CAPÍTOL 8

LLARGADA DE PARAULA

Índex Capítol

CAPÍTOL 8

LLARGADA DE PARAULA

8.1 Introducció

Tal i com s'ha comentat en la descripció de la base de dades del capítol 4, la llargada de paraula es mesura en nombre de lletres. L'estudi s'ha fet tant pels capítols de més de 200 paraules com pels blocs de 1000 paraules consecutives: s'ha fet primer una anàlisi exploratòria de la llargada mitjana i de les proporcions de paraules en cadascuna de les 10 categories en que s'ha classificat les paraules segons la seva llargada, per observar si es detecta una frontera d'estil. Aquesta anàlisi fa servir tant gràfics de control com l'Anàlisi de Correspondències. Pel cas dels capítols, a més, s'ha estudiat la relació entre les quantitats analitzades i la llargada de capítol.

A continuació s'estima el punt de canvi i es realitza una Anàlisi Cluster amb l'objectiu d'obtenir una agrupació dels capítols o blocs en 2 conjunts homogenis, a partir dels perfils fila descrits en el capítol 5.

8.2 Estudi de la llargada de paraula per capítol

Les variables que s'han obtingut de la quantificació de la llargada de paraula per als capítols són:

- el nombre de paraules i la proporció que representa respecte al total del capítol, per paraules de *i* lletres, per *i=1,2,...,17*. Les paraules de llargada 10 i superior s'han agrupat en una sola categoria, que anomenem *10+*;
- la llargada mitjana de paraula per als capítols,

En el cas dels capítols, i a diferència del que passa pels blocs, la llargada del capítol N_i no és constant. La distribució del nombre de paraules en cada categoria depèn molt de la llargada de capítol. Per a les anàlisis s'han considerat només els capítols de més de 200 paraules.

La taula 8.1 mostra el tros corresponent als capítols 321-366 de la taula que farem servir per a l'anàlisi de la llargada de paraula pels capítols. La taula sencera té 425 files que corresponen als capítols de més de 200 paraules, i 13 columnes: la primera columna indica el número de capítol, la segona el nombre total de paraules, N_i i la darrera columna conté la llargada mitjana de paraula per capítol. Les deu columnes centrals formen una taula de contingència que conté el nombre d'ocurrències en cadascuna de les *l=10* categories en que s'han agrupat les paraules en funció de la seva llargada, per cada capítol. La columna que conté el nombre de paraules de deu lletres o superior s'ha etiquetat com "*10+*". A la taula podem llegir, per exemple, com pel capítol *321* el nombre total de paraules és N_i=1454, hi ha 146 paraules d'una lletra, 315 de dues lletres, 340 paraules de tres lletres, 41 de nou lletres i 33 paraules de deu o més lletres.

Anomenem y_{ji} al nombre de paraules de *j* lletres, per *j=1,2,...,10+*, en el capítol i-èssim, $\overline{ll}_i$ a la mitjana de les llargades en el capítol i-èssim, i *ll* a la variable aleatòria llargada d'una paraula. El vector de comptatges $\boldsymbol{y_i}=(y_{1i},y_{2i,...},y_{10+i})$ es distribueix $Mult(N_i, \pi_i)$, on $\pi_i=(\pi_{1i},\pi_{2i},...,\pi_{10+i})$ és el vector tal π_{ji} és la probabilitat que una paraula presa a l'atzar en el capítol i-èssim tingui *j* lletres.

Cap.	N_i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10+**	$\overline{ll}_i$
321	1454	146	315	340	135	159	144	81	60	41	33	3,960
322	448	40	107	104	41	45	41	30	18	10	12	3,971
323	569	46	136	133	47	60	55	35	31	11	15	4,023
324	705	57	196	135	88	60	59	39	34	17	20	3,935
325	570	48	146	106	57	51	58	37	25	23	19	4,137
326	289	24	67	59	29	34	19	12	23	14	8	4,194
327	615	51	156	138	59	56	51	31	30	26	17	4,018
328	775	94	128	188	66	82	76	53	44	25	19	4,114
329	613	67	139	130	53	66	56	36	35	20	11	3,987
330	1092	121	228	232	110	110	106	72	56	37	20	4,026
331	478	46	118	87	61	67	38	23	19	9	10	3,866
332	279	17	72	68	16	34	37	13	11	8	3	3,957
333	693	61	153	155	78	79	66	30	32	26	13	4,006
334	1619	182	320	335	187	180	154	98	76	54	33	4,039
335	701	77	135	158	90	82	64	35	27	24	9	3,913
336	216	12	63	51	17	21	21	5	12	3	11	4,005
337	1033	111	220	228	106	117	119	46	28	30	28	3,942
338	269	24	54	59	22	24	28	22	18	8	10	4,327
339	970	100	207	225	99	94	111	71	32	15	16	3,894
340	2083	241	353	476	235	237	225	113	85	79	39	4,054
341	252	22	61	61	25	17	29	10	14	10	3	3,925
343	1338	147	241	308	142	157	140	81	58	35	29	4,030
344	954	109	203	185	108	107	94	61	43	28	16	3,980
345	785	84	163	186	86	72	80	38	29	28	19	3,948
346	479	48	107	107	51	33	41	29	34	9	20	4,100
347	710	65	145	167	76	74	62	41	30	26	24	4,107
348	235	26	57	36	22	28	18	20	15	8	5	4,115
349	2269	234	446	446	246	253	254	126	104	84	76	4,193
350	1117	121	232	240	104	122	120	54	49	38	37	4,106
351	428	42	96	96	30	49	33	27	33	10	12	4,107
352	217	23	43	40	22	27	24	17	8	6	7	4,207
353	738	71	156	156	86	60	58	42	55	31	23	4,224
354	852	77	198	188	92	73	71	39	46	44	24	4,112
355	1556	145	361	313	154	140	163	102	81	47	50	4,140
356	428	29	106	98	35	41	40	26	31	12	10	4,129
357	1002	87	231	217	102	96	105	50	42	31	41	4,142
358	279	21	67	58	38	19	26	18	16	9	7	4,100
359	311	38	67	69	27	32	27	21	19	5	6	3,900
360	416	37	91	89	45	34	45	26	24	9	16	4,188
362	446	44	87	100	43	35	58	24	23	15	17	4,226
363	244	21	53	46	40	31	20	11	6	10	6	4,016
364	349	24	84	80	34	38	30	11	26	10	12	4,132
365	261	18	69	54	27	29	22	12	17	5	8	4,061
366	579	68	134	105	69	52	66	31	20	14	20	3,991

Taula 8.1: Part de la taula de contingència per la llargada de paraula pels capítols de més de 200 paraules: cada fila correspon a un capítol, la primera columna dóna el número de capítol, la segona el nombre de paraules per capítol, les següents el nombre de paraules de *j* lletres, per *j=1,2,...,10+*, i la darrera conté la llargada mitjana de paraula per capítol.

8.2.1 Relació de la distribució de llargades amb N_i

Tant la llargada mitjana de paraula, $\overline{ll}_i$, com les proporcions de paraules de llargada *j*,

$$\hat{\pi}_{ji} = \frac{y_{ji}}{N_i},$$

per *j=1,2,...,10+*, són estimadors no esbiaixats del valor esperat de la llargada mitjana i de π_{ji}. La variança dels estimadors no és constant, perquè depèn de la llargada del capítol, N_i.

Essent la llargada mitjana de paraula un promig de més de 200 observacions, pel teorema central del límit, es té que:

$$\overline{ll}_i \underset{Aprox.}{\sim} N\left(E(ll_i), \frac{Var(ll_i)}{N_i}\right),$$

on ll_i és la variable aleatòria llargada d'una paraula del capítol i-èssim, i la proporció, $\hat{\pi}_{ji}$, és tal que:

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

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

Observem com els valors esperats de ll_i i de $\hat{\pi}_{ji}$ no depenen de la llargada del capítol N_i mentre que la variança d'ambdós sí que en depèn, i tant el valor esperat com la variança de ll_i i de $\hat{\pi}_{ji}$ depenen del capítol, *i*.

A nivell empíric això s'observa en la figura 8.1: la variança de la llargada mitjana de paraula per capítol és proporcional a $1/N_i$, és a dir, que és més gran pels capítols més curts i més petita pels més llargs, mentre que el seu valor esperat no es veu afectat per la llargada del capítol. El segon gràfic de la figura 8.1 mostra la relació entre y_{1i} i N_i, com a exemple de la relació lineal entre el nombre d'ocurrències en una categoria i la llargada del capítol, mentre que el tercer mostra la relació entre $\hat{\pi}_{1i}$ i N_i, en el que, de nou, es veu com el valor esperat de la proporció no depèn de la llargada del capítol mentre que la seva variança depèn de N_i.

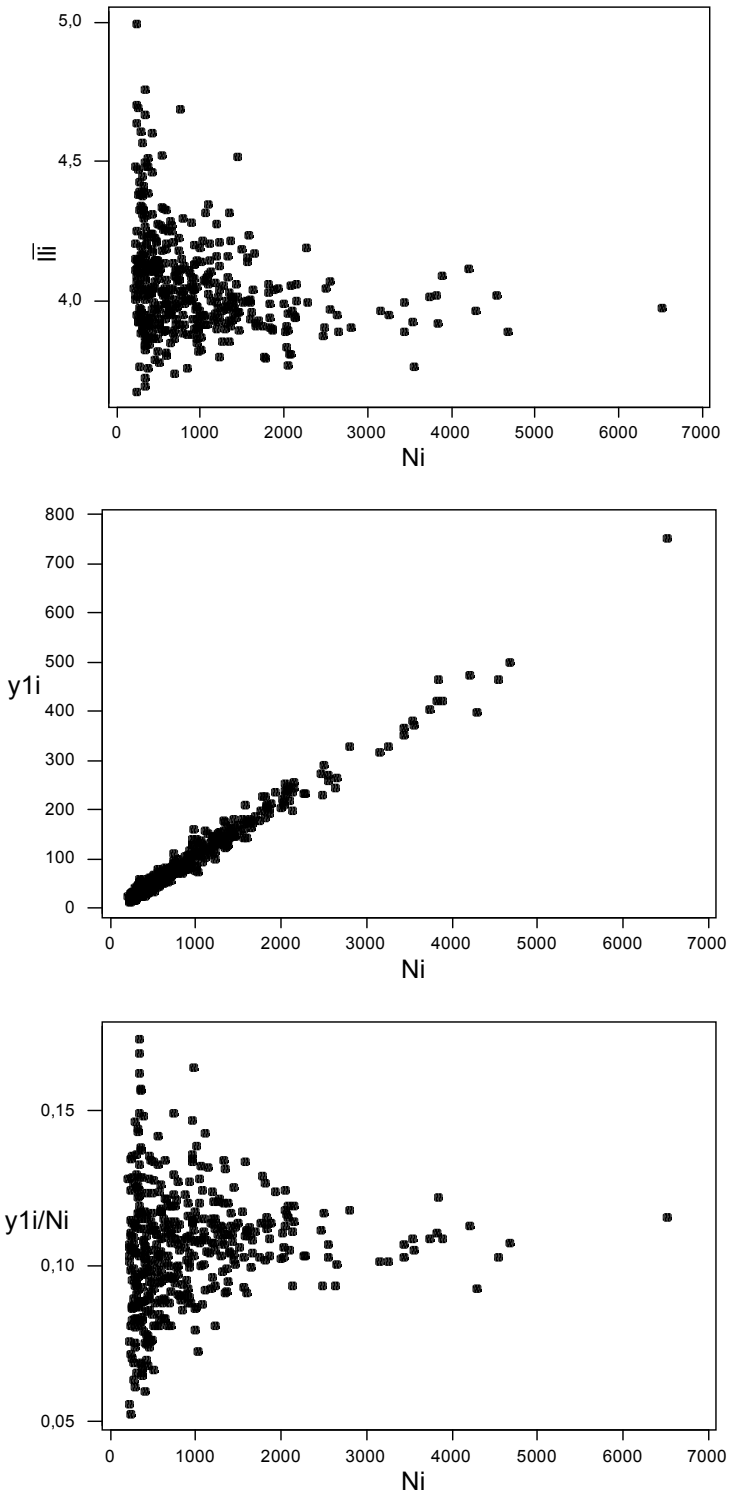

Figura 8.1: Relació de la llargada mitjana de paraula per capítol, del nombre de paraules de llargada 1 i de la proporció de paraules d'una lletra per capítol amb la llargada del capítol N_i.

8.2.2 Gràfics de Control per la llargada de paraula

A la figura 8.2 hi ha representada l'evolució de la llargada mitjana de paraula al llarg del *Tirant*, que ha de donar una idea de si l'estil és homogeni o be hi ha algun canvi.

S'observa com al voltant del capítol 345 canvia el valor esperat de $\overline{ll}_i$: pels capítols anteriors la majoria de les llargades mitjanes de paraula es troben lleugerament per sota de la línia central mentre que per els capítols posteriors la majoria dels valors cauen molt per sobre de la mitjana global. Aquest canvi ha estat marcat en el gràfic amb un canvi de color: els capítols des del primer al 345 estan en color verd, mentre que els que van del 346 al final estan en color vermell.

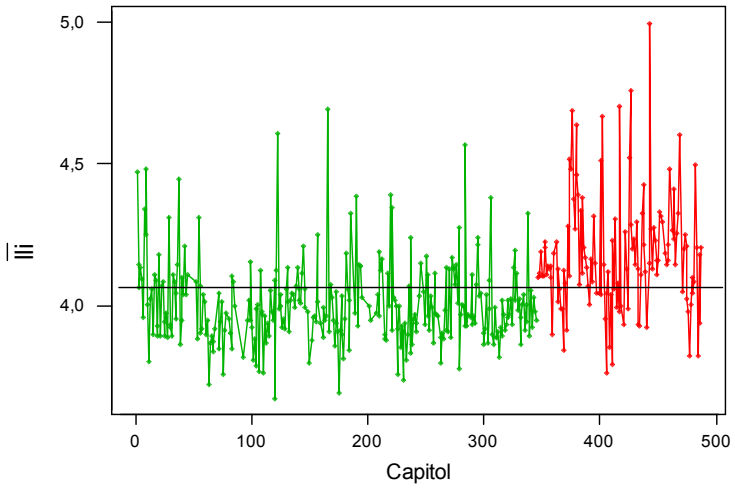

Figura 8.2: Gràfic de l'evolució temporal de la llargada mitjana de paraula per als capítols de més de 200 paraules. El canvi de color es troba entre els capítols 345 i 346.

A la figura 8.3 es presenta el gràfic que fem servir per estudiar les evolucions temporals de les proporcions de paraules de j lletres, $\hat{\pi}_{ji}$, per $j=1,...10+$. Sota la hipòtesi de que no hi ha variacions estilístiques al llarg del *Tirant*, el valor esperat de la proporció de paraules en cada categoria serà constant. El canvi de colors s'ha mantingut en el capítol 345**.**

La categoria que mostra més clarament el canvi al voltant del capítol 345 és la de paraules de deu o més lletres. El gràfic de les proporcions de paraules de nou lletres mostra també amb claredat el canvi, tot i que es troba retardat fins al capítol 371. Per paraules de vuit lletres , el canvi és menys evident que per a les anteriors, i es pot localitzar en el capítol 345. Per paraules de vuit, nou i deu o més lletres la proporció augmenta després del canvi. En canvi, per paraules de llargades intermitges, de 4 a 7 lletres, els gràfics de proporcions no mostren canvis clars en el nivell. Pel que fa a les paraules de llargades curtes, en les proporcions de paraules de una, dues i de tres lletres el canvi es troba en el capítol 371. Tant per llargada dos com tres després del canvi les proporcions són inferiors que abans del canvi. Per llargada u el canvi de proporcions es dóna en sentit contrari, augmentant després del punt de canvi.

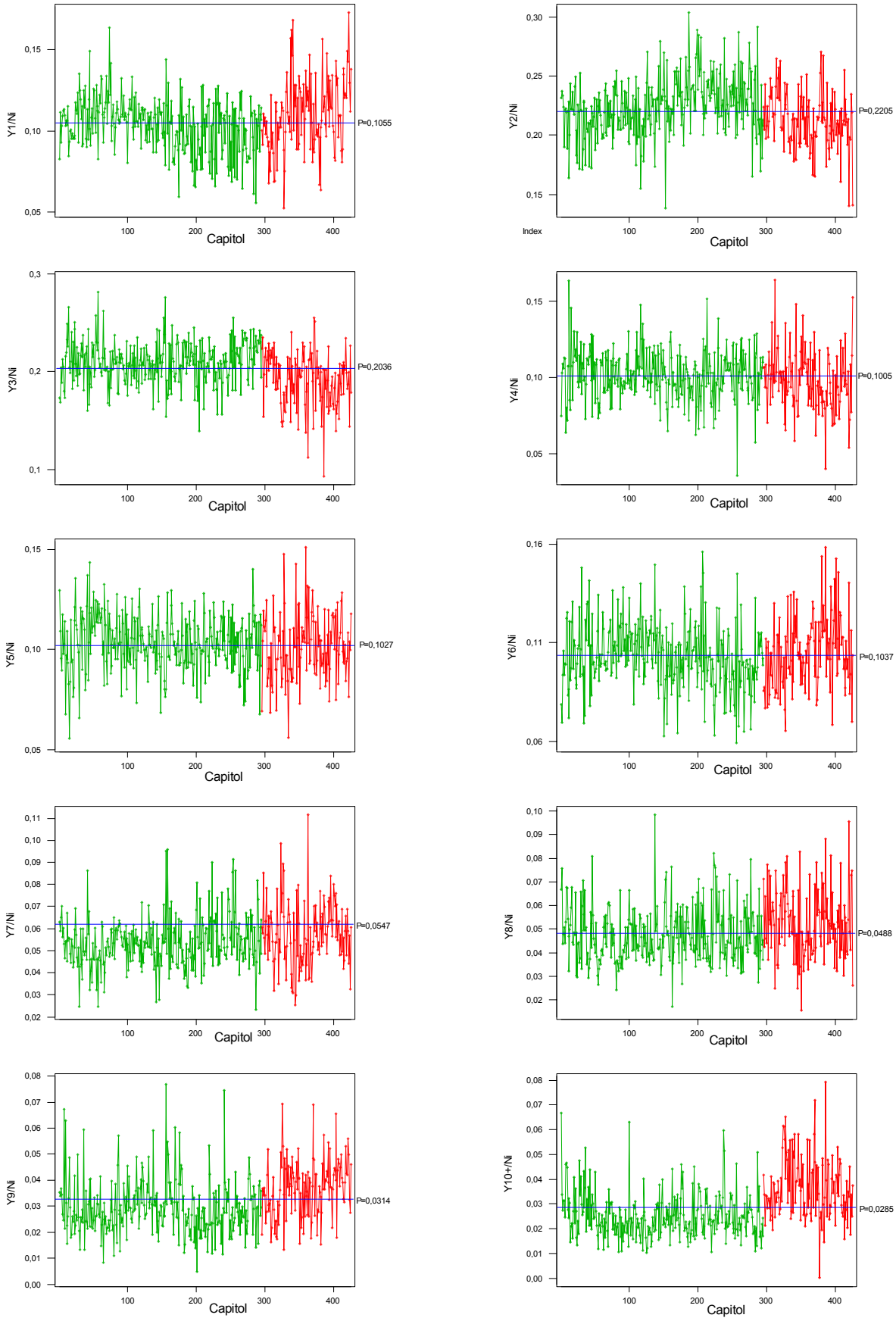

Figura 8.3: Evolució de la proporció de paraules de j lletres, per j=1,2,...10+,en els capítols de més de 200 paraules. El color verd correspon als capítols 1-345, el vermell als capítols 346-487.

La seqüència del vector multinomial amb el nombre de paraules de j lletres, per $j=1,2,...,10+$, $\mathbf{y_i}=(y_{1i},y_{2i},y_{3i},...,y_{10+i})$, es pot representar en un gràfic de control Chi-quadrat, com el de la figura 8.4. Els valors de la distància χ^2 per cada capítol s'han obtingut prenent, com a valor esperat de les proporcions, $\hat{\pi}_j$, les proporcions promig de paraules de j lletres en tots els capítols:

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

En el gràfic es pot observar com el punt de canvi es troba prop del capítol 371. S'observa, com ja s'havia vist en el cas de la llargada mitjana de paraula, com després del punt de canvi varia el nivell i augmenta la variabilitat. Això provoca que alguns capítols posteriors al 371 tinguin valors de l'estadístic Chi-quadrat més propers a la mitjana dels capítols anteriors al 371 que a la dels posteriors.

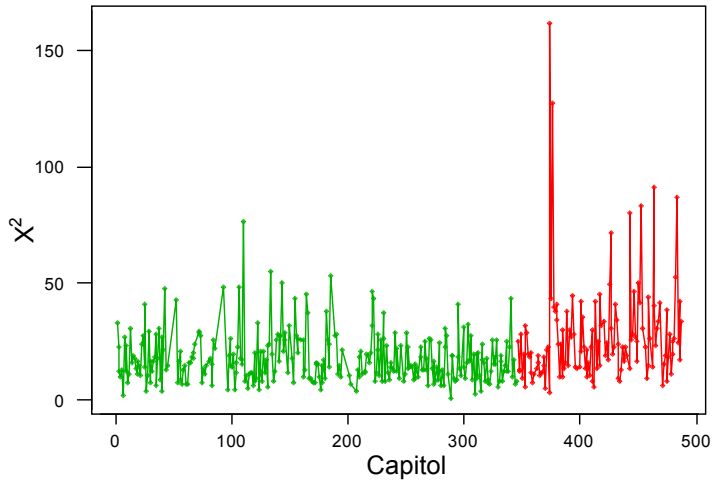

Figura 8.4: Gràfic Chi-quadrat per a la seqüència multinomial de llargades de paraula en els capítols de més de 200 paraules del *Tirant*. Els punts en verd corresponen als capítols 1-345, mentre que els de color vermell són els capítols del 346 al final.

8.2.3 Anàlisi de correspondències per la llargada de paraula

El punt de partida per a l'anàlisi de la llargada de paraula a través de l'Anàlisi de Correspondències està format per les distribucions de llargada de paraula tal i com apareixen a la taula de contingència 8.1. Els detalls sobre l'anàlisi de correspondències s'han descrit en el capítol 5 de la tesi.

El gràfic simètric per columnes, de l'esquerra de la Figura 8.5, mostra com a la primera component les categories estan ordenades per llargada: les paraules curtes es troben a la dreta de l'eix i les llargues en la part esquerra, amb l'única excepció de les d'una lletra. Aquesta primera component representa el 27% de la inèrcia total. La proporció de paraules d'una lletra no segueix la "ordenació" al llarg de la primera component. La segona component explica el 16% de la inèrcia total.

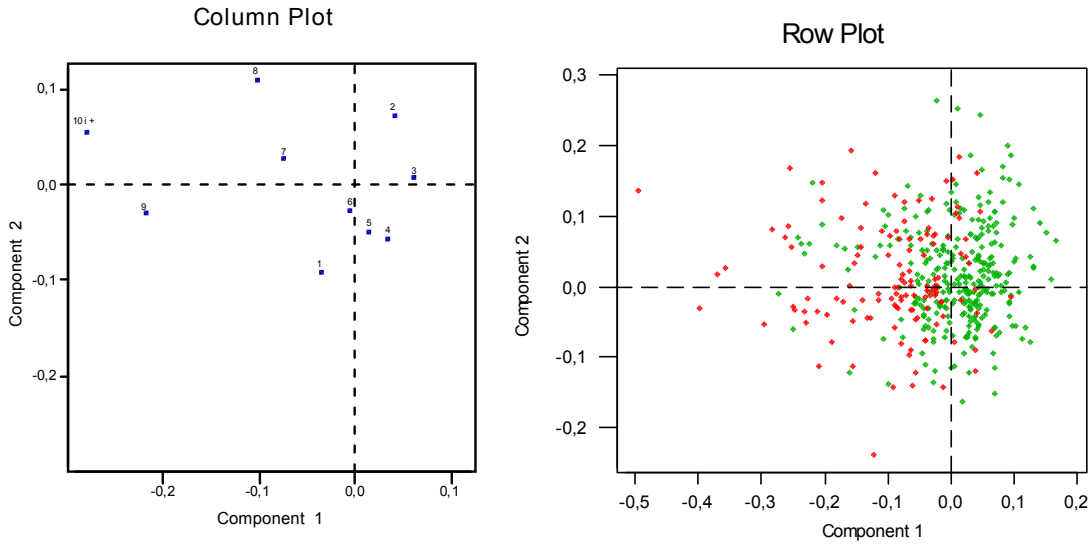

Figura 8.5: Gràfic simètric mostrant les columnes a l'esquerra i les files a la dreta, per a l'anàlisi de correspondències de la llargada de paraula per al capítols de més de 200 paraules. En el gràfic de la dreta, s'han representat en color verd els capítols 1-345 i en vermell els capítols 346-487.

El gràfic simètric que només mostra les files de la figura 8.5 (dreta) mostra com la majoria de punts que corresponen a capítols posteriors al 346, en vermell al gràfic, tenen valors de la primera component negatius, indicant que en aquests capítols la proporció de paraules llargues és major que la proporció promig, mentre que la proporció de paraules de 2, 3 i 4 lletres és menor que la mitjana, mentre que gairebé tots els punts amb primera component positiva corresponen a capítols anteriors al 346. Per aquests capítols, en general la proporció de paraules llargues és menor i la de paraules de 2, 3 i 4 lletres és més gran que la proporció mitjana. Els punts que en el gràfic simètric per files de la figura 8.5 (dreta) es troben més a l'esquerra corresponen als capítols en els que les proporcions de paraules de 8, 9 i 10 o més lletres són majors que les proporcions promig, $\hat{\pi}_j$, per j=*8, 9, 10+*, mentre que per aquests capítols paraules les proporcions de paraules de 2, 3 i 4 lletres són menors que els promitjos. Els punts més a la dreta en el gràfic, corresponen a capítols en els que les desigualtats van en sentit contrari.

Tant el gràfic simètric com els asimètrics per files i per columnes, com els biplots que mostren en el mateix gràfic files i columnes, són poc útils, ja que el fet de tenir 425 punts en un espai molt reduït en dificulta extraordinàriament la interpretació.

El valor esperat de la primera component de l'anàlisi de correspondències pot dependre del capítol però no depèn de la llargada del capítol N_i. En canvi, com en el cas de la llargada mitjana de paraula, la seva variança sí que en depèn, tal i com es pot veure de forma empírica en la figura 8.6.

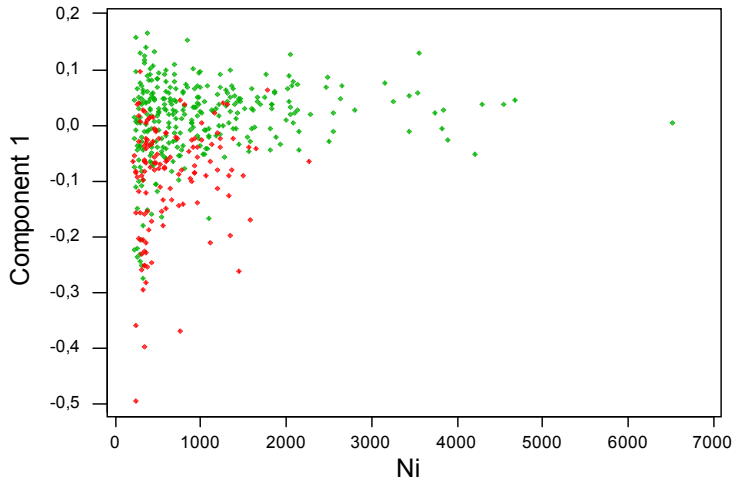

Figura 8.6: Relació entre la primera component de l'anàlisi de correspondències per la llargada de paraula amb la llargada de capítol N_i.

L'evolució temporal de la primera component per files, representada en el gràfic de la figura 8.7, mostra la gran diferència que existeix entre els capítols posteriors al 371 i els anteriors al 345, i com alguns els capítols que queden en el mig prenen valors de la primera component que els fan semblants als anteriors al 345 i els altres són semblants als posteriors al capítol 371. Cal observar que la mitjana global de la primera component principal per files és igual a *0*, com s'observa del gràfic de la figura 8.7, i com ja s'havia comentat en el capítol 5, atès que el 0 representa el perfil fila mig.

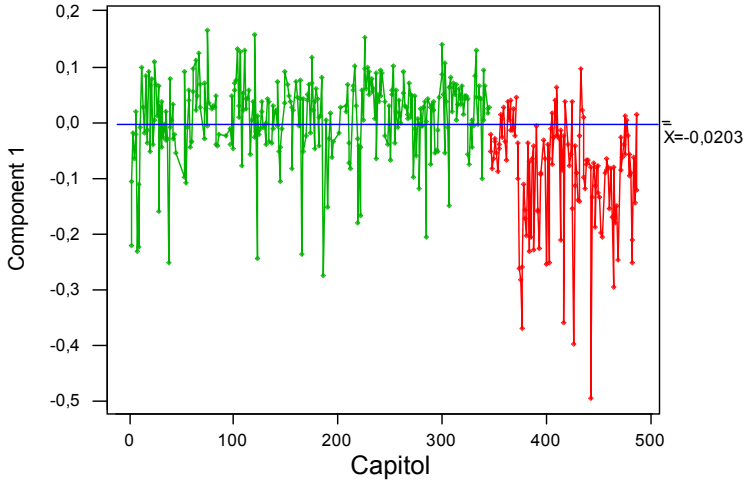

Figura 8.7: Evolució temporal de la primera component de l'anàlisi de correspondències, per capítols de més de 200 paraules. El color verd correspon als capítols 1-345 i el color vermell al 346 i posteriors.

8.2.4 Estimació del punt de canvi

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 de llargades mitjanes de paraula i del valor de la primera component de l'anàlisi de correspondències. A continuació, s'obtenen deu estimacions del punt de canvi a partir de les deu seqüències de proporcions de paraules de *j* lletres, per *j=1,2,...,10+*, per separat mitjançant l'aproximació al problema basada en models logístics. A continuació, s'estima el punt de canvi en la seqüència de

dades multinomials contingudes en la taula 8.1 via l'ajust de models politòmics i resumint la informació continguda en les deu seqüències binomials en una sola estimació.

8.2.4.1 Punt de canvi en seqüències Normals

S'ha estimat el punt de canvi per la llargada mitjana de paraula fent servir les tècniques proposades per a dades Normals descrites a la secció 6.2. Tot i que la distribució de la llargada és discreta i, per tant, no és normal, la hipòtesi de normalitat en la que es sustenten els models lineals s'acompleix de forma aproximada gràcies al teorema central del límit.

Sota les hipòtesi de independència i Normalitat, les llargades mitjanes de paraula , $\overline{ll}_i$, es distribueixen:

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

per *r=1,2,..., 424*, on $Ind_i^{(r)}$ 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 $w_i=N_i$. S'ha escollit l'estadístic *F* com a criteri per a la selecció del millor model, però podria haver-se triat el *t-ratio* de la comparació de dues mostres, l'estadístic *t* per fer el test de significativitat de la $\beta_1^{(r)}$, la versemblança del model, la variança residual s_R^2, o el coeficient de determinació R^2. Sota qualsevol d'aquests criteris, el punt de canvi estimat hauria sigut el mateix.

El gràfic de la figura 8.8 mostra l'evolució de F_r en funció de *r*. La millor estimació del punt de canvi per a la seqüència de llargades mitges per capítols de més de 200 paraules es troba en el màxim de la figura, a $\hat{r}_N$ =345.

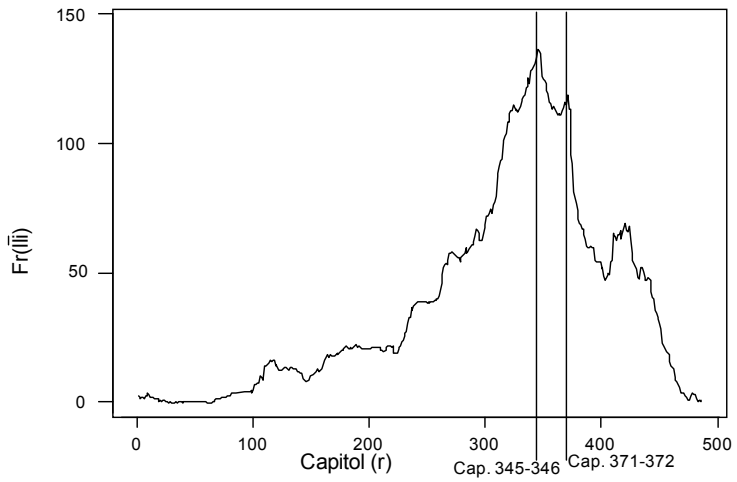

Figura 8.8: Gràfic de l'evolució de F_r en funció de *r* per a l'estimació del punt de canvi en la seqüència de llargades mitjanes de paraula pels capítols de més de 200 paraules. El màxim de F_r es té per *r=345*.

S'ha fet servir la mateixa metodologia per a estimar el punt de canvi en la seqüència de valors de la primera component de l'anàlisi de correspondències, obtinguts a la secció 8.2.3. El màxim de F_r es troba per *r=371*, tot i que també hi ha un màxim local, amb valor de F_r no gaire inferior a l'anterior, per a *r=345*, tal i com es pot observar en el gràfic de la figura 8.9.

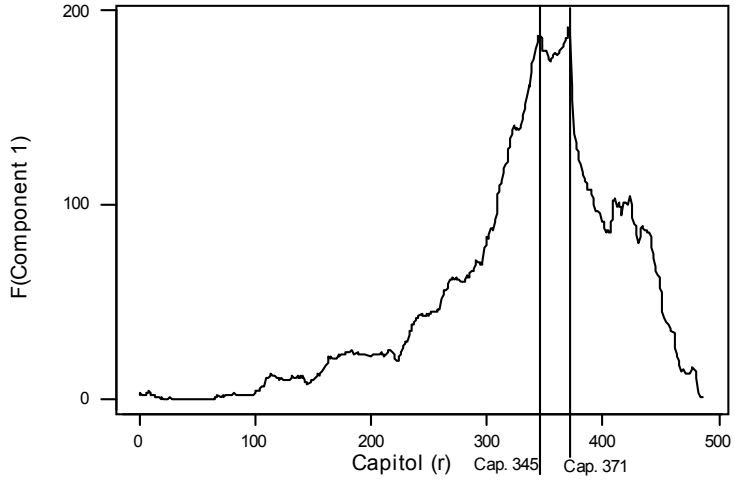

Figura 8.9: Evolució de F_r en funció de *r* en l'estimació del punt de canvi per a la primera component de l'anàlisi de correspondències fet a l'apartat 8.2.3. El màxim de F_r es té per *r=371*.

8.2.4.2 Punt de canvi en seqüències binomials

En aquest apartat s'estudien per separat les seqüències de nombres de paraules en cadascuna de les deu categories en que s'han classificat d'acord amb la seva llargada. D'aquesta manera s'obtenen 10 punts de canvi, possiblement diferents.

Suposant que existeix un punt de canvi a *r*, per una categoria *j* fixada, amb *j=1,2,...10+*, y_{ij} té una distribució *binomial* (N_i,π_{ja}) per *i=1,...,r* i *binomial* (N_i,π_{jd}) per *i=r+1,...,n*. Per cadascuna de les 10 categories s'ajusta el model logístic (6.2):

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

La figura 8.25 conté els gràfics de $L_j\left(r, \hat{\beta}^{(r)}\right)$, el màxim del logaritme de la versemblança, en funció de *r* per a les 10 categories en què s'ha subdividit la llargada de paraula.

Les seqüències de paraules curtes, d'una, dues i tres lletres, presenten el màxim en *r=327*, les seqüències de paraules de vuit i deu o més lletres tenen el màxim en el capítol 371, mentre que les de nou lletres el tenen en el capítol 345. Pel que fa a les llargades intermitges, de quatre a set lletres, els màxims no són coincidents, i van del capítol 145 al 427.

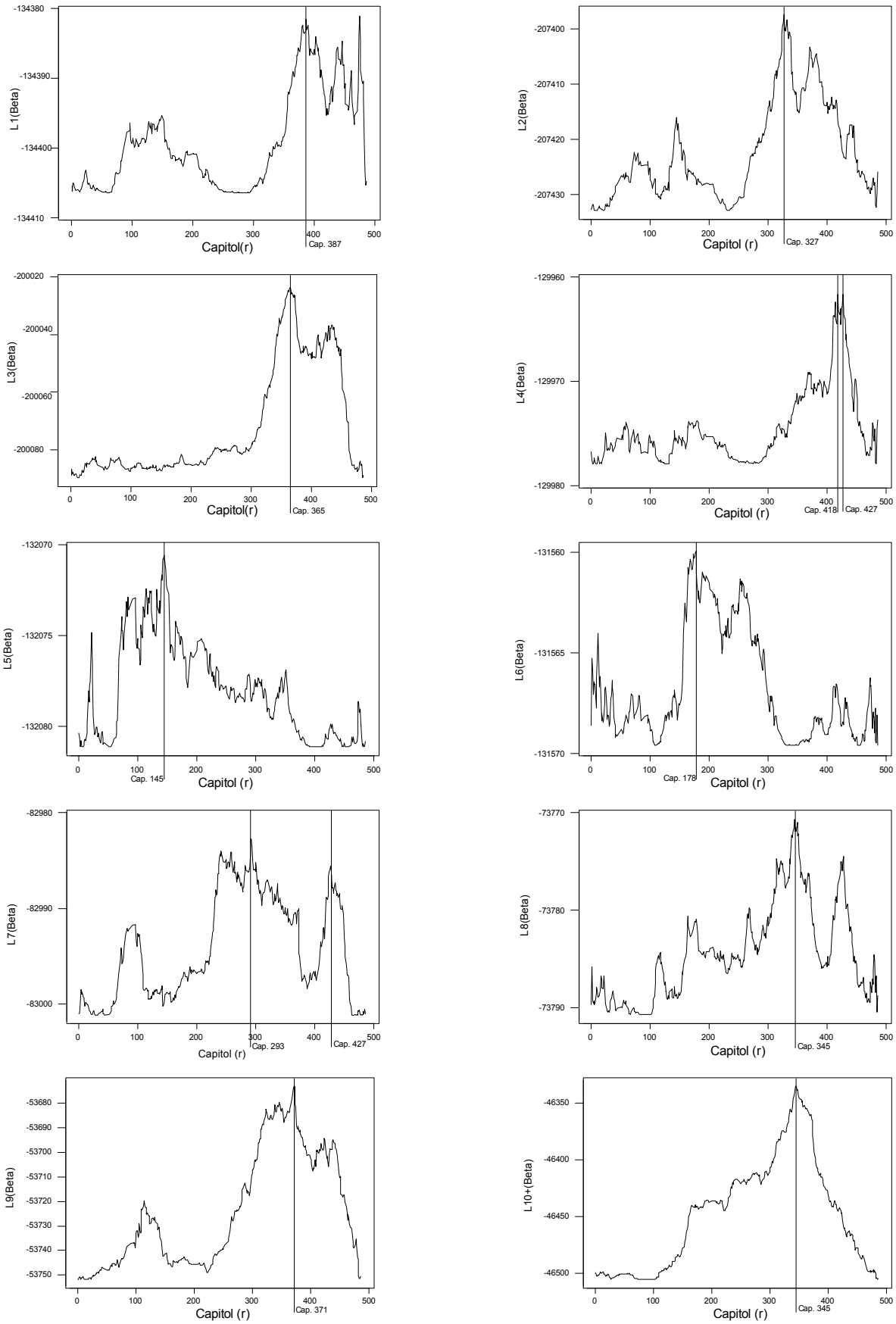

Figura 8.10: Evolució del màxim del logaritme de la versemblança del model logístic en funció de r per a les seqüències de paraules de *j* lletres, per *j=1,2,...,10+*. La línia vertical indica el punt de canvi estimat per cadascuna de les llargades.

8.2.4.3 Punt de canvi en seqüència multinomial; model politòmic

El problema d'estimar el punt de canvi en la seqüència multinomial formada per les 10 categories en que s'ha subdividit la llargada de paraula, s'ha abordat primer mitjançant l'ajust de models de regressió politòmica i, posteriorment, a partir de les 10 estimacions individuals obtingudes en l'apartat 8.2.4.2, emprant la metodologia explicada en el capítol 6 de la tesi. Si y_{ji} és el nombre de paraules de j lletres en el capítol i, $\mathbf{y_i}=(y_{1i},y_{2i},....,y_{10+i})$ està distribuït $Mult(N_i,\pi_i)$, on N_i és la llargada del capítol i $\pi_i=(\pi_{1i},\pi_{2i},...,\pi_{10+i})$ és el vector amb les probabilitats que una paraula sigui assignada a cadascuna de les l categories.

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 $j=1,...,10+$, 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:

$$L\left(r,\hat{\beta}^{(r)}\right)=\sum_{i=1}^{n}\left(\sum_{j=2}^{10+}y_{ji}\left(\hat{\beta}_{0j}^{(r)}+\hat{\beta}_{1j}^{(r)}Ind_{1i}^{(r)}\right)-\ln\left(1+\sum_{j=2}^{10+}e^{\left(\hat{\beta}_{0j}^{(r)}+\hat{\beta}_{1j}^{(r)}Ind_{1i}^{(r)}\right)}\right)\right),$$

és màxim. El gràfic de la figura 8.11 mostra l'evolució de $L\left(r,\hat{\beta}^{(r)}\right)$ en funció de r. L'estimació del punt de canvi és $\hat{r}_M=371$, tot i que hi ha un màxim local de valor semblant al màxim global per $r=345$.

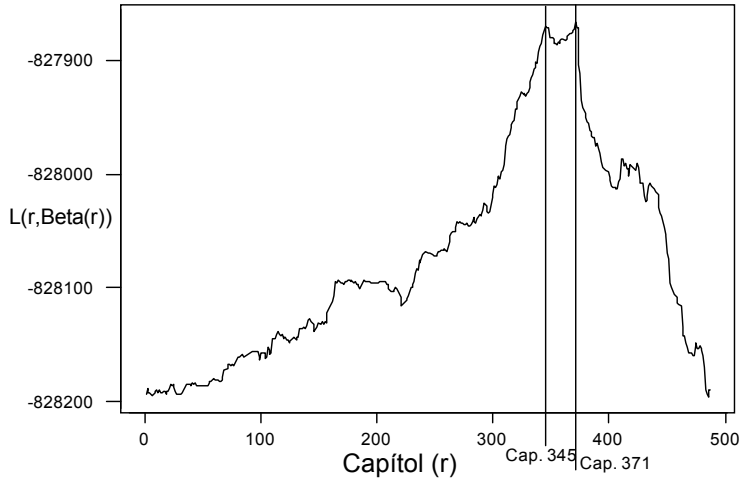

Figura 8.11: Evolució del màxim del logaritme de la versemblança en funció de r. El màxim del màxim es té per $\hat{r}_M=371$, i s'observa un màxim local per $r=345$.

Estimar r en una seqüència multinomial via l'ajust de models de regressió politòmica és equivalent a ajustar-lo via màxim versemblant. De totes les vies utilitzades en aquesta secció 8.2.4, aquesta és el manera més correcta de fer-ho i la més elegant.

8.2.4.4 Punt de canvi en seqüència multinomial; combinació de binomials

Una manera alternativa d'estimar el punt de canvi en una seqüència de multinomials és calculant les deu estimacions individuals en cadascuna de les deu seqüències binomials de paraules de *j* lletres, obtingudes a 8.2.4.2. Això es pot fer adaptant els estadístics proposats per Wolfe i Chen (1990), que resumeixen les *l=10* estimacions en les seqüències binomials en un sol valor. Tres estimadors alternatius de *r*, proposats en el capítol 6 de la tesi i que resumeixen les *l=10* estimacions en un sol valor, són $\hat{r}_{ML1}$, $\hat{r}_{ML2}$ i $\hat{r}_{ML3}$. Per calcular $\hat{r}_{ML1}$ ens fixem en les estimacions del punt de canvi per a les *10* seqüències i escollim aquella *r* per la que obtenim el màxim global de la versemblança, $\hat{r}_{ML2}$ s'obté sumant per a cada *r=1,...,424* les versemblances pels *10* models ajustats i prenent el màxim, i $\hat{r}_{ML3}$ s'obté de forma semblant, sumant ara els quadrats de les versemblances.

La figura 8.12 mostra l'evolució de:

$$\max_{1\le j\le 10} L_j\left(r,\hat{\beta}^{(r)}\right),$$

$$\sum_j L_j\left(r,\hat{\beta}^{(r)}\right),$$

i

$$\sum_j L_j^2\left(r,\hat{\beta}^{(r)}\right),$$

en funció de *r*. Els màxims dels tres gràfics senyalen $\hat{r}_{ML1}$, $\hat{r}_{ML2}$ i $\hat{r}_{ML3}$, respectivament. Tal i com et pot observar en la figura 8.12, $\hat{r}_{ML1}$=345, $\hat{r}_{ML2}$=371 i $\hat{r}_{ML3}$=371.

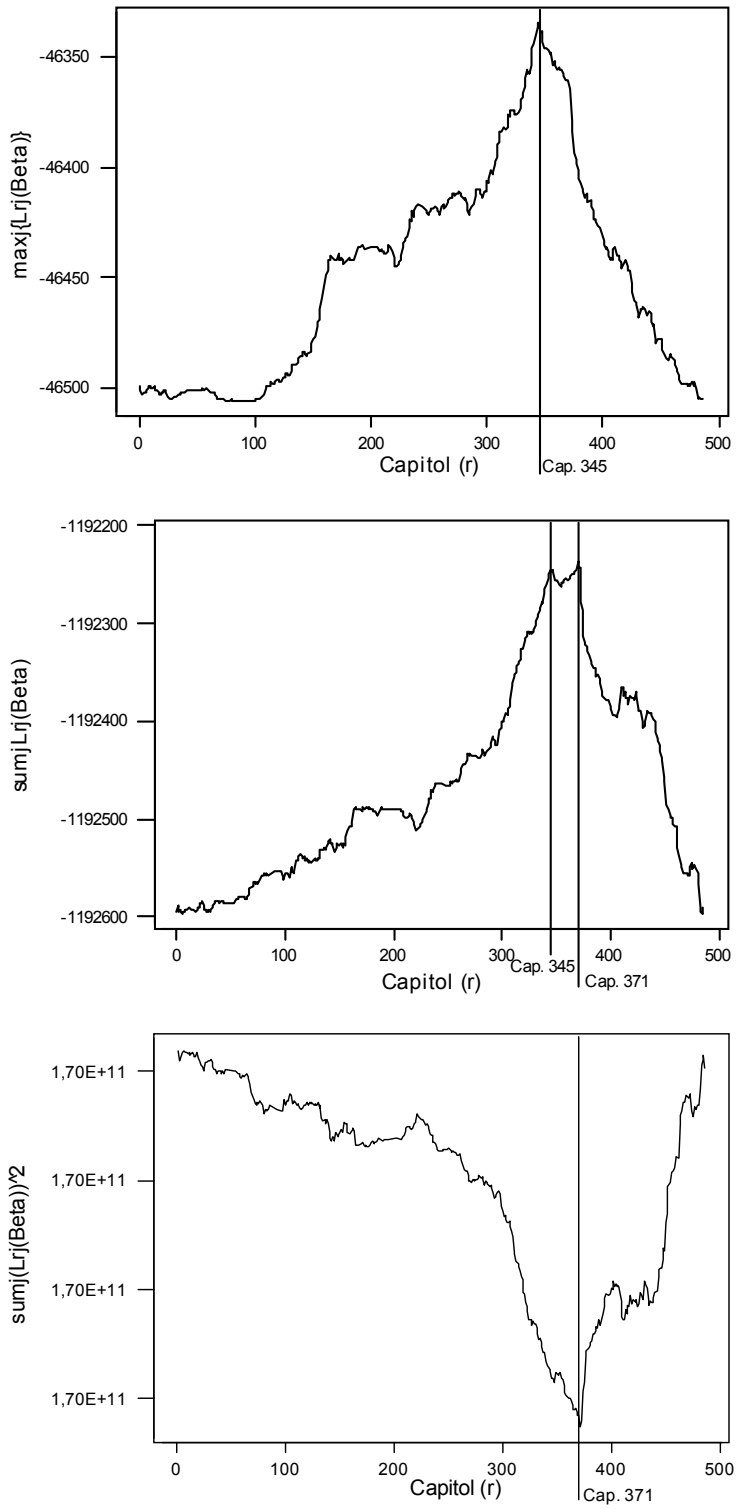

Figura 8.12: Gràfics que representen l'evolució de $\max_{1\le j\le 10} L_j\left(r,\hat{\beta}^{(r)}\right)$, $\sum_j L_j\left(r,\hat{\beta}^{(r)}\right)$ i de $\sum_j L_j^2\left(r,\hat{\beta}^{(r)}\right)$, en funció de r. Els tres extrems corresponen a $\hat{r}_{ML1}$, $\hat{r}_{ML2}$ i $\hat{r}_{ML3}$.

8.2.4.5 Conclusions

S'observa l'existència d'un punt de canvi molt clar en l'estil a través de la llargada de paraula. L'estimador del punt de canvi no és sempre el mateix, depenent del mètode o la unitat d'estadística textual analitzada, ballant entre el capítol 345 i el 371. Això podria indicar que en la zona entre aquests dos capítols es barregen els dos estils.

En general, en els primers capítols la llargada mitjana de paraula és mes breu que al final, atès que hi abunden més paraules de dues i tres lletres mentre que són menys abundants les paraules d'una i de set o més lletres.

La taula 8.2 mostra un quadre resum amb les estimacions del punt de canvi obtingudes en l'estudi de la llargada de paraula pels capítols de més de 200 paraules.

Distribució	**Seqüència**	**Punt de canvi ppal.**	**Punt de canvi sec.**
Normal	*Llargada mitjana*	$\hat{r}_N$ =345	r=371
	1ª Comp. A. Corresp.	$\hat{r}_N$ =371	r=345
Binomial	*1 lletra (y_{1i})*	$\hat{r}_{L_1}$ =387	
	2 lletres (y_{2i})	$\hat{r}_{L_2}$ =327	
	3 lletres (y_{3i})	$\hat{r}_{L_3}$ =365	
	4 lletres (y_{4i})	$\hat{r}_{L_4}$ =427	
	5 lletres (y_{5i})	$\hat{r}_{L_5}$ =145	
	6 lletres (y_{6i})	$\hat{r}_{L_6}$ =178	
	7 lletres (y_{7i})	$\hat{r}_{L_7}$ =293	r=427
	8 lletres (y_{8i})	$\hat{r}_{L_8}$ =345	
	9 lletres (y_{9i})	$\hat{r}_{L_9}$ =371	
	10 o més lletres (y_{10+i})	$\hat{r}_{L_{10+}}$ =345	
Multinomial	*Combinació de binom.* $\max_{1\le j\le 10} L_j\left(r,\hat{\beta}^{(r)}\right)$	$\hat{r}_{ML1}$ =345	
	Combinació de binom. $\sum_j L_j\left(r,\hat{\beta}^{(r)}\right)$	$\hat{r}_{ML2}$ =371	r=345
	Combinació de binom. $\sum_j L_j^2\left(r,\hat{\beta}^{(r)}\right)$	$\hat{r}_{ML3}$ =371	
	Model Politòmic	$\hat{r}_M$ =371	r=345

Taula 8.2: Quadre resum de les estimacions del punt de canvi obtingudes en l'anàlisi de la llargada de paraula per als capítols de més de 200 paraules.

8.2.5 Anàlisi cluster de les files d'una taula de contingència

De l'anàlisi gràfica de la llargada de paraula en els capítols s'ha pogut observar com alguns capítols posteriors al punt de canvi tenen unes característiques que els fan més semblants als anteriors que als posteriors. Aquesta situació es dóna per totes les variables que hem considerat relacionades amb la llargada de paraula. L'estudi del punt de canvi ha revelat que hi poden haver dos grups diferenciats, però hi ha la possibilitat que hi hagi barreja d'estils després del punt de canvi.

S'ha decidit fer una anàlisi cluster, per veure si és possible recol·locar algun capítol del final. Es faran servir les tècniques d'anàlisi cluster exposades en el capítol 7, basades en dos criteris d'ajust de models per dades politòmiques. El punt de partida de l'anàlisi és la taula de contingència 8.1, on les files són els 425 capítols de més de *200* paraules i les 10 columnes representen els nombres de paraules d'1, 2, 3, ..., 9 i 10 o més lletres.

8.2.5.1 Anàlisi cluster basada en la distància χ^2

S'ha executat 1000 vegades l'algorisme cluster basat en la distància χ^2, amb assignacions inicials aleatòries de capítols a un dels dos grups. Els resultats han portat a set solucions diferents, no convergint a una única solució. Si el criteri triat és el de quedar-nos amb la solució per la que el valor de χ^2_B, *entre clusters*, és màxim, més del 77% de les execucions condueixen a la mateixa solució. De les altres sis solucions obtingudes, tres, que representen l'11% de les execucions, són clarament pitjors en donar valors de χ^2_B inferiors i, alhora, valors χ^2_W, *dintre dels clusters*, superiors. Una solució, que representa el 2,5% de les execucions, dóna un valors de χ^2_B bastant inferior mentre que el valor χ^2_W és pràcticament igual i, per tant, no millora l'assignació. En les altres dues, que representen al voltant del 10% de les execucions de l'algorisme, s'obtenen valors de χ^2_B bastant inferiors, però els valors χ^2_W són també inferiors, pel que podrien ser solucions escollides si el criteri fos el de minimitzar χ^2_W. Triem la solució per la que χ^2_B és màxim i que, al mateix temps, és a la que ha convergit l'algorisme més vegades.

A l'Annex A8.1 hi ha les assignacions dels capítols a cadascun dels 2 clusters. En total, 55 capítols anteriors al *345* i 24 posteriors al *371* han canviat d'assignació. S'observa com els capítols entre el *345* i el *371* han sigut assignats a ambdós clusters: entre el *346* i el *355* estan assignats majoritàriament al cluster dels capítols del final, essent les excepcions els capítols *347* i el *350*, mentre que entre el *356* i el *359* i del *363* al *371* han sigut assignats al cluster dels capítols del principi. El gràfic de la figura 8.13 mostra l'assignació de capítols a clusters.

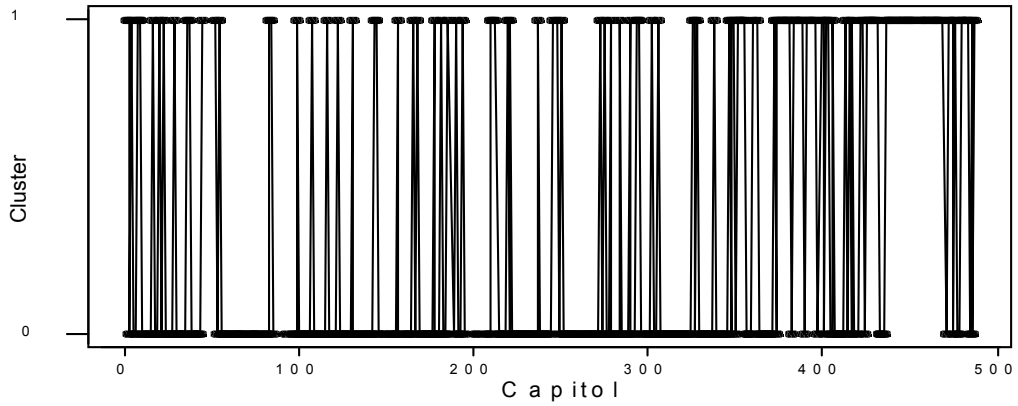

Figura 8.13: Assignació dels capítols a un dels dos clusters, després de l'anàlisi cluster basada en la distància χ^2 per la llargada de paraula en els capítols de més de 200 paraules del *Tirant*.

L'estadístic χ^2 per la taula de contingència val χ^2=8408,3, mentre que el valor entre clusters és de χ^2_B =1304,9, el que dóna una relació χ^2_B / χ^2 = 0,155, és a dir la divisió en aquests dos clusters explica el 15,5% de la "no-homogeneïtat" entre tots els capítols en funció de la distribució de la llargada de paraula. Aquest valor, tal i com s'ha avançat en el capítol 7, és menor que la proporció de inèrcia explicada per la primera component principal, que és del 27%.

Els clusters obtinguts de l'aplicació de l'algorisme separen els capítols del Tirant en funció de la primera component de l'Anàlisi de Correspondències, com es pot apreciar del gràfic de la figura 8.14, en la que s'han representat en colors diferents els capítols assignats a clusters diferents.

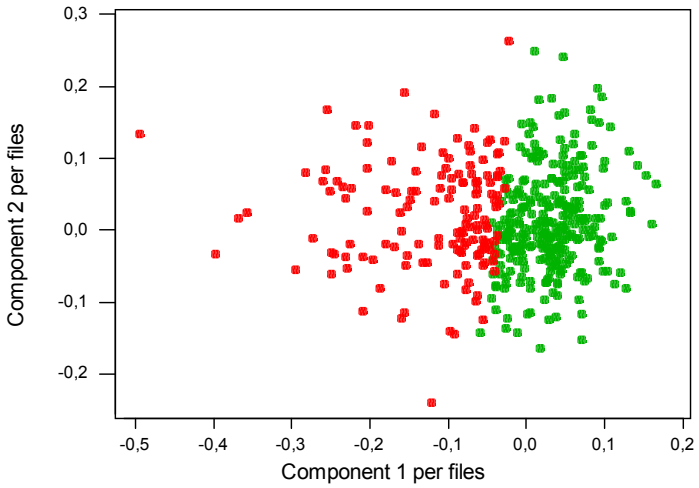

Figura 8.14: Projecció dels perfils fila en el pla format per les dues primeres components de l'anàlisi de correspondències en l'estudi de la llargada de paraula realitzat a la 8.2.3. Els punts en vermell corresponen als capítols assignats al cluster *0*, el que conté majoritàriament els capítols del principi del llibre, i els punts en vermell als capítols assignats al cluster *1*, el que conté els capítols del final.

S'han validat els resultats obtinguts en l'anàlisi cluster, ajustant, el model politòmic:

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

on $Ind_{1i}^{(c)}$ és la variable indicadora que pren valor *0* pels blocs assignats al cluster *0* i pren valor *1* pels blocs assignats al cluster *1*, i es calcula el màxim del logaritme de la versemblança del model.

De forma anàloga s'ajusta per cadascuna de les l=10 categories el model logístic:

$$y_{ji} \sim Binomial\left(N_i, \pi_{ji} = \frac{e^{\beta_{j0}^{(c)} + \beta_{j1}^{(c)} Ind_{1i}^{(c)}}}{1 + e^{\beta_{j0}^{(c)} + \beta_{j1}^{(c)} Ind_{1i}^{(c)}}}\right),$$

on $Ind_i^{(c)}$ és la mateixa variable indicadora, i per cadascun d'aquests models es guarda el màxim del logaritme de la versemblança, $L_j\left(\hat{\beta}^{(c)}\right)$, amb *j=1,...,10+*, i es calculen els estadístics:

$$\max_{1 \le j \le 10} L_j\left(\hat{\beta}^{(c)}\right),$$

$$\sum_{j=1}^{10} L_j\left(\hat{\beta}^{(c)}\right),$$

i

$$\sum_{j=1}^{10} L_j^2\left(\hat{\beta}^{(c)}\right).$$

La comparació de les bondats d'ajust del model que suposa un únic punt de canvi i el que suposa l'existència de clusters, calculats fent servir la distància χ^2 es mostren el la taula 8.3. En ella es pot veure com l'anàlisi cluster millora, tot i que de forma poc notable, els estadístics, pel que es pot assegurar que l'estimació del punt de canvi proposa un bona separació en grups.

Model Ajustat	**Estadístic**	**Punt de Canvi**	**Cluster χ^2**
Politòmic	$L\left(r, \hat{\beta}^{(r)}\right)$	-827865	-827580
10 logístics	$\max_{1\le j\le l} L_j\left(r, \hat{\beta}^{(r)}\right)$	-46335	-46297
10 logístics	$\sum_{j=1}^{l} L_j\left(r, \hat{\beta}^{(r)}\right)$	-1192238	-1191930
10 logístics	$\sum_{j=1}^{l} L_j^2\left(r, \hat{\beta}^{(r)}\right)$	170094932228	170038869913

Taula 8.3: Comparació dels estadístics emprats per calcular la bondat d'ajust del model que suposa un punt de canvi i dels models que suposen l'existència de clusters calculada a partir de la distància χ^2. Els primer estadístic és suposa que el punt de canvi és $\hat{r}_M$, i els altres tres suposen que és $\hat{r}_{ML1}$, $\hat{r}_{ML2}$ i $\hat{r}_{ML3}$.

Analitzant com queden la llargada mitjana de paraula i la primera component de l'anàlisi de correspondències després de l'anàlisi cluster es veu, de la mateixa manera que a la figura 8.14 com el cluster agrupa els capítols segons el valor de la primera component de l'anàlisi de correspondències. A la figura 8.15 s'hi ha representat l'evolució temporal d'aquestes dues variables, representant en colors diferents els capítols assignats a clusters diferents. La majoria de capítols anteriors al punt de canvi pertanyen al cluster en verd en la figura 8.15, mentre que la majoria dels blocs posteriors, amb excepcions, pertanyen al cluster en vermell en aquesta figura.

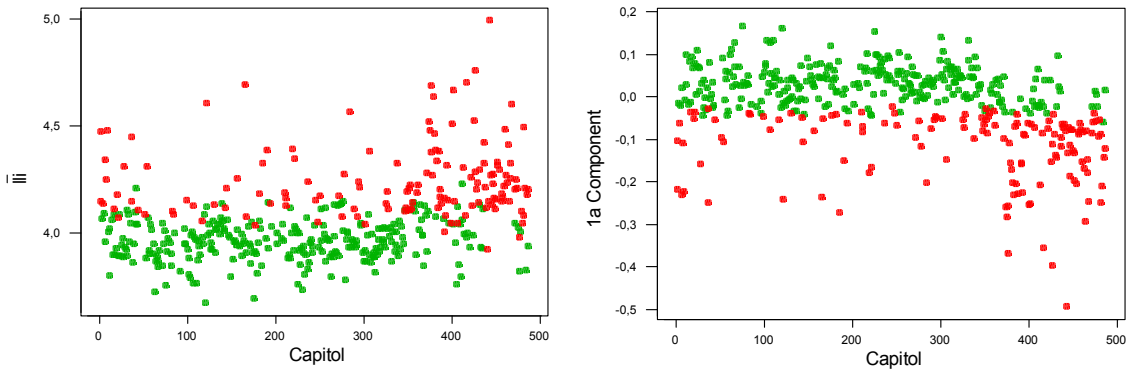

Figura 8.15: evolució temporal de la llargada mitjana de paraula, a l'esquerra, i del valor de la primera component principal de l'anàlisi de correspondències per les files, a la dreta, després de l'anàlisi cluster. Els punts en verd i en vermell corresponen als capítols agrupats en un dels dos clusters.

8.2.5.2 Anàlisi cluster basada en la deviança

L'execució de l'algorisme cluster basat en la deviança dels models de regressió politòmica porta a resultats molt semblants als obtinguts amb l'algorisme basat en la distància χ^2. Hi ha sis capítols que l'algorisme classifica en el cluster 1, el que conté majoritàriament els capítols del final del llibre que l'algorisme basat en la distància χ^2 classifica en l'altre cluster. Aquests capítols són: 42, 254, 270, 369, 412, 432. Es pot apreciar com només tres d'ells són posteriors a l'estimació del punt de canvi donada per $\hat{r}_M^*$. Hi ha un capítol, el 399, que l'algorisme col·loca en el cluster que conté majoritàriament els capítols del principi del llibre, que l'algorisme basat en la distància χ^2 col·loca en l'altre. A l' Annex A8.1 hi ha les assignacions que fan cadascun dels dos mètodes dels capítols de més de 200 paraules als 2 clusters.

El guany que s'obté amb aquest mètode respecte a l'aplicació de l'algorisme cluster basat en la distància χ^2, en termes del màxim del logaritme de la versemblança, és extraordinàriament petit.

8.3 Estudi de la llargada de paraula per blocs

8.3.1 Notació i forma de les dades

La taula 8.4 és l'anàloga per blocs a la taula 8.1 per capítols. Mostra el tros corresponent als blocs 303-349 de la taula que es farà servir per a l'anàlisi de la llargada de paraula pels blocs. La taula sencera té 396 files, que corresponen als blocs, i 13 columnes: la primera columna indica el número de bloc, la segona el nombre total de paraules, N_i que, a diferència del que passava pels capítols, és constant i igual a 1000 paraules per tots els blocs, i la darrera columna conté la llargada mitjana per bloc, $\overline{ll}_i$. Les deu columnes centrals formen una taula de contingència que conté el nombre de paraules de *j* lletres, per *j=1,2,...,10+*. La columna que agrupa el nombre de paraules de 10 lletres o superior s'ha etiquetat com "*10+*". A la taula podem llegir, per exemple, com per al bloc 303 hi ha 102 paraules d'una lletra, 222 de dues lletres, 210 paraules de tres lletres, 26 de nou lletres i 16 de deu o més lletres.

Anomenem y_{ji} al nombre de paraules de *j* lletres, per *j=1,2,...,10+*, en el bloc i-èssim, $\overline{ll}_i$ a la mitjana de les llargades en el bloc i-èssim, i *ll* a la variable aleatòria llargada d'una paraula. Sota la hipòtesi de independència, y_{ji} es distribueix *binomial*(N_i=*1000*, π_{ji}), on π_{ji} és la probabilitat que una paraula agafada a l'atzar en el bloc i-èssim pertanyi a la categoria j-èssima. El vector de comptatges $\boldsymbol{y_i}=(y_{1i},y_{2i,...},y_{10+i})$ es distribueix *Mult*(N_i, $\boldsymbol{\pi_i}$), on $\boldsymbol{\pi_i}=(\pi_{1i},\pi_{2i},...,\pi_{10+i})$ és el vector tal π_{ji} és la probabilitat que una paraula presa a l'atzar en el bloc i-èssim tingui *j* lletres.

Bloc	**N_i**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10+**	$\overline{ll}_i$
303	1000	102	222	210	129	109	96	50	40	26	16	3,894
304	1000	96	215	232	112	111	98	42	41	31	22	3,944
305	1000	110	210	218	104	111	107	51	30	27	32	3,982
306	1000	97	208	230	103	95	116	73	42	18	18	3,985
307	1000	113	165	239	126	113	103	51	42	34	14	3,996
308	1000	115	177	219	103	119	107	58	41	41	20	4,088
309	1000	97	207	227	101	108	105	51	48	35	21	4,046
310	1000	109	177	237	112	100	113	61	45	26	20	4,027
312	1000	104	213	244	109	86	95	50	43	29	27	3,932
313	1000	98	201	211	112	102	89	69	48	37	33	4,174
314	1000	110	222	170	114	115	95	63	54	30	27	4,098
315	1000	101	173	206	105	116	114	56	40	45	44	4,327
316	1000	103	212	217	99	102	118	45	42	31	31	4,067
317	1000	108	205	214	76	120	100	61	59	28	29	4,139
318	1000	95	215	217	121	88	78	51	65	39	31	4,153
319	1000	102	230	209	90	87	87	56	57	54	28	4,148
320	1000	84	234	197	102	89	114	68	55	23	34	4,191
321	1000	79	236	222	94	85	99	60	59	36	30	4,160
322	1000	84	232	209	114	101	93	55	45	32	35	4,119
323	1000	109	214	210	105	87	112	55	53	22	33	4,065
324	1000	74	236	225	114	106	89	37	59	28	32	4,071
325	1000	99	254	182	109	88	115	54	43	24	32	4,016
326	1000	91	252	193	102	95	102	64	49	20	32	4,025
327	1000	99	252	198	108	96	88	44	60	26	29	3,989
328	1000	103	211	192	93	107	104	56	58	35	41	4,267
329	1000	100	184	179	86	107	93	87	65	51	48	4,538
330	1000	112	229	155	95	90	93	86	47	40	53	4,353
331	1000	116	189	155	92	91	95	75	58	68	61	4,626
332	1000	78	233	167	114	101	85	76	65	31	50	4,403
333	1000	105	234	162	95	108	107	43	64	42	40	4,262
334	1000	116	188	205	111	78	97	65	52	43	45	4,256
335	1000	104	202	177	91	100	123	75	60	42	26	4,306
336	1000	101	189	237	113	116	93	53	43	32	23	4,034
337	1000	137	202	183	108	83	106	48	49	42	42	4,115
338	1000	145	205	193	76	85	114	35	55	44	48	4,130
339	1000	112	178	199	115	102	124	62	40	40	28	4,194
340	1000	116	199	181	96	119	121	43	47	33	45	4,225
341	1000	106	208	198	92	104	103	54	49	43	43	4,246
342	1000	101	219	176	99	102	99	76	45	46	37	4,267
343	1000	101	204	218	117	106	89	47	50	38	30	4,095
344	1000	129	236	193	105	91	94	45	39	29	39	3,922
345	1000	118	227	200	128	105	98	44	30	24	26	3,851
346	1000	113	229	211	96	106	92	48	37	31	37	3,987
347	1000	118	188	222	115	109	101	42	35	44	26	4,021
348	1000	127	217	213	118	115	87	47	43	18	15	3,762
349	1000	107	209	218	117	104	89	59	45	24	28	3,992

Taula 8.4: Part de la taula de contingència per la llargada de paraula pels blocs de 1000 paraules: cada fila correspon a un bloc, la primera columna dóna el número de bloc, la segona el nombre de paraules per bloc, les següents el nombre de paraules de *j* lletres, per *j=1,2,...,10+*, i la darrera conté la llargada mitjana de paraula per bloc.

8.3.2 Anàlisi descriptiva de les dades

La llargada mitjana de paraula per bloc, $\overline{ll}_i$, és una mitjana de 1000 observacions. Si la distribució de les llargades de paraula fos la mateixa al llarg de tot el llibre, pel teorema central del límit $\overline{ll}_i$ hauria de tenir distribució Normal $N(\mu,\sigma^2)$, es a dir, si l'estil no canviés en tot el llibre, el valor esperat d'aquest valor seria constant, i com que la llargada dels blocs és sempre $N=1000$, la variança d'aquesta variable també hauria de ser constant. L'histograma de les $\overline{ll}_i$ de la figura 8.16 mostra una distribució no Normal, el que ens porta a pensar que ens trobem davant una mescla de distribucions, que possiblement implica la presència de més d'un estil.

La mitjana de les $\overline{ll}_i$ és de 4,018 lletres per paraula, la desviació tipus és de 0,141 i els valors màxim i mínim són, respectivament, 4,626 i 3,654 lletres per paraula.

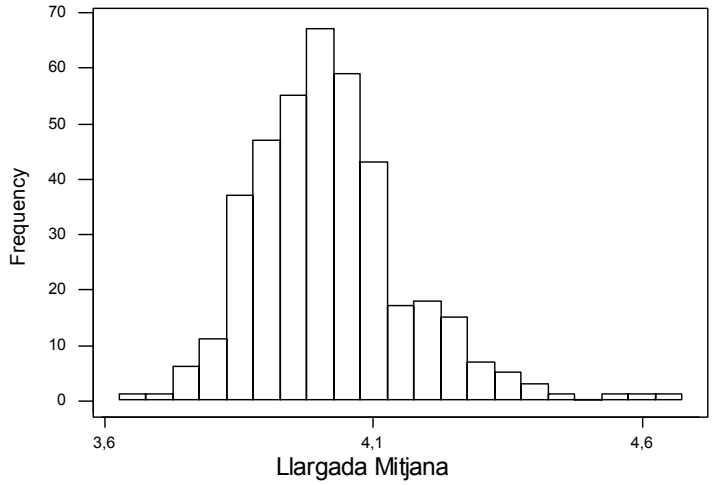

Figura 8.16: histograma de la llargada mitjana de paraula pels blocs de 1000 paraules.

8.3.3 Gràfics de Control per la llargada de paraula

Obtenim una sola mesura contínua per bloc per la llargada mitjana de paraula, $\overline{ll}_i$. La representació gràfica de l'evolució temporal de la llargada mitjana de paraula al llarg del *Tirant* reflexarà els possibles canvis en el nivell i en la variabilitat i, per tant, ha de donar una idea de si l'estil és sempre el mateix o be hi ha algun canvi.

La figura 8.17 mostra l'evolució de la llargada mitjana de paraula al llarg del *Tirant* mitjançant un gràfic de observacions individuals. Els límits de control s'han obtingut a partir de totes les dades, tal i com s'ha explicat en el capítol 5 de la tesi. S'observa un canvi en el nivell del procés, entre els blocs 312 i 313, que correspon, aproximadament, al capítol 346. Abans del canvi els valors de la llargada mitjana es troben al voltant o lleugerament per sota del valor mig, mentre que després del canvi estan clarament per sobre de la mitjana, inclús en alguns casos per sobre del Límit Superior de Control. Tot i així, s'aprecia que hi ha algun bloc després del canvi que té un valor de la llargada mitjana més semblant als anteriors al canvi que no pas als posteriors. Aquest canvi ha estat marcat en el gràfic de la figura 8.17 amb un canvi de color, els blocs que van des del primer al 312 estan en color verd, mentre que els que van del 313 al final esta en color vermell.

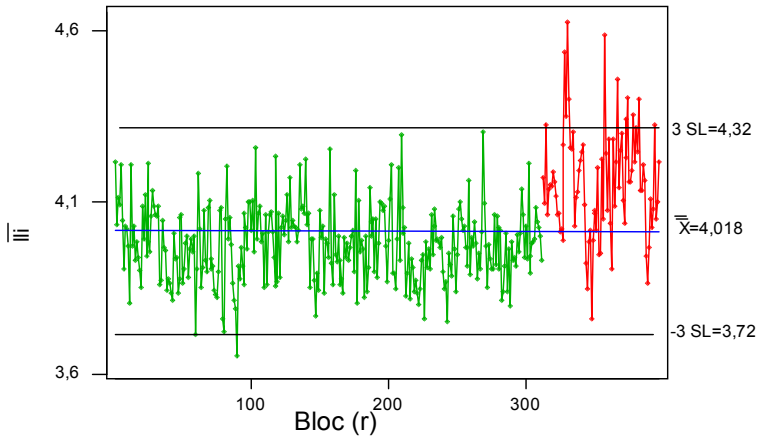

Figura 8.17: Gràfic de control per a Observacions Individuals per a la llargada mitjana de paraula per blocs. El canvi de color es troba entre els blocs 312 i 313, que correspon, aproximadament, al capítol 346.

La Figura 8.18 mostra el gràfic Cusum per la llargada mitjana de paraula. En el capítol 5 s'ha descrit com aquests gràfics representen l'evolució temporal de:

$$S_m = \sum_{i=i}^{m} \left(\overline{ll}_i - \mu_0 \right),$$

per *m=1,2,..., 396*, on S_m rep el nom de suma acumulada fins al bloc *m* i μ_0 és el valor nominal o objectiu. En aquest cas no es disposa de valor objectiu, i s'ha triat com a μ_0 la mitjana de les llargades mitjanes per a tots els blocs que, a més a més, és la mitjana de les llargades de paraula per tot el *Tirant*. El gràfic mostra com la suma acumulada té una clara tendència a disminuir fins al bloc 312, on hi ha el mínim, per augmentar a partir d'aquest punt. Això indica que, en aquest punt, hi ha un canvi en el valor esperat de la llargada mitjana, i que la llargada mitjana pels blocs anteriors al mínim és inferior a la mitjana global, mentre que pels blocs posteriors al mínim és més gran.

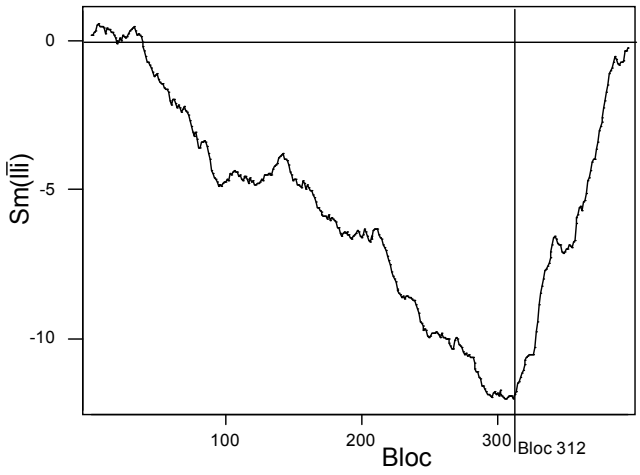

Figura 8.18: Gràfic CUSUM per la llargada mitjana de paraula pels blocs de 1000 paraules. S'ha pres com a valor objectiu la mitjana de les llargades mitjanes per bloc. El mínim de la suma acumulada es té pel bloc 312.

Per tal d'estudiar l'evolució temporal del nombre de paraules de j lletres, per $j=1,2,...,10+$, s'han representat les deu seqüències y_{ji} en els gràfics de control de tipus *NP* de la figura 8.19, en tenir distribució binomial i $N_i=1000$ constant. Sota la hipòtesi de que no hi ha variacions estilístiques al llarg del *Tirant*, tant el valor esperat com la variança per aquestes unitats seran constants.

La categoria que mostra més clarament el canvi al voltant del bloc 313 és la de paraules de deu o més lletres. Els gràfics del nombre de paraules de vuit i nou lletres mostren també amb claredat aquest canvi, tot i que per la llargada 8 el canvi està en el bloc 317. El nombre de paraules que comptem en cadascuna d'aquestes tres categories, augmenta després del bloc 313.

Els gràfics per les llargades intermitges (de quatre a set lletres) no mostren canvis clars en el nivell. Només per la llargada 7 sembla que hi hagi un canvi de nivell, però aquest no apareix fins al bloc 358.

Pel que fa a les llargades curtes (una, dues i tres lletres), per les paraules de dues lletres el canvi de nivell s'avança uns blocs fins al 302, mentre que per les de tres lletres s'endarrereix uns blocs, fins al 325. En ambdós casos després del canvi la mitjana és inferior que abans del canvi, d'acord amb el fet que la llargada mitjana augmenti i amb el fet que la proporció de paraules de vuit, nou i deu o més lletres augmenti. Per les paraules d'una lletra el canvi es dóna cap el bloc 337, augmentant y_{1i} després del canvi.

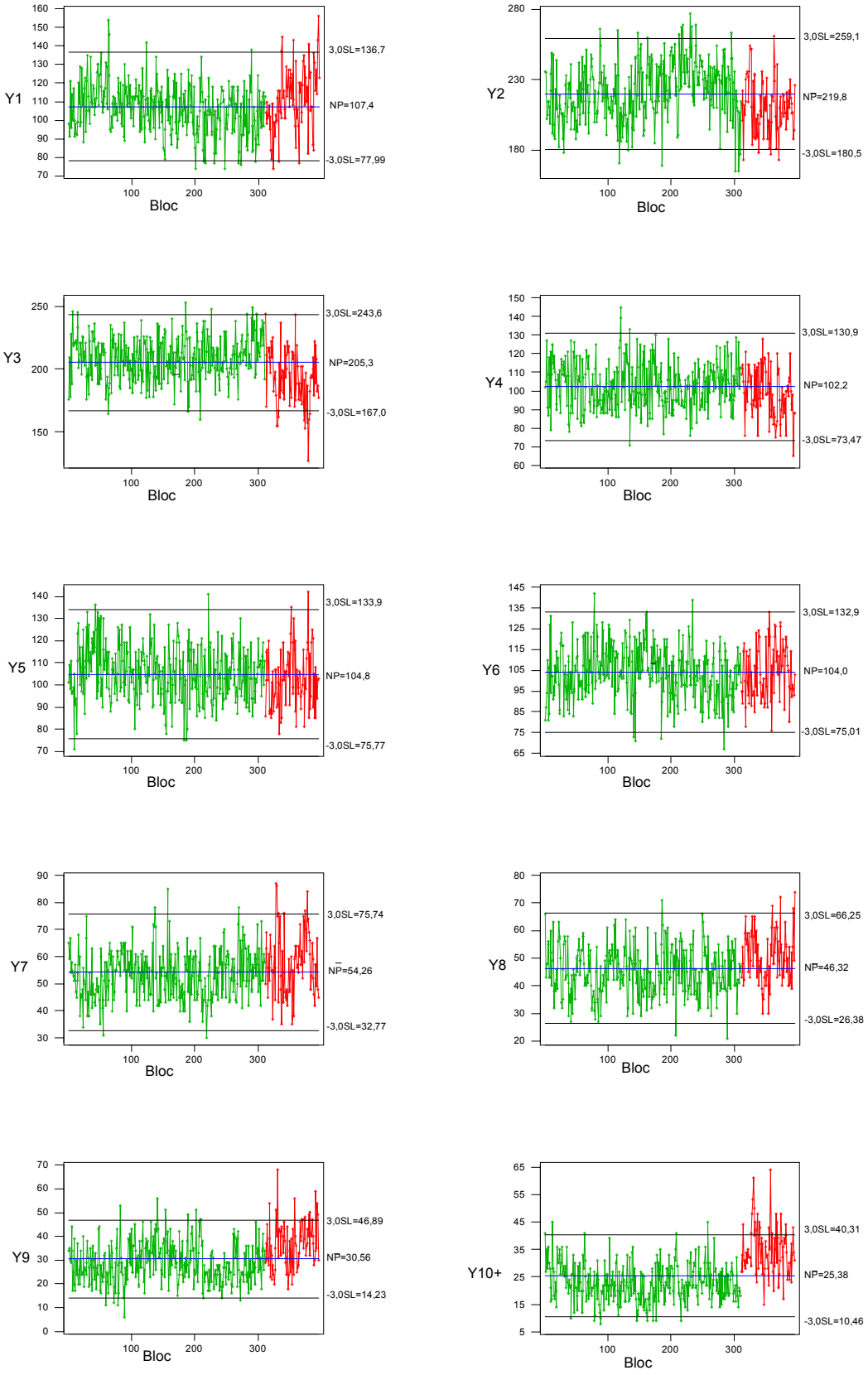

Figura 8.19b: Evolució del nombre de paraules de *j* lletres, per *j=1,2,...,10+*, per blocs de 1000 paraules. El color verd correspon als blocs 1-312 i el vermell als blocs 313-final.

La seqüència $\boldsymbol{y_i}=(y_{1i},y_{2i},y_{3i},...,y_{10+i})$, per $i=1,...,396$, es pot representar en un gràfic de control Chi-quadrat com el de la figura 8.20, tal i com s'ha explicat en el capítol 5. Els valors de la distància χ^2 per cada capítol s'han obtingut prenent, com a valor esperat de la proporció, la proporció promig de paraules de j lletres en tots els capítols de més de 200 paraules:

$$\bar{\pi}_j = \frac{1}{n}\sum_{i=1}^{n}\frac{y_{ji}}{1000},$$

per $n=425$.

En el gràfic es pot observar com el punt de canvi sembla ser prop del bloc 313. S'observa que, com passava per la llargada mitjana de paraula, després del punt de canvi varia el nivell i augmenta la variabilitat. Això provoca que alguns blocs posteriors al 313 tinguin valors de l'estadístic Chi-quadrat amb nivell més proper a la mitjana dels blocs anteriors que a la dels posteriors. Aquesta anàlisi coincideix amb la feta pels capítols de més de 200 paraules.

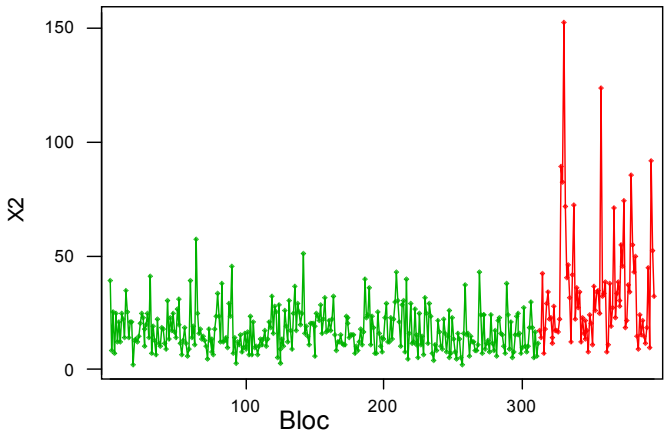

Figura 8.20: Gràfic Chi-quadrat per a la seqüència multinomial de llargades de paraula en els capítols de més de 200 paraules del *Tirant*. Els punts en verd corresponen als blocs des de l'inici fins al 312, mentre que els de color vermell són els del 313 al final.

8.3.4 Anàlisi de correspondències per la llargada de paraula

El punt de partida per a l'anàlisi de la llargada de paraula a través de l'anàlisi de correspondències està format per les distribucions de llargada de paraula tal i com apareixen en la taula de contingència 8.2.

Els resultats de l'anàlisi de correspondències per blocs de 1000 paraules coincideixen amb els que s'han obtingut per als capítols de més de 200 paraules. El gràfic simètric de l'esquerra de la Figura 8.21 mostra com a la primera component les categories estan ordenades per llargada: les paraules curtes es troben a la dreta de l'eix i les llargues en la part esquerra amb l'excepció de les paraules d'una lletra, que no segueixen la "ordenació" al llarg de la primera component.

El gràfic simètric que només mostra les files de la figura 8.21 (dreta) mostra com la majoria de punts que corresponen a blocs posteriors al 313, en vermell al gràfic, tenen

valors de la primera component negatius, indicant que en aquests capítols la proporció de paraules llargues és major que la proporció promig, mentre que la proporció de paraules de dues, tres i quatre lletres és menor que la mitjana, mentre que gairebé tots els punts amb primera component positiva corresponen a blocs anteriors al 312. Per aquests capítols, en general la proporció de paraules llargues és menor i la de paraules de dues, tres i quatre lletres és més gran que la proporció mitjana.

Com ja s'havia observat en l'estudi dels capítols, tant el gràfic simètric com els asimètrics per files i per columnes, com els biplots que mostren en el mateix gràfic files i columnes, són poc útils, el fet de tenir 396 punts en un espai molt reduït en dificulta extraordinàriament la interpretació.

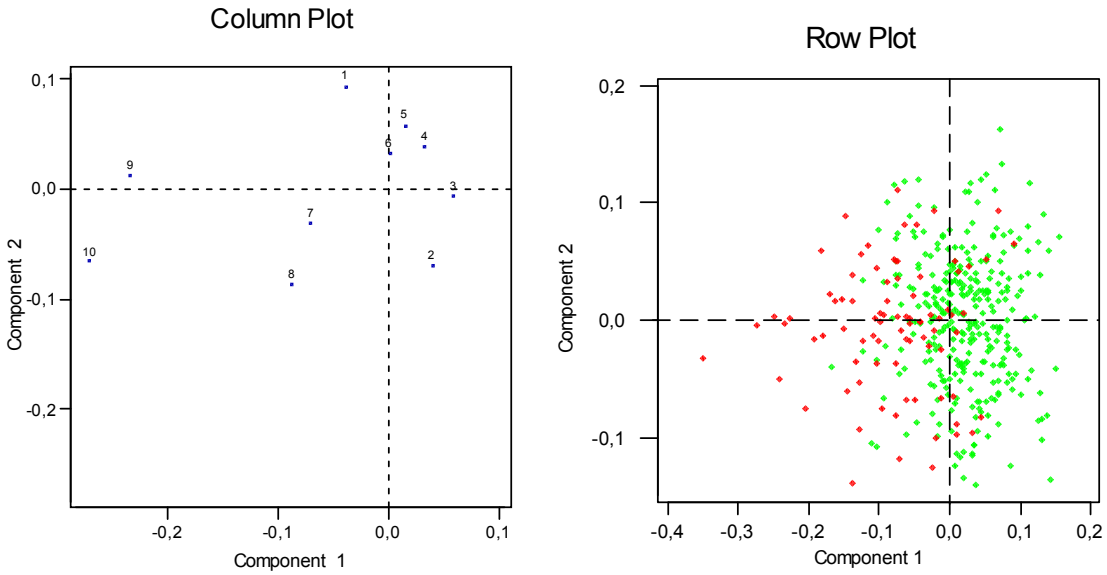

Figura 8.21: Gràfic simètric mostrant columnes de la taula 8.4 a l'esquerra i files a la dreta, per a l'anàlisi de correspondències de la llargada de paraula. S'han representat en color verd els primers 312 blocs i en vermell els blocs 313-396.

El valor esperat de la primera component principal de l'anàlisi de correspondències pot dependre del bloc, però no depèn de la llargada del capítol N_i. En canvi, com en el cas de la llargada mitjana de paraula, la seva variança sí que en depèn, tal i com es pot veure de forma empírica en la figura 8.6.

El valor de l'estadístic χ^2, que mesura el grau d'associació entre blocs i llargades de paraula, per a l'anàlisi de la taula de contingència on les files són els 396 capítols i les columnes el nombre de paraules de *j* lletres, per *j=1,2,...9 i 10+*, és de 7840,02. Com més gran és aquest valor més discrepants són els valors observats dels esperats i, per tant, menys convençuts estem que els perfils columna siguin homogenis al llarg de totes les files. Si es compara aquest valor amb la distribució χ^2 amb $\nu=3555$ graus de llibertat, emprant la aproximació a la $Normal(\nu,(2\cdot\nu)^{-1/2})$, dóna una probabilitat de *0* d'error en rebutjar la hipòtesi nul·la d'homogeneïtat, o, dit d'una altra manera, la hipòtesi que els blocs tenen tots distribucions de llargada de paraula iguals. Cal doncs pensar que hi ha diferències entre els blocs el que ens pot indicar diferències d'estil.

Si l'estil és el mateix en tot el llibre, el valor esperat i la variança de la primera component de l'anàlisi de correspondències no depenen de la llargada. L'evolució temporal de la primera component per files, representada en el gràfic de control per observacions

individuals de la figura 8.22, mostra la gran diferència que existeix entre el valor de la primera component pels blocs posteriors al 327 i els anteriors al 312, i com alguns els blocs que queden en el mig prenen valors de la primera component que els fan semblants als anteriors al 312 i els altres són semblants als posteriors al bloc 327.

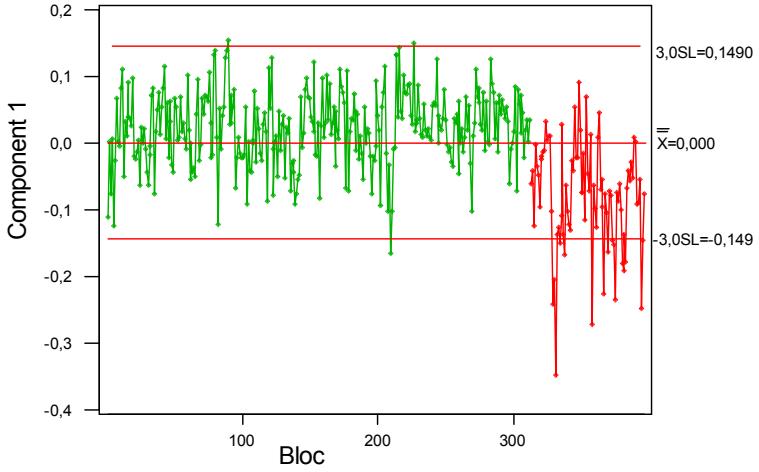

Figura 8.22: Gràfic de control d'observacions individuals per a la 1ª component de l'anàlisi de correspondències. El color verd correspon als blocs anteriors al bloc 313 i el color vermell als blocs posteriors.

8.3.5 Estimació del punt de canvi

La determinació del punt de canvi s'ha fet servir la mateixa metodologia emprada per als capítols. En primer lloc es fa servir el mètode proposat per seqüències de Normals, i s'aplica a les seqüències de llargades mitjanes de paraula i del valor de la primera component de l'anàlisi de correspondències. A continuació, s'obtenen deu estrimacions del punt de canvi a partir de les deu seqüències de proporcions de paraules de *j* lletres, per *j=1,2,...,10+*, per separat mitjançant l'aproximació al problema basada en models logístics. Per últim, s'estima el punt de canvi en la seqüència de dades multinomials contingudes en la taula 8.4 ajustant models politòmics.

8.3.5.1 Punt de canvi en seqüències Normals

S'ha estimat el punt de canvi per la llargada mitjana de paraula fent servir les tècniques proposades per a dades Normals. Tot i que la distribució de la llargada és discreta i, per tant, no és normal, les hipòtesis en les que es sustenten els models lineals s'acompleixen de forma aproximada, gràcies al teorema central del límit.

Suposem que les llargades mitjanes de paraula, $\overline{ll}_i$, són independents i es distribueixen:

$$\overline{ll}_i \sim N(\beta_0^{(r)} + \beta_1^{(r)} Ind_{1i}^{(r)}, \sigma^2),$$

on $Ind_i^{(r)}$ és una variable indicadora que pren valor *0* per *i=1,2,..,r*, i pren valor *1* per *i=r+1,...,n=396*. La hipòtesi de variança constant aquí és sustenta en el fet que tots els blocs són de 1000 paraules. 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 amb variable indicadora $Ind_{1i}^{(r)}$.

El gràfic de la figura 8.23 mostra l'evolució de F_r en funció de r. El màxim de F_r i, per tant, la millor estimació del punt de canvi per a la seqüència de llargades mitjans per bloc es té per $\hat{r}_N$ =312, que correspon al capítol 346. El gràfic mostra un màxim a $\hat{r}_N$ clarament diferenciat.

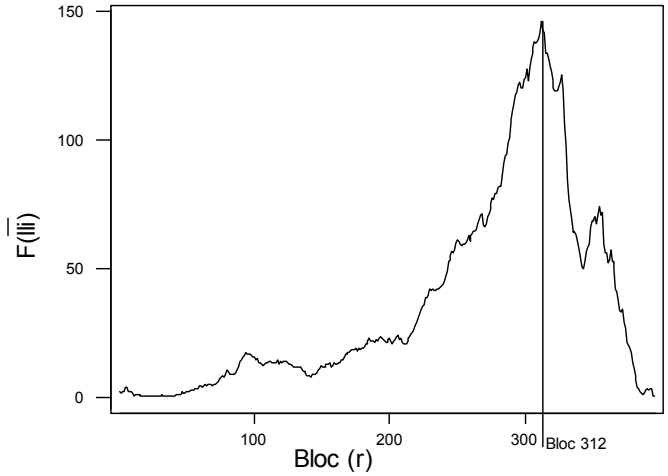

Figura 8.23: Gràfic que representa l'evolució de F_r en funció de r per a l'estimació del punt de canvi en la seqüència de llargades mitjanes de paraula per blocs de 1000 paraules. El màxim de F_r es té per *r=312*, que correspon al capítol 346.

Es fa servir el mateix procediment per estimar el punt de canvi en la seqüència de valors de la primera component de l'anàlisi de correspondències realitzat a l'apartat 8.3.4. El màxim de F_r es troba per *r=327*, que correspon al capítol 371, tot i que també hi ha un màxim local, amb valor de F_r no gaire inferior a l'anterior, per a la *r=312*, tal i com es pot observar en el gràfic de la figura 8.24.

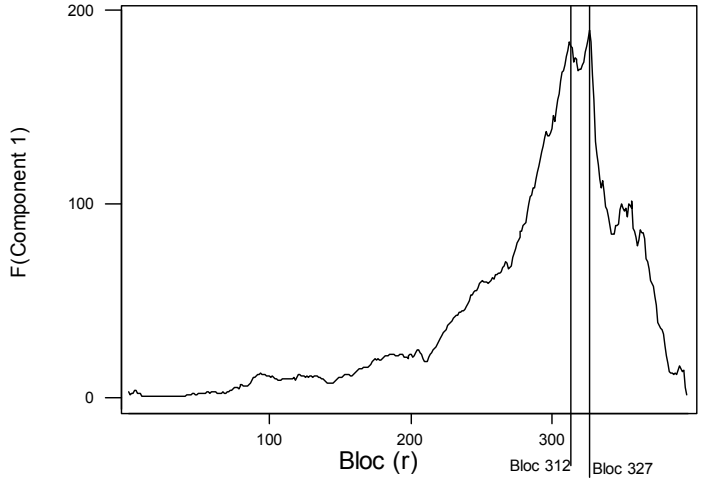

Figura 8.24: Evolució de F_r en funció de r per a la seqüència de valors de la primera component de l'anàlisi de correspondències. El màxim de F_r es té per *r=327*, que correspon al capítol 371.

8.3.5.2 Punt de canvi en les seqüències de binomials

En aquest apartat s'estudien per separat les seqüències de nombres de paraules en cadascuna de les 10 categories en que s'han classificat les paraules en funció de la seva llargada. Aquest procediment dóna 10 punts de canvi, possiblement diferents.

Suposant que existeix un punt de canvi a *r*, per una categoria *j* fixada, amb *j=1,2,...10+*, y_{ij} té una distribució *binomial* $(N=1000,\pi_{ja})$ per *i=1,...,r* i *binomial* $(N=1000,\pi_{jd})$ per *i=r+1,...,396*. Per cadascuna de les 10 categories s'ajusta el model logístic (6.2):

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

La figura 8.25 conté els gràfics de $L_j\left(r, \hat{\beta}^{(r)}\right)$, el màxim del logaritme de la versemblança, en funció de *r* per a les 10 categories en què s'ha subdividit la llargada de paraula.

Les 10 estimacions del punt de canvi són diferents. Per les seqüències de paraules curtes, d'una, dues i tres lletres, i llargues, de vuit ,nou i deu o més lletres, els màxims es troben entre els capítols 301 i 336, mentre que per les llargades intermitges, de quatre a set lletres, els màxims cauen fora d'aquest interval.

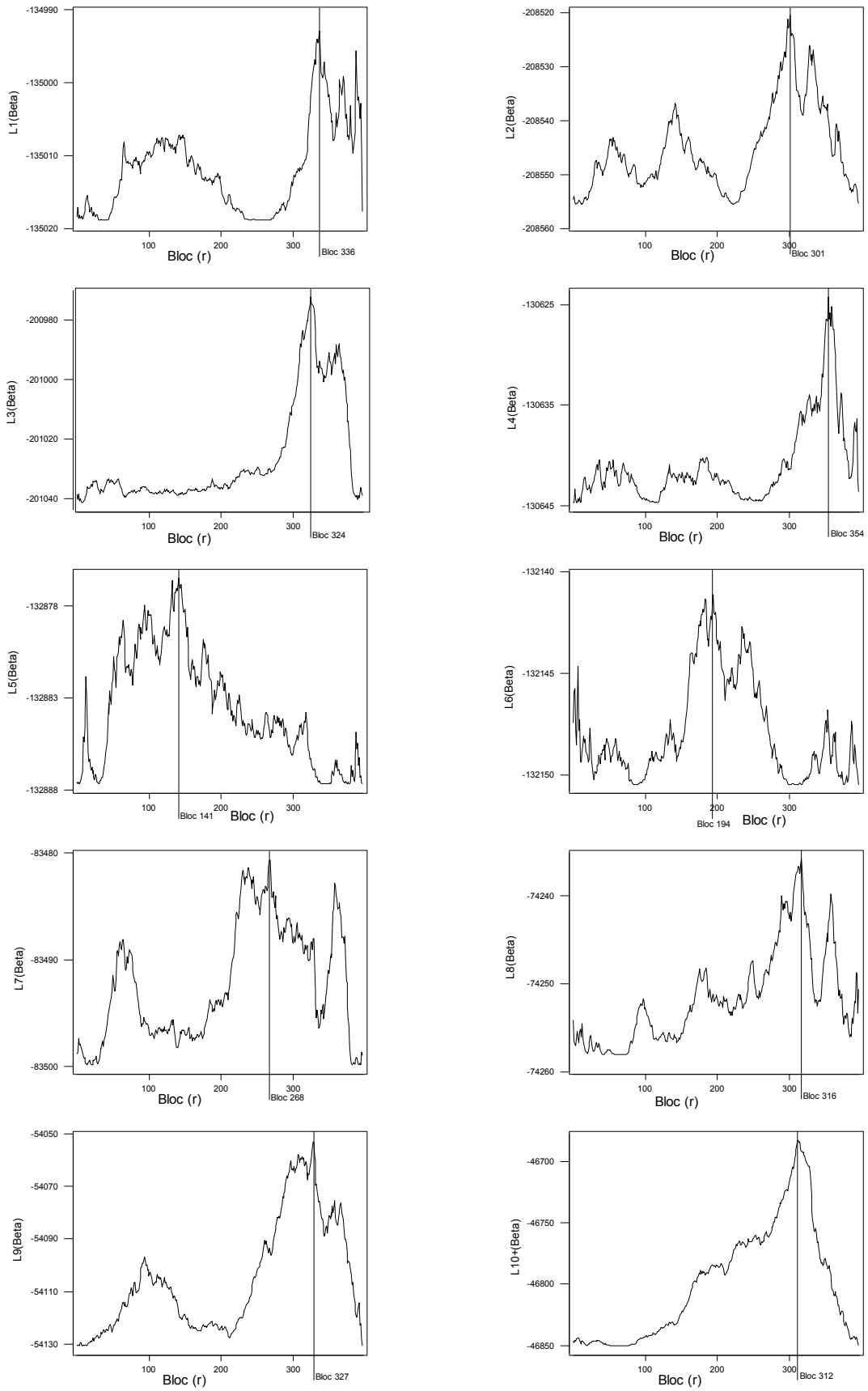

Figura 8.25: Evolució del màxim logaritme de la versemblança del model logístic en funció de *r* per a les seqüències de paraules de *j* lletres, per *j=1,2,...,10+*. La línia vertical indica el punt de canvi estimat per cadascuna de les llargades.

8.3.5.3 Punt de canvi en seqüència de multinomial; model politòmic

Per a estimar el punt de canvi en la seqüència multinomial $\boldsymbol{y_i}=(y_{1i},y_{2i},y_{3i},...,y_{10+i})$, per $i=1,...,425$, s'ha seguit la mateixa metodologia que pels capítols: en primer lloc s'ha fet servir l'ajust de models politòmics, i posteriorment a s'ha estimat a partir de les 10 estimacions individuals obtingudes a 8.2.5.2.

Si $\boldsymbol{y_i}$ està distribuït $Mult(N=1000,\pi_i)$, on $\pi_i=(\pi_{1i},\pi_{2i},...,\pi_{10+i})$ és el vector amb les probabilitats que una unitat sigui assignada a cadascuna de les deu categories. El model politòmic suposa que:

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

per $j=1,...,10+$, i on $g_j(\cdot)$ suposarem que és una generalització de la funció *link logit* emprada per dades binomials, i descrita a l'apartat 6.3.1, al cas multinomial. En general tindrem:

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

S'estima el punt de canvi, r, com aquella r per la que el màxim del logaritme de la versemblança, $L\left(r,\hat{\beta}^{(r)}\right)$, és màxim. El gràfic de la figura 8.26 mostra l'evolució del màxim del logaritme de la versemblança en funció de r. L'estimació del punt de canvi és $\hat{r}_M=327$, tot i que hi ha un màxim local de valor semblant al màxim global per $r=312$.

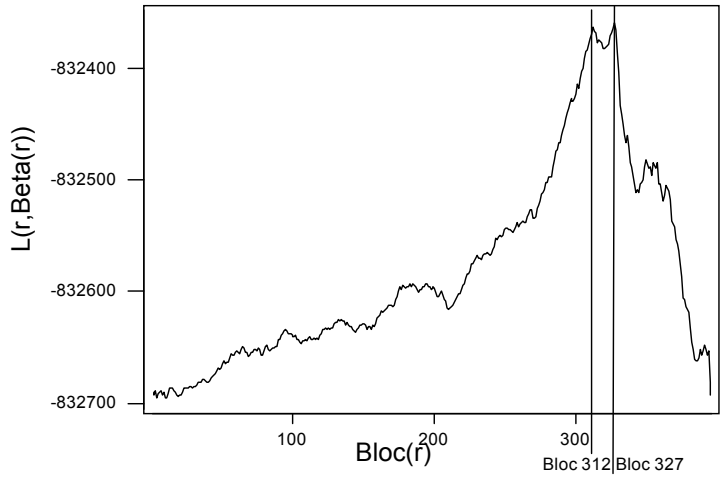

Figura 8.26: Gràfic de l'evolució de $L\left(r,\hat{\beta}^{(r)}\right)$ en funció de r. El màxim del màxim es té per $\hat{r}_M=327$, tot i que s'observa un màxim local per r=312.

Com s'ha descrit en el capítol 6 i ja s'ha aplicat en el cas dels capítols de més de 200 paraules, estimar r en una seqüència multinomial via l'ajust de models de regressió politòmica és equivalent a ajustar-lo via màxim versemblant. Aquesta és el manera correcta de fer-ho i la més elegant.

8.3.5.4 Punt de canvi en seqüència multinomial; combinació de binomials

Una manera alternativa d'estimar el punt de canvi en una seqüència de multinomials és calculant les deu estimacions individuals en cadascuna de les deu seqüències binomials de paraules de *j* lletres, obtingudes a 8.3.4.2. Això es pot fer adaptant els estadístics proposats per Wolfe i Chen (1990), que resumeixen les *l=10* estimacions en les seqüències binomials en un sol valor. Tres estimadors alternatius de *r*, proposats en el capítol 6 de la tesi, que corresponen, respectivament, a les fórmules (6.8), (6.9) i (6.10) i que resumeixen les *l=10* estimacions en un sol valor, són $\hat{r}_{ML1}$, $\hat{r}_{ML2}$ i $\hat{r}_{ML3}$.

La figura 8.27 mostra l'evolució de:

$$\max_{1\le j\le 10} L_j\left(r,\hat{\beta}^{(r)}\right),$$

$$\sum_j L_j\left(r,\hat{\beta}^{(r)}\right)$$

i

$$\sum_j L_j^2\left(r,\hat{\beta}^{(r)}\right),$$

en funció de *r*. Els màxims dels tres gràfics indiquen $\hat{r}_{ML1}$, $\hat{r}_{ML2}$ i $\hat{r}_{ML3}$, respectivament. Tal i com s'observa en la figura $\hat{r}_{ML1}$=312, $\hat{r}_{ML2}$=327 i $\hat{r}_{ML3}$=327, que corresponen, respectivament, als capítols 345, 371 i 371.

Les estimacions del punt de canvi no coincideixen amb les que donen els estimadors proposats per Wolfe i Chen (1990), tot i que són força semblants. Els estimadors que s'obtenen són, respectivament $\hat{r}_{W1}=324$, $\hat{r}_{W2}=327$ i $\hat{r}_{W3}$=327.

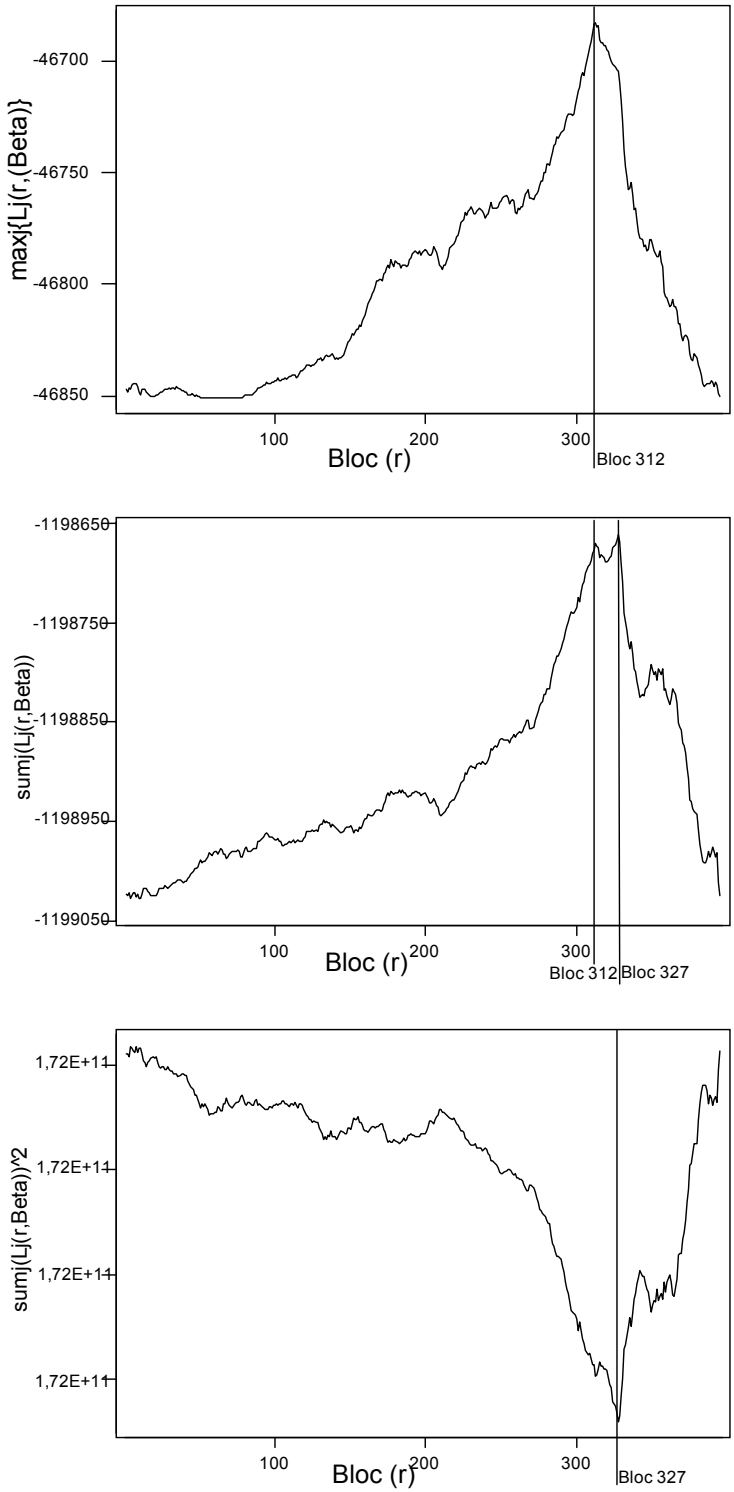

Figura 8.27: Gràfics que representen l'evolució de $\max_{1\le j\le 10} L_j\left(r,\hat{\beta}^{(r)}\right)$, $\sum_j L_j\left(r,\hat{\beta}^{(r)}\right)$ i de $\sum_j L_j^2\left(r,\hat{\beta}^{(r)}\right)$, en funció de r. Els tres extrems corresponen a $\hat{r}_{ML1}$, $\hat{r}_{ML2}$ i $\hat{r}_{ML3}$.

8.3.5.5 Conclusions

La conclusió d'aquesta anàlisi és que s'observa l'existència d'un punt de canvi en la seqüència de llargades de paraula. L'estimador del punt de canvi no és sempre el mateix, depenent del mètode o la unitat analitzada, però sempre balla entre el bloc 312 i el 327, és a dir, entre els capítols 345 i 371. Això podria indicar que en la zona entre aquests dos blocs es barregen els dos estils. Els resultats coincideixen amb el que havíem obtingut en l'estudi de la llargada de paraula en els capítols.

En general, en els primers capítols la llargada mitjana de paraula és mes breu que al final, hi abunden més paraules de dues i tres lletres mentre que són menys abundants les paraules d'una i de set o més lletres.

La taula 8.5 mostra un quadre resum amb les estimacions del punt de canvi obtingudes en l'estudi de la llargada de paraula pels capítols de més de 200 paraules.

Distribució	**Seqüència**	**Punt de canvi ppal.**	**Punt de canvi sec.**
Normal	*Llargada mitjana*	$\hat{r}_N$ =312	
	1ª Comp. A. Corresp.	$\hat{r}_N$ =327	312
Binomial	*1 lletra (y_{1i})*	$\hat{r}_{L_1}$ =336	
	2 lletres (y_{2i})	$\hat{r}_{L_2}$ =301	
	3 lletres (y_{3i})	$\hat{r}_{L_3}$ =324	
	4 lletres (y_{4i})	$\hat{r}_{L_4}$ =354	
	5 lletres (y_{5i})	$\hat{r}_{L_5}$ =141	
	6 lletres (y_{6i})	$\hat{r}_{L_6}$ =194	
	7 lletres (y_{7i})	$\hat{r}_{L_7}$ =268	
	8 lletres (y_{8i})	$\hat{r}_{L_8}$ =316	
	9 lletres (y_{9i})	$\hat{r}_{L_9}$ =327	
	10 o més lletres (y_{10+i})	$\hat{r}_{L_{10+}}$ =312	
Multinomial	*Combinació de binom.* $\max_{1\le j\le 10} L_j\left(r,\hat{\beta}^{(r)}\right)$	$\hat{r}_{ML1}$ =312	
	Combinació de binom. $\sum_j L_j\left(r,\hat{\beta}^{(r)}\right)$	$\hat{r}_{ML2}$ =327	345
	Combinació de binom. $\sum_j L_j^2\left(r,\hat{\beta}^{(r)}\right)$	$\hat{r}_{ML3}$ =327	
	Model Politòmic	$\hat{r}_M$ =327	312

Taula 8.5: Quadre resum de les estimacions del punt de canvi obtingudes en l'anàlisi de la llargada de paraula per als capítols de més de 200 paraules. Els blocs 312 i 327 corresponen, respectivament, als capítols 345 i 371.

8.3.6 Anàlisi cluster de les files d'una taula de contingència

De l'anàlisi gràfica de la llargada de paraula en els blocs, com ja s'ha comentat pels capítols, s'observa com alguns blocs posteriors al punt de canvi tenen uns trets que els fan més semblants als anteriors que a la majoria dels posteriors, es a dir, hi ha la possibilitat que hi hagi barreja d'estils després del punt de canvi. Aquesta situació es dóna per totes les variables que hem considerat relacionades amb la llargada de paraula..

Per aquest motiu, s'ha decidit fer una anàlisi cluster, per veure si és possible recol·locar algun bloc del final. Es faran servir les tècniques exposades en el capítol 7 d'anàlisi cluster, basades en dos criteris d'ajust de models per dades politòmiques: χ^2 i deviança. El punt de partida de l'anàlisi és la taula de contingència continguda en la taula 8.4, on les files són els 396 blocs de *1000* paraules i les 10 columnes representen els nombres de paraules d'una, dues, tres, ..., nou i deu o més lletres.

8.3.6.1 Anàlisi cluster basada en la distància χ^2

S'ha executat més de 1000 vegades l'algorisme cluster basat en la distància χ^2 amb assignacions inicials aleatòries i els resultats han convergit sempre a la mateixa solució. A l'annex A8.2 hi ha les assignacions dels blocs a cadascun dels 2 clusters. En total, 51 blocs anteriors al 312 i 24 posteriors han canviat d'assignació. Pel que fa als blocs entre el 313 i el 327 veiem com han sigut assignats a ambdós clusters: els blocs entre el 313 i el 319, amb l'excepció del 316, han sigut assignats al cluster dels blocs del final, mentre que els altres han sigut assignats al cluster dels blocs del principi. El gràfic de la figura 8.28 mostra l'assignació de blocs a clusters.

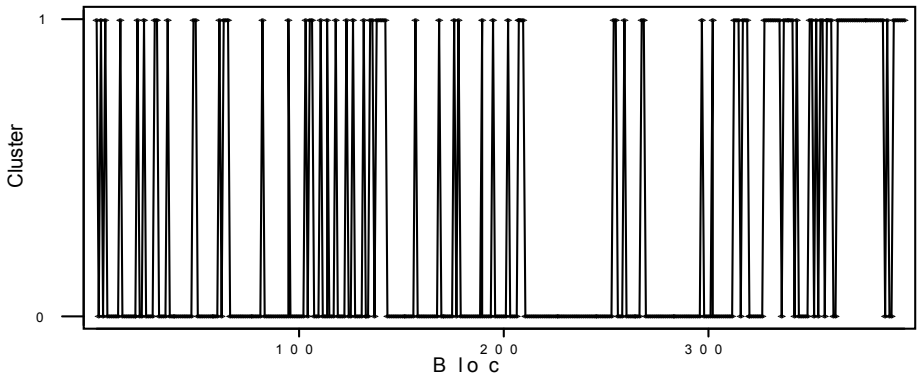

Figura 8.28: Assignació dels blocs a un dels dos clusters pel mètode basat en la distància χ^2 per a la distribució de llargades de paraula en els blocs de 1000 paraules del *Tirant*.

L'estadístic χ^2 per la taula de contingència completa val $\chi^2=7840$, mentre que el valor de l'estadístic entre clusters val $\chi^2_B = 1341{,}65$, el que dóna una relació $\chi^2_B / \chi^2 = 0{,}171$, és a dir la subdivisió de blocs en aquests dos clusters explica el 17,1% de la "no-homogeneïtat" entre tots els blocs en funció de la distribució de la llargada de paraula. Aquest valor, tal i com s'ha avançat en el capítol 7, és menor que la proporció de inèrcia explicada per la primera component principal, que és del 28%.

Els clusters obtinguts de l'aplicació de l'algorisme separen els capítols del *Tirant* en funció de la primera component de l'anàlisi de correspondències, realitzat a l'apartat 8.3.4, com es pot apreciar del gràfic de la figura 8.29.

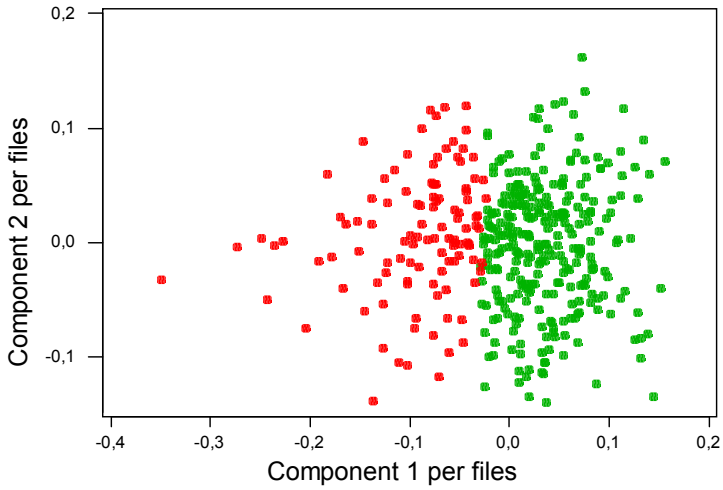

Figura 8.29: Projecció dels perfils fila en el pla format per les dues primeres components de l'anàlisi de correspondències realitzat a la 8.3.4. Els punts en vermell corresponen als blocs assignats al cluster 0, el que conté majoritàriament els capítols del principi del llibre, en l'anàlisi basada en la distància χ^2, i els punts en vermell als blocs assignats al cluster 1, el que conté els capítols del final.

S'ha fet una validació dels resultats obtinguts en l'anàlisi cluster. S'ha ajustat, el model politòmic:

$$g_j(\pi_i) = \log\left(\frac{\pi_{ji}}{\pi_{1i}}\right) = \beta_{j0}^{(c)} + \beta_{j1}^{(c)} Ind_{1i}^{(c)}, \qquad 0$$

on $Ind_{1i}^{(c)}$ és una variable indicadora que pren valor *0* pels blocs assignats al cluster *0* i pren valor *1* pels blocs assignats al cluster *1*, i es calcula el màxim de la versemblança del model, $L_c\left(\hat{\beta}^{(c)}\right)$.

De forma anàloga s'ajusta, per cadascuna de les *l=10* categories, el model logístic:

$$y_{ji} \sim Binomial\left(N_i, \pi_{ji} = \frac{e^{\beta_{j0}^{(c)} + \beta_{j1}^{(r)} Ind_{1i}^{(c)}}}{1 + e^{\beta_{j0}^{(c)} + \beta_{j1}^{(c)} Ind_{1i}^{(c)}}}\right),$$

on $Ind_{1i}^{(c)}$ és la mateixa variable indicadora. Per cadascun, d'aquests models es guarda el màxim del logaritme de la versemblança, $L_j\left(\hat{\beta}^{(c)}\right)$, per *j=1,...,10+*, i es calculen els estadístics:

$$\max_{1 \le j \le 10} L_j\left(\hat{\beta}^{(c)}\right),$$

$$\sum_{j=1}^{10} L_j\left(\hat{\beta}^{(c)}\right),$$

i

$$\sum_{j=1}^{10} L_j^2\left(\hat{\beta}^{(c)}\right).$$

La comparació de les bondats d'ajust del model que suposa un únic punt de canvi i el que suposa l'existència de clusters, calculats fent servir la distància χ^2 es mostren el la taula 8.6. En ella es pot veure com l'anàlisi cluster millora, tot i que de forma poc notable, els estadístics, pel que es pot assegurar que l'estimació del punt de canvi proposa un bona separació en grups.

Model Ajustat	Estadístic	Punt de Canvi	Cluster χ^2
Politòmic	$L\left(r, \hat{\beta}^{(r)}\right)$	-832.055	-832.358
10 logístics	$\max_{1\le j\le l} L_j\left(r, \hat{\beta}^{(r)}\right)$	-46.648	-46.682
10 logístics	$\sum_{j=1}^{l} L_j\left(r, \hat{\beta}^{(r)}\right)$	-1.198.326	-1.198.660
10 logístics	$\sum_{j=1}^{l} L_j^2\left(r, \hat{\beta}^{(r)}\right)$	171.799.365.846	171.871.976.621

Taula 8.6: Comparació dels estadístics emprats per calcular la bondat d'ajust del model que suposa un punt de canvi i dels models que suposen l'existència de clusters calculada a partir de la distància χ^2. Els primer estadístic és suposa que el punt de canvi és $\hat{r}_M$, i els altres tres suposen que és $\hat{r}_{ML1}$, $\hat{r}_{ML2}$ i $\hat{r}_{ML3}$.

Analitzant com queden la llargada mitjana de paraula i la primera component de l'anàlisi de correspondències després de l'anàlisi cluster basat en la distància χ^2 s'observa com el cluster agrupa els blocs segons el valor de la primera component de l'anàlisi de correspondències. A la figura 8.30 s'hi ha representat l'evolució temporal d'aquestes dues variables, representant en colors diferents els capítols assignats a clusters diferents. La majoria de capítols anteriors al punt de canvi pertanyen al cluster en verd, mentre que la majoria dels blocs posteriors, amb excepcions, pertanyen al cluster en vermell en la figura.

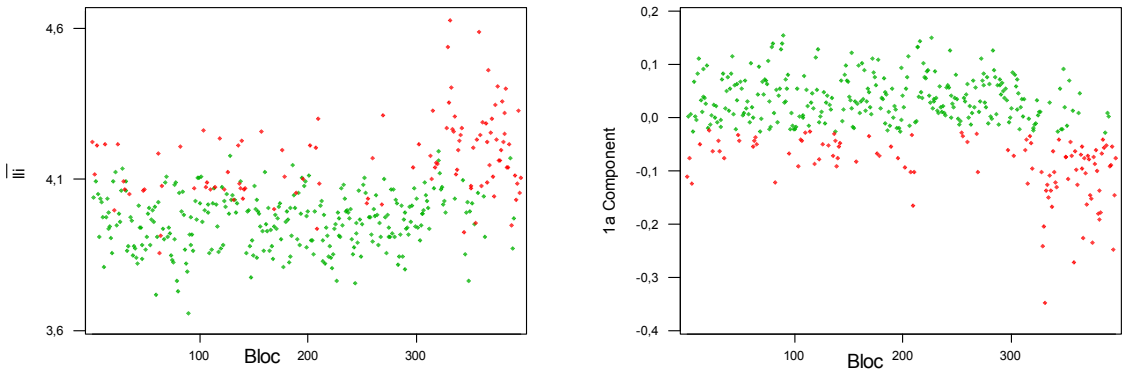

Figura 8.30: evolució temporal de la llargada mitjana de paraula i del valor de la primera component principal de l'cnàlisi de correspondències per les files després de l'anàlisi cluster. Els punts en verd i en vermell corresponen als blocs classificats en un dels dos clusters.

8.3.6.2 Anàlisi cluster basada en la deviança

L'execució de l'algorisme cluster basat en la deviança de models de regressió politòmica porta a resultats molt semblants als obtinguts amb l'algorisme basat en la distància χ^2, com ja passava pels capítols. Hi ha 15 blocs que l'algorisme basat en models classifica en el cluster que conté majoritàriament els blocs del final del llibre, que l'algorisme basat en la distància χ^2 assigna a l'altre cluster. Aquests blocs són: 31, 68, 77, 78, 96, 98, 100, 101,

128, 194, 196, 308, 343, 347, 387. Es pot apreciar com només tres d'ells són posteriors a l'estimació del punt de canvi.

El guany que s'obté amb aquest mètode respecte a l'aplicació de l'algorisme cluster basat en la distància χ^2, en termes del màxim del logaritme de la versemblança, és extraordinàriament petit.