un segon màxim per *r*=114.
- per a la seqüència de les 14 grafies diferents de les vocals, el màxim es troba a $\hat{r}_N$=169. Aquí també hi ha un màxim local per *r*=381.

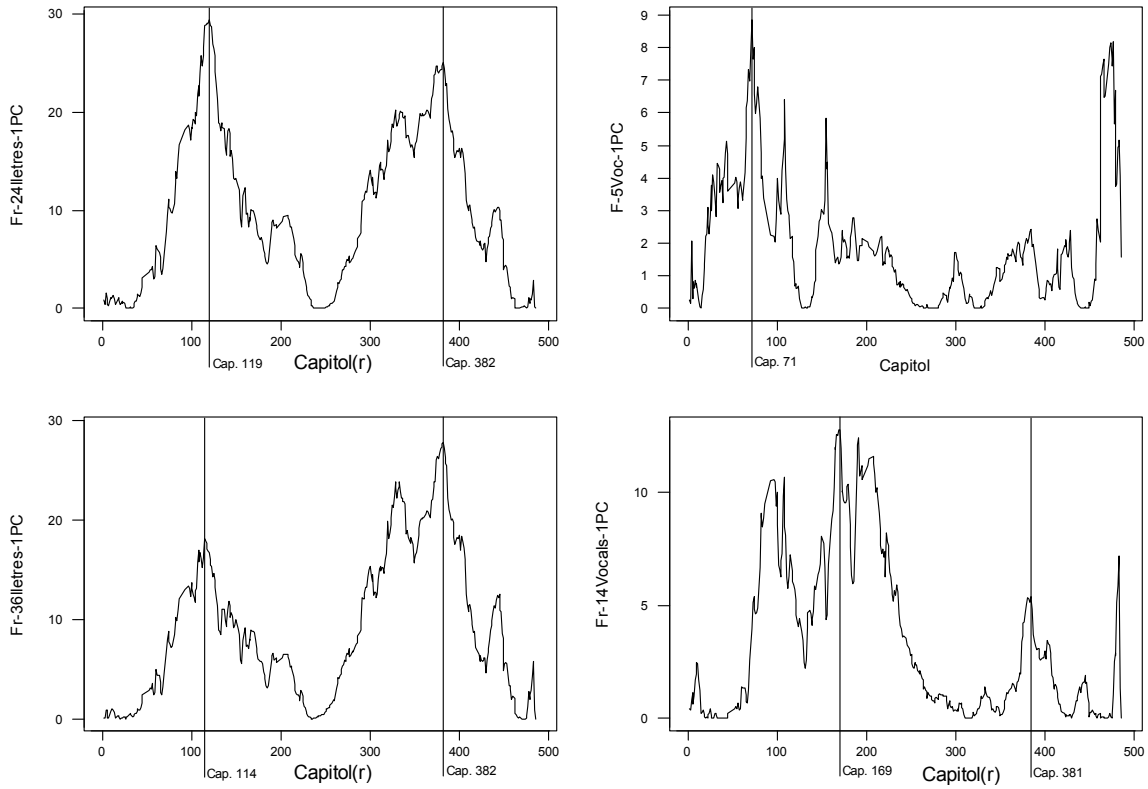

Figura 11.4: Gràfic de F_r en funció de *r* per les seqüències multinomials composades per les 24 lletres (esquerra) i 5 vocals (dreta) de l'alfabet a dalt i 36 lletres (esquerra) i 14 vocals (dreta) amb grafia diferent a baix.

11.5.2 Punt de canvi en seqüència binomial

En aquest apartat s'estima el punt de canvi en la seqüència binomial del nombre de vocals en el capítol, nv_i.

Suposant que existeix un punt de canvi a *r* en la seqüència, nv_i té una distribució *binomial* (N_i, π_{ja}) per $i=1,...,r$ i *binomial* (N_i, π_{jd}) per $i=r+1,...,n$. S'ajusta el model logístic (6.2):

$$nv_i \sim Binomial\left(N, \pi_i = \frac{e^{\beta_0^{(r)} + \beta_1^{(r)} Ind_{1i}^{(r)}}}{1 + e^{\beta_0^{(r)} + \beta_1^{(r)} Ind_{1i}^{(r)}}} \right).$$

La figura 11.5 conté el gràfic de $L\left(r, \hat{\beta}^{(r)}\right)$, el màxim del logaritme de la versemblança, en funció de *r*. El màxim del gràfic es troba a $\hat{r}_L$=99.

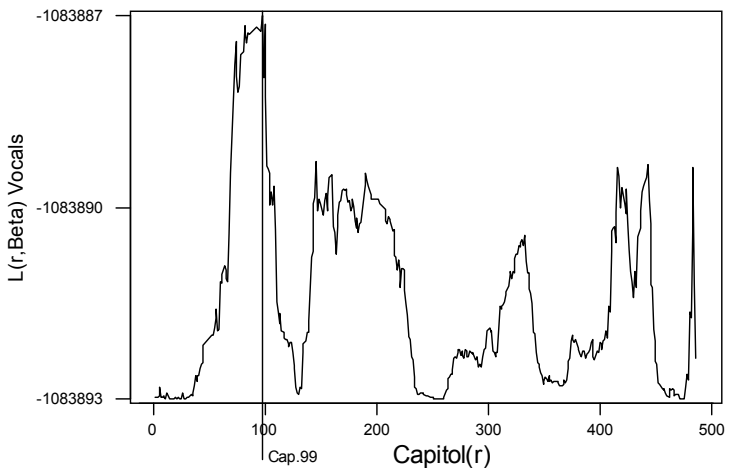

Figura 11.5: Evolució del màxim del logaritme de la versemblança del model logístic en funció de *r* per a la seqüència del nombre de vocals. L'estimació del punt de canvi és $\hat{r}_L$=99.

11.5.3 Punt de canvi en les seqüències multinomials; model politòmic

Recordem que y_{ji} és el nombre d'ocurrències de la lletra *j*-èssima en el capítol *i*, per *j=1,2,...l*, amb *l=5, 24, 14* o *36*, segons quina de les quatre seqüències analitzem. $\boldsymbol{y_i}=(y_{1i},y_{2i,...}y_{li})$ es distribueix $Mult(N_i,\boldsymbol{\pi_i})$, on $\boldsymbol{\pi_i}=(\pi_{1i},\pi_{2i},...,\pi_{li})$ és el vector tal que π_{ji} és la probabilitat que una lletra presa a l'atzar en el capítol i-èssim sigui la *j*-èssima.

El model politòmic suposa que:

$$g_j(\pi)=\log\left(\frac{\pi_{ji}}{\pi_{1i}}\right)=\beta_{j0}^{(r)}+\beta_{j1}^{(r)}Ind_{1i}^{(r)},$$

per *j=1,...,l*, i on $g_j(\cdot)$ suposarem que és una generalització de la funció *link logit* emprada per dades binomials al cas multinomial, descrita a l'apartat 6.3.1. En general tindrem:

$$g(\pi_i)=(g_2(\pi_i),....,g_l(\pi_i)).$$

on $g_j(\cdot)$, per *j=2,...,l*, podria ser qualsevol funció link que pugui ser utilitzada per a modelar dades binomials. S'estima el punt de canvi, *r*, com aquell en el que el màxim del logaritme de la versemblança:

$$L\left(r,\hat{\beta}^{(r)}\right)=\sum_{i=1}^{n}\left(\sum_{j=2}^{l}y_{ji}\left(\hat{\beta}_{0j}^{(r)}+\hat{\beta}_{1j}^{(r)}Ind_{1i}^{(r)}\right)-\ln\left(1+\sum_{j=2}^{l}e^{\left(\hat{\beta}_{0j}^{(r)}+\hat{\beta}_{1j}^{(r)}Ind_{1i}^{(r)}\right)}\right)\right),$$

és màxim. El gràfic de la figura 11.6 mostra l'evolució de $L\left(r,\hat{\beta}^{(r)}\right)$ en funció de *r* per a les quatre seqüències multinomials estudiades. S'obtenen estimacions diferents en els quatre casos, i diferents a les obtingudes a 11.5.1 i 11.5.2:

- per a la seqüència de les 24 lletres que formen l'alfabet, $\hat{r}_M$=300.
- per a la seqüència de les 5 vocals, el màxim es troba a $\hat{r}_M$=371.
- per a la seqüència de les 36 lletres amb grafia diferent, $\hat{r}_M$=299, tot i que existeix una zona fins a *r*=369 en la que el valor màxim del logaritme de la versemblança roman pràcticament constant.

- per a la seqüència de les 14 grafies diferents de les vocals, el màxim es troba a $\hat{r}_M$ =368. Un valor molt semblant per al màxim del logaritme de la versemblança s'obté per *r*=383.

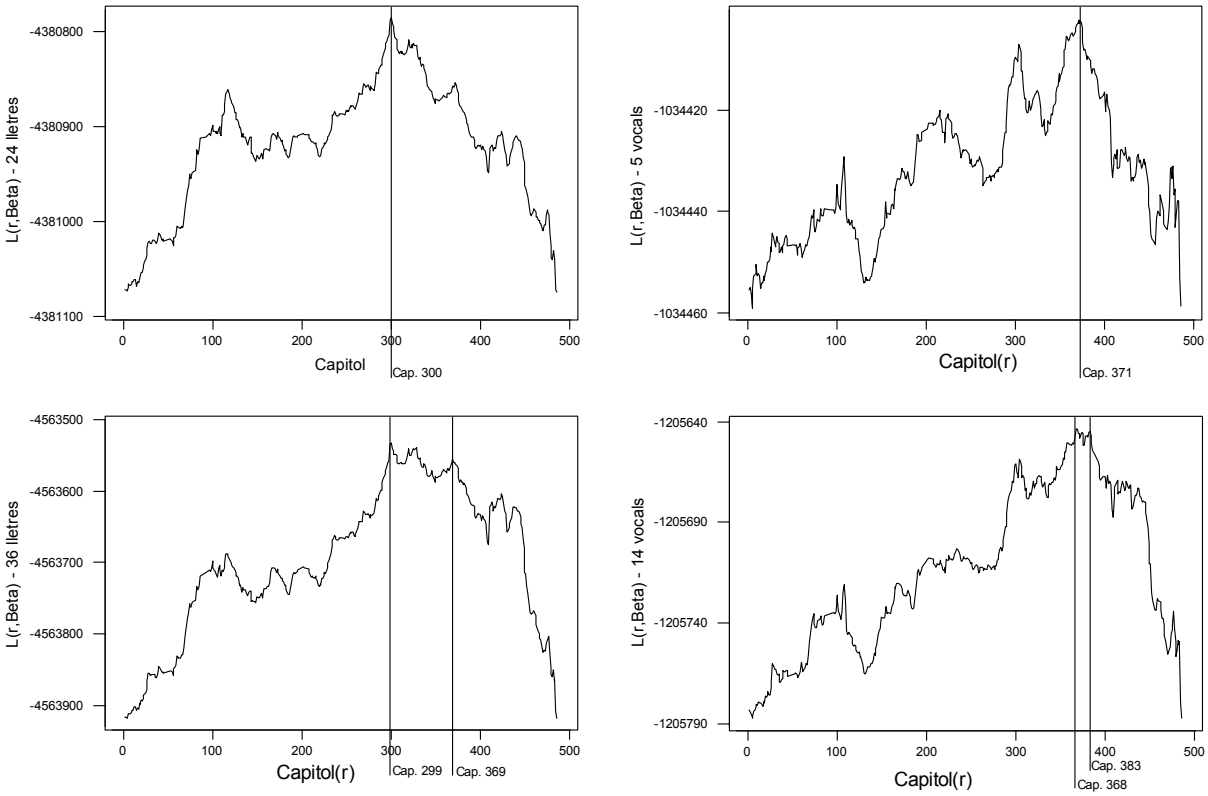

Figura 11.6: Evolució del màxim del logaritme de la versemblança en funció de *r* per les seqüències multinomials composades per les 24 lletres (esquerra) i 5 vocals (dreta) de l'alfabet a dalt i 36 lletres (esquerra) i 14 vocals (dreta) amb grafia diferent a baix.

11.6 Conclusions

De l'anàlisi gràfica de la seqüència temporal de les proporcions de lletres no és evident que hi hagi un canvi que pugui senyalar una manca de homogeneïtat en l'estil. A més a més, l'estudi analític del punt de canvi dóna estimacions diferents segons la seqüència estudiada. La majoria d'aquestes estimacions donen punts de canvi després del capítol 300, i algunes coincideixen amb els resultats que s'obtenen en l'estudi d'altres unitats d'estadística textual:

- Els dos màxims, a *r=119* i *r=382*, de F_r per l'estudi de la primera component de l'anàlisi de correspondències per les seqüències de composades per les 24 lletres i les 36 lletres amb grafia diferent coincideixen amb els que s'obtenen en l'estudi de la Diversitat del llenguatge al capítol 13,
- Els màxims del màxim del logaritme de la versemblança en funció de *r* per les seqüències multinomials composades per les 5 vocals de l'alfabet i les 36 lletres i 14 vocals amb grafia diferent són força semblants als que s'obtenen de l'estudi de la llargada de paraula al capítol 8 i de les paraules més freqüents al capítol 12.