CAPÍTOL 12

ÚS DE PARAULES EN EL *TIRANT*

Índex Capítol

CAPÍTOL 12

ÚS DE PARAULES EN EL *TIRANT*

12.1 Introducció

L'anàlisi de l'ús de paraules ofereix grans possibilitats per a la discriminació entre estils i entre autors. Es sabut que la proporció d'algunes paraules varia molt en diferents obres d'un mateix autor, sovint depenent de l'argument, mentre que altres paraules presenten molta estabilitat en totes les seves obres. A l'hora de discriminar entre autors calen paraules tals que la seva presència sigui tan independent com es pugui del context, per no confondre l'efecte de l'autor amb diferències en l'argument, que anomenem *paraules eina*. S'ha dedicat l'apartat 2.2.2 a llistar exemples en els que les *paraules eina* s'han emprat per ajudar a discriminar entre autors. En el aquest capítol es faran servir per estudiar la homogeneïtat d'estil al llarg del *Tirant*.

L'estudi s'ha fet tant pels capítols de més de 200 paraules com pels blocs de 1000 paraules consecutives. Com ja s'ha explicat en la descripció de la base de dades del capítol 4, s'ha comptat el nombre de vegades que apareix cadascuna de les formes (paraules) en cada capítol i en cada bloc. en l'annex A12.1 hi ha un llistat amb el nombre de vegades que apareixen les 104 paraules més abundants en el *Tirant*.

Tant per capítols com per blocs s'ha fet primer una anàlisi exploratòria de les proporcions d'ús de les paraules més freqüents, per observar si es detecta una frontera d'estil. Aquesta anàlisi fa servir els tant gràfics de control com l'anàlisi de correspondències. Pel cas dels capítols, a més a més, s'ha estudiat la relació entre el nombre d'ocurrències de les paraules eina 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 dos conjunts homogenis, a partir dels perfils fila descrits en el capítol 5. A partir de l'estimació del punt de canvi es buscaran, entre les 100 paraules més freqüents, les que ajuden a discriminar entre estils en el *Tirant*. Per aquesta part, no es consideraran el substantius, en ser sensibles al context.

Es consideraran per a l'estudi les 25 paraules lliures de context més abundants en el llibre: *e, de, la, que, lo, en, a, per, no, l, los, com*, *ab, les, d, li, qui, del, se, molt, és, gran, jo, si, dix*. La tria és arbitrària, i s'ha basat en un compromís entre la manejabilitat de la taula de contingència, de forma que no hi hagués "massa" paraules, i la representativitat, de forma que no n'hi hagués massa poques. Del llistat de paraules més freqüents al llarg del llibre, no s'han considerat per a l'estudi de l'homogeneïtat d'estil dues que es sospita que el seu ús depèn del context: *Tirant* i *rei*. La paraula *Tirant* apareix 2.943 vegades, és la setzena més freqüent, i no apareix per primer cop fins al capítol 29, *Rei* és la vint-i-cinquena paraula més freqüent i apareix 1.929 vegades.

L'estudi es fa en paral·lel per a les dotze paraules eina més freqüents (de *e* a *com*) i per a les vint-i-cinc paraules eina més freqüents (de *e* a *dix*).

12.2 Estudi de les paraules eina per Capítols

Es consideren només els capítols de més de 200 paraules.

12.2.1 Notació i forma de les dades

La taula 12.1 mostra el tros corresponent als capítols 350-400 de la taula que es fa servir per a l'anàlisi de l'ús de les 12 paraules més abundants. La taula sencera té les 425 files que corresponen a capítols de més de 200 paraules, i 15 columnes: la primera indica el número de capítol, la segona el nombre total de paraules, N_i, les dotze centrals formen una taula de contingència que conté el nombre d'ocurrències de cadascuna de les dotze paraules en cada capítol, i la darrera columna conté la suma del nombre de vegades que apareixen totes les altres paraules en el capítol, de manera que la suma del contingut d'aquesta columna més les dotze anteriors és N_i. Aquesta darrera columna s'ha etiquetat com a *Resta.* Per exemple, a la taula podem llegir com pel capítol 351 hi ha un total de $N_i=428$ paraules, la paraula *e* hi apareix 23 vegades, *de* apareix 18 vegades, *la* 18 vegades, ... *com* hi apareix 4 vegades, i la suma de totes les altres paraules etiquetada com *Resta* és 134. La taula que es farà servir per a 25 paraules eina és anàloga, amb 25 columnes centrals i valor per a la columna *Resta* diferent.

Per cada paraula eina j existeix un valor π_j que és la proporció de vegades que apareixeria en un text del mateix autor de llargada infinita, sota la hipòtesi que tant la llista de paraules com els valors π_j són constants al llarg de l'obra de l'autor en estudi, suposició que pot ser discutible quan els textos de l'autor comprenen èpoques o gèneres molt diferents. Denotem per y_{ji} el nombre d'ocurrències de la paraula j-èssima en el capítol i-èssim i per $\hat{\pi}_{ji} = y_{ji}/N_i$ la proporció. Cadascuna de les N_i paraules d'un capítol pot ser classificada en una de les $l+1$ categories, amb $l=12$ *o* 25. La categoria addicional és la formada per la variable etiquetada com a *Resta*.

$\boldsymbol{y_i}=(y_{1i},y_{2i,\ldots},y_{l+1i})$ satisfà:

$$\sum_{j=1}^{l+1} y_{ij} = N_i \ ,$$

i es distribueix: $Mult(N_i,\pi_i)$, on $\pi_i=(\pi_{1i},\pi_{2i},...,\pi_{l+1i})$ és el vector tal que π_{ji} és la probabilitat de que una paraula agafada a l'atzar del capítol i-èssim pertanyi a cadascuna de les $l+1$ categories, per a $i=1,...,425$.

Cap.	Ni	e	de	la	que	lo	en	a	per	no	l	los	com	Resta
350	1117	61	39	52	47	20	20	26	16	7	9	10	19	326
351	428	23	18	18	21	10	7	9	6	9	2	7	4	134
352	217	14	9	3	6	3	5	6	3	7	1	6	2	65
353	738	33	29	21	29	9	12	14	14	14	7	10	7	199
354	852	51	37	32	24	13	16	14	21	21	4	7	8	248
355	1556	71	61	55	39	26	38	26	20	31	15	13	18	413
356	428	13	13	13	20	8	8	5	11	11	0	4	3	109
357	1002	48	41	45	31	18	24	15	20	17	12	17	10	298
358	279	14	8	10	10	9	5	4	4	1	1	1	4	71
359	311	22	8	9	16	11	7	3	4	8	3	3	3	97
360	416	20	9	17	17	13	4	5	6	4	4	3	4	106
362	446	23	16	25	12	6	12	7	7	6	6	3	4	127
363	244	12	15	9	8	4	4	1	3	4	7	2	4	73
364	349	11	17	8	10	12	8	7	8	6	0	3	4	94
365	261	7	10	14	14	6	4	3	4	4	1	0	2	69
366	579	40	24	28	17	9	21	9	11	7	8	6	7	187
367	579	40	24	28	17	9	21	9	11	7	8	6	7	187
368	331	20	6	12	19	9	8	6	5	7	2	2	2	98
369	463	7	27	10	11	7	9	7	11	15	6	7	2	119
370	505	29	28	26	18	9	6	11	13	8	5	3	3	159
371	750	36	27	31	35	11	17	14	15	10	3	7	12	218
372	892	47	34	39	36	17	19	16	13	10	13	8	8	260
373	388	23	16	18	19	1	6	7	6	8	5	1	6	116
374	1441	86	49	57	39	12	26	28	17	34	10	15	11	384
375	357	33	9	15	7	4	4	8	1	8	0	6	3	98
376	752	55	24	25	11	12	13	12	12	15	10	7	6	202
377	303	25	14	13	2	6	5	1	7	9	4	5	3	94
378	523	27	14	14	15	9	11	8	11	16	5	10	2	142
379	230	8	8	8	4	4	7	0	3	9	4	3	1	59
380	414	10	16	17	7	9	12	2	8	10	5	8	4	108
381	272	6	12	10	9	6	2	8	6	4	1	4	0	68
382	330	19	14	8	9	9	2	9	8	5	3	3	5	94
383	309	25	21	10	9	3	8	3	5	1	7	2	2	96
384	512	39	22	18	10	15	16	9	13	2	2	9	6	161
385	287	14	16	14	6	9	3	4	6	1	6	7	1	87
386	588	38	29	32	17	17	5	7	9	6	3	11	7	181
387	1639	92	70	57	63	45	30	25	27	23	14	37	12	495
388	357	24	14	15	6	10	6	10	7	1	19	3	6	121
390	454	28	19	17	17	14	9	7	11	3	3	3	8	139
391	333	25	21	18	15	2	5	8	1	3	16	2	6	122
392	287	20	15	12	7	2	4	9	1	3	7	3	10	93
393	327	29	18	19	9	8	5	8	5	1	14	1	2	119
394	1059	62	38	51	19	22	14	19	29	15	8	20	11	308
395	257	24	9	11	7	4	3	6	2	0	3	4	6	79
397	356	25	9	14	21	9	12	4	6	7	6	2	10	125
399	203	17	8	10	6	3	10	2	5	2	5	4	3	75
400	362	16	12	15	6	11	7	8	6	3	1	5	1	91

Taula 12.1: Part corresponent als capítols 350-400 de les dades utilitzades en l'estudi de l'ús de 12 *paraules eina* per capítols. Cada fila correspon a un capítol, la primera columna dóna el número de capítol, la segona el nombre de paraules per capítol N_i, les següents el nombre de vegades que apareixen les *12* paraules més abundants per capítol, i la darrera, *Resta*, és la suma de totes les altres paraules.

12.2.2 Relació de la proporció de paraules eina amb N_i

Per cada paraula i per cada capítol tenim que:

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

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

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

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

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

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

A nivell empíric, observem la relació de y_{ji} i de $Var(y_{ji})$ amb N_i en la figura 12.1: en el gràfic de l'esquerra es mostra com, per la paraula *e,* y_{ei} i N_i tenen una relació lineal molt forta, mentre que el gràfic de la dreta mostra la relació de y_{ei}/N_i amb N_i. En el gràfic s'observa com el valor esperat de y_{ei}/N_i no es veu afectat per la llargada del capítol mentre que la variança és proporcional a $1/N_i$, és a dir, que és més gran pels capítols més curts i més petita pels més llargs. Gràfics semblants s'obtenen amb les altres paraules eina.

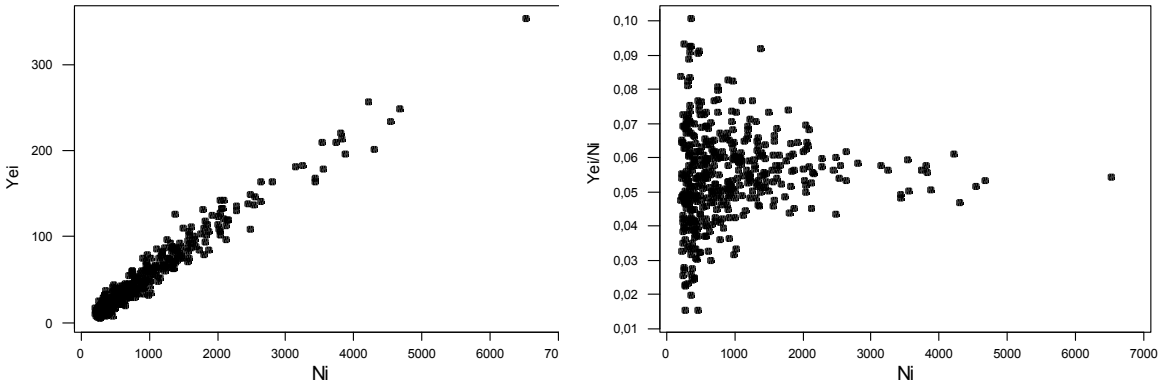

Figura 12.1: Relació del nombre i de la proporció d'ocurrències de la paraula *e*, y_{ei} i y_{ei}/N_i respectivament, amb la llargada del capítol N_i.

12.2.3 Gràfics de Control per les proporcions de paraules

Representem gràficament l'evolució de les proporcions d'ús de les paraules eina, y_{ji}/N_i, al llarg del *Tirant* per observar si les π_{ji} són constants i, per tant, de si l'estil és homogeni o be hi ha algun punt en el que canvi. Si no hi ha canvi d'estil, el valor esperat de y_{ji}/N_i serà constant al llarg del llibre.

A la figura 12.2 hi ha representades, mitjançant gràfics *P*, l'evolució de les proporcions d'ocurrències de cadascuna de les 25 paraules eina més freqüents al llarg del *Tirant*. Els

capítols 1-382 estan en color verd, mentre que els que van del 346 al final estan en color vermell. Els límits de control han sigut obtinguts a partir de les proporcions en tots els capítols, tal i com s'ha explicat en el capítol 5.

De l'anàlisi de cadascun dels gràfics per separat tenim que:

- *e*: en el capítol 382, aproximadament, s'observa un augment en el nivell,
- *de*: en el capítol 382, aproximadament, s'observa un augment en el nivell,
- *la*: en el capítol 355, aproximadament, s'observa un augment en el nivell,
- *que*: en el capítol 382, aproximadament, s'observa una disminució en el nivell,
- *lo*: en el capítol 100, aproximadament, s'observa una disminució en el nivell,
- *en*: no s'observa cap canvi significatiu en el nivell,
- *a*: no s'observa cap canvi significatiu en el nivell,
- *per*: no s'observa cap canvi significatiu en el nivell,
- *no*: s'hi observen dos possibles punts de canvi, el primer cap al capítol 110 i el segon cap el 383. Abans del primer punt de canvi el nivell cau lleugerament per sota de la mitjana global, entre els dos punts de canvi el nivell està clarament per sobre de la mitjana global, mentre que després del segon canvi el nivell està per sota de la mitjana global i per sota del nivell d'abans del primer punt de canvi,
- *l*: en el capítol 383 la variabilitat augmenta,
- *los*: no s'observa cap canvi significatiu en el nivell,
- *com*: en el capítol 450, aproximadament, s'observa una disminució en el nivell,
- *ab*: no s'observa cap canvi significatiu en el nivell,
- *les*: en el capítol 200, aproximadament, s'observa una lleugera disminució en el nivell,
- *d:* no s'observa cap canvi significatiu en el nivell,
- *li:* no s'observa cap canvi significatiu en el nivell,
- *qui:* en el capítol 382, aproximadament, s'observa una disminució en el nivell,
- *del:* en el capítol 400, aproximadament, s'observa un augment en el nivell,
- *se:* no s'observa cap canvi significatiu en el nivell,
- *molt:* en el capítol 383, aproximadament, s'observa un augment en el nivell,
- *és*: en el capítol 383, aproximadament, s'observa una disminució en el nivell,
- *gran:* en el capítol 383, aproximadament, s'observa un augment en el nivell,
- *jo:* en el capítol 383, aproximadament, s'observa una disminució en el nivell,
- *si:* en el capítol 383, aproximadament, s'observa una disminució en el nivell,
- *dix:* en el capítol 350, aproximadament, s'observa una disminució en el nivell.

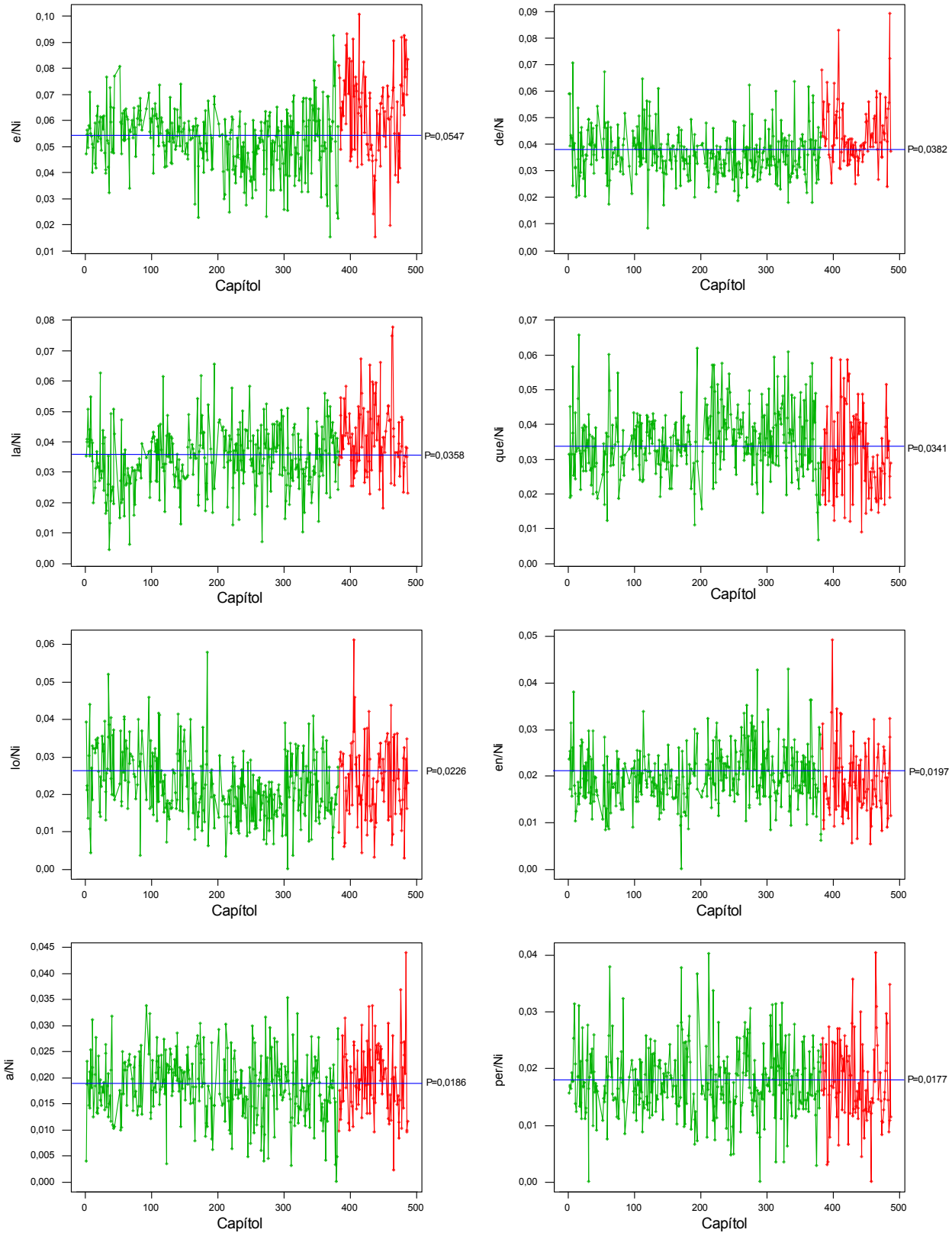

Figura 12.2a: Evolució temporal del nombre d'ocurrències de les paraules eina: *e, de, la, que, lo, en, a, per* en cada capítol. El color verd correspon als capítols 1-382, el vermell als capítols 383-487.

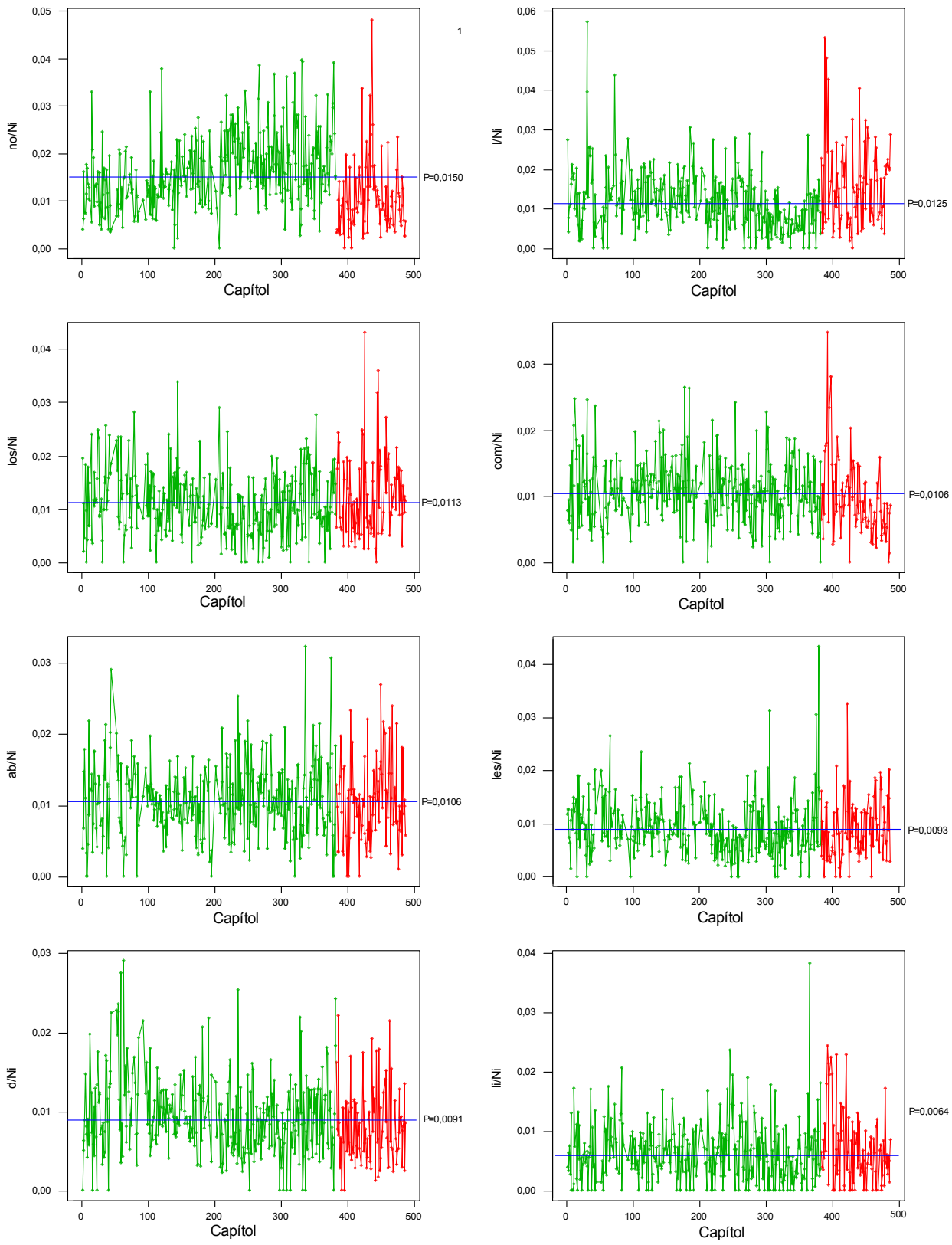

Figura 12.2b: Evolució temporal del nombre d'ocurrències de les paraules eina *no, l, los, com, ab, les, d, li* en cada capítol. El color verd correspon als capítols 1-382, el vermell als capítols 383-487.

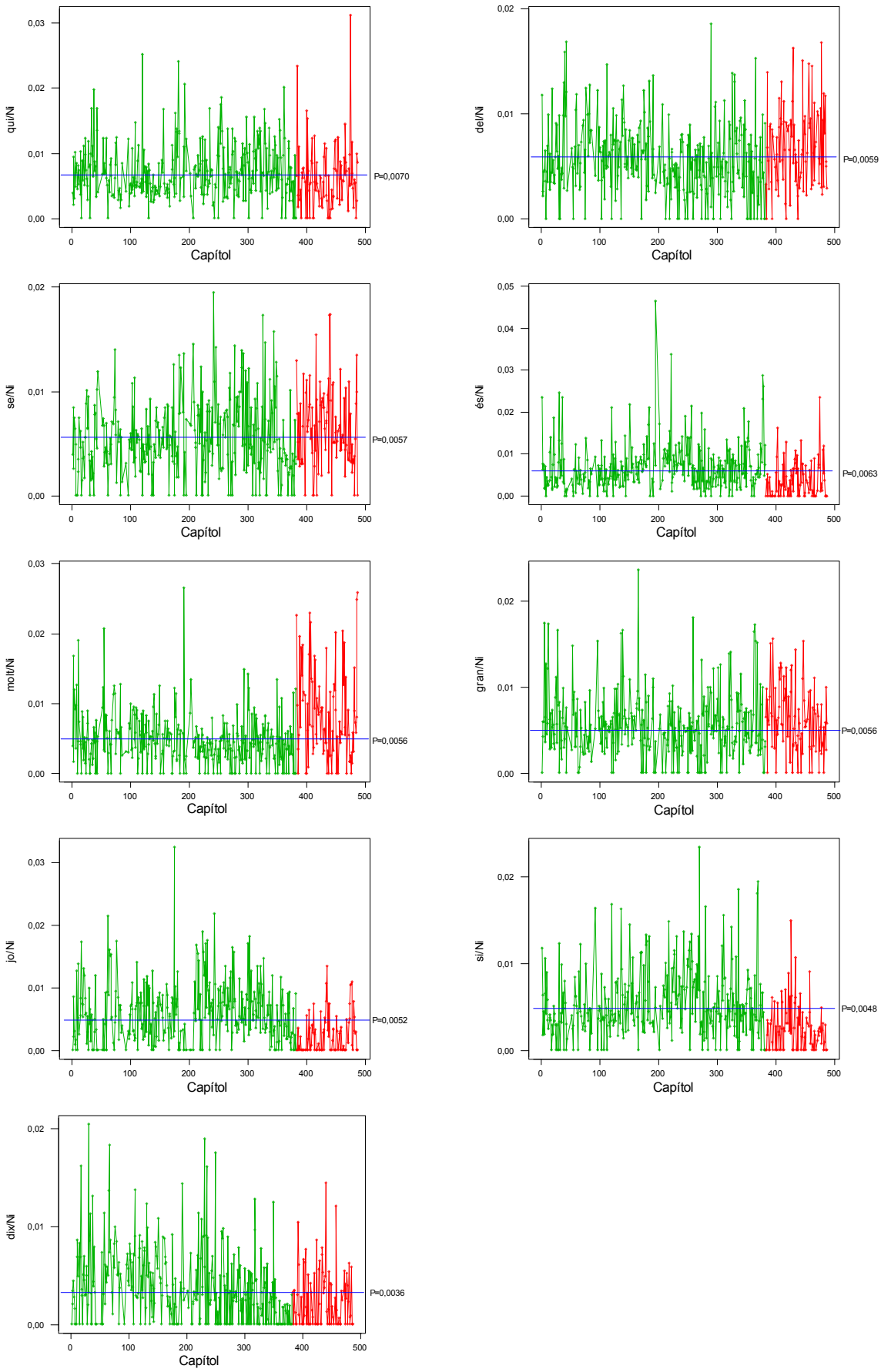

Figura 12.2c: Evolució temporal del nombre d'ocurrències de les paraules eina: *qui, del, se, és, molt, gran, jo, si, dix* en cada capítol. El canvi de color està entre els capítols 382 i 383.

La figura 12.3 mostra l'evolució de la proporció de paraules diferents a les *12* i *25* paraules eina considerades, $Resta_i$. En totes dues seqüències es pot apreciar una forta disminució en el nivell i un augment en la variabilitat en el capítol 333, que correspon, aproximadament, al capítol 383. Aquest augment es pot interpretar com que a partir d'aquest punt s'incrementa l'ús de les paraules que són més abundants.

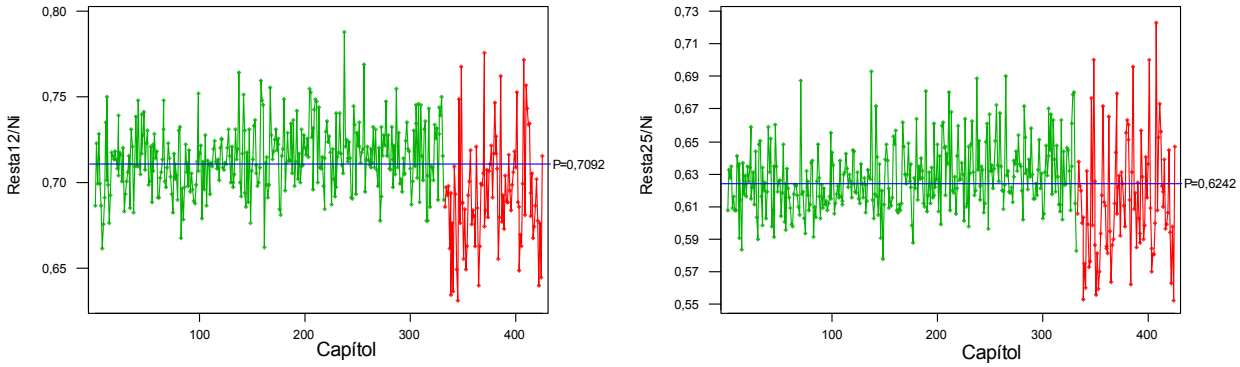

Figura 12.3: Evolució temporal de la proporció de paraules diferents a les 12 (esquerra) i 25 (dreta) *paraules eina* considerades. S'han representat en verd els primers 382 capítols, en vermell del 383 al final.

La figura 12.4 mostra l'evolució temporal de les proporcions de les paraules *Tirant* i *rei*, les dues paraules que, tot i estar entre les més freqüents, no s'han considerat en l'estudi per dependre del context. Observem com la paraula *Tirant* no surt fins al capítol 29, i com hi ha parts del llibre on apareix freqüentment la paraula *rei*, mentre n'hi ha d'altres en les que apareix de forma molt més escassa, en funció de l'argument.

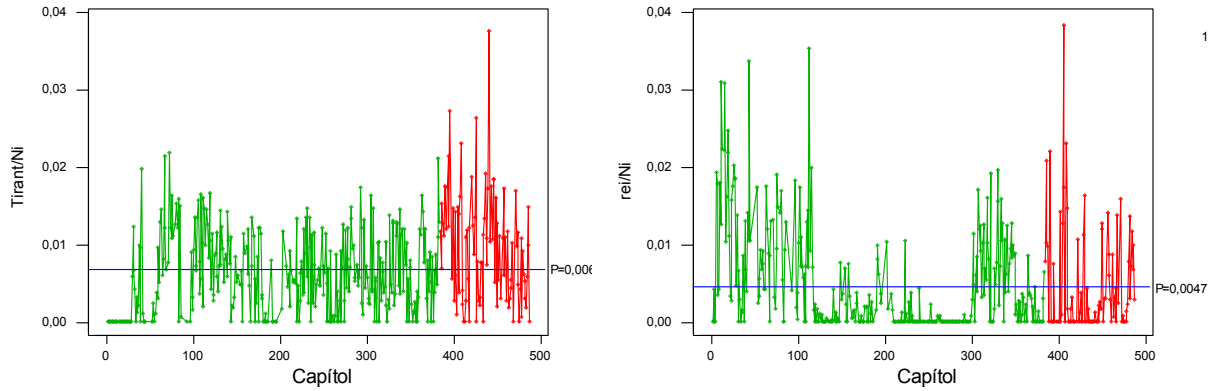

Figura 12.4: Evolució temporal de les paraules descartades de l'estudi entre les més freqüents (*Tirant* i *rei*), en considerar que el seu ús depèn del context. S'han representat en verd els primers 382 capítols, en vermell del 383 al final.

La seqüència del vector multinomial amb el nombre d'ocurrències per la paraula *j-èssima*, per $j=1,2,\ldots,l+1$, $\boldsymbol{y_i}=(y_{1i},y_{2i},y_{3i},\ldots,y_{l+1i})$, es pot representar en un gràfic de control Chi-quadrat, com el de la figura 12.5. Els valors de la distància χ^2 per cada capítol,

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

s'han calculat prenent, com a valor esperat de les proporcions, $\hat{\pi}_j$, les proporcions promig de paraules de *j* lletres en tots els capítols:

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

amb *l=12* o *l=25*, per *i=1,2,...n*. En els gràfics es pot observar com el punt de canvi sembla coincidir amb el canvi de color, en el capítol *382*. S'observa com després del punt de canvi no només varia el nivell si no que, a més a més, augmenta la variabilitat.

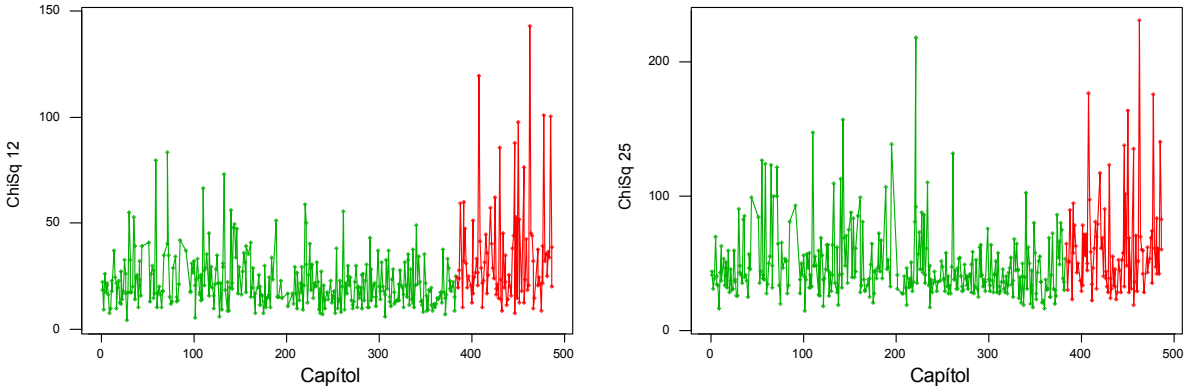

Figura 12.5: Gràfic Chi-quadrat per a les seqüències multinomials composades per 12 (esquerra) i 25 (dreta) paraules eina. Els punts en verd corresponen als capítols des de l'inici fins al 382 i els de color vermell són els capítols del 383 al final.

12.2.4 Anàlisi de correspondències per l'ús de paraules

L'anàlisi de correspondències és una tècnica que permet representar gràficament gràfic les dades contingudes en una taula de contingència. El punt de partida per a l'estudi de l'ús de paraules a través de l'anàlisi de correspondències està format per la taula de contingència 12.1. Els detalls sobre l'anàlisi de correspondències s'han descrit en el capítol 5.

En els gràfics de la dreta de la figura 12.6 es troben els gràfics simètrics per les files. Per *l=12* es pot veure com la majoria de punts que corresponen a capítols posteriors al 382, en vermell en la figura, prenen valors en la segona component negatius. Pel que fa a la primera component no sembla discriminar entre els capítols anteriors i els posteriors al 382, tot i que la majoria dels capítols posteriors al 382 prenen valors positius. Quan s'analitzen 25 paraules eina, el gràfic mostra com la majoria de punts en vermell tenen valors de la primera component positius, mentre que la segona component no sembla discriminar entre els capítols anteriors i els posteriors al 382.

El gràfic simètric per columnes (a l'esquerra en el gràfic) per *l=12* mostra com les paraules que estan en el primer quadrant (*e* i *de*) són per les que observàvem un canvi més acusat al voltant del capítol 382, augmentant la proporció d'aquestes paraules després del canvi. De les paraules que es troben en el tercer quadrant: *no*, que és la que es troba més allunyada de l'origen, té el canvi molt marcat en el capítol 382 però en sentit contrari: la proporció de *no* disminueix després del canvi; *que* té el canvi menys marcat i també disminueix després del capítol 382. Les altres dues paraules que es troben en el tercer quadrant, *per* i *en*, es troben molt més a prop de l'origen i en la seqüència temporal no s'observa cap canvi en el nivell. Pel que fa a les paraules que es troben en el segon quadrant, *l*, la més allunyada de l'origen, té un canvi de nivell proper

al capítol 382, augmentant la proporció després del punt de canvi, *la* té un comportament semblant, mentre que per *a*, molt més propera a l'origen, no s'observa un canvi en el nivell. De les paraules que es troben en el quart quadrant, *los, lo* i *com*, en cap d'elles s'observa un canvi de nivell al voltant del capítol 382. Per *l=25* el mostra com les paraules amb valor absolut de la primera component més elevat són algunes de les que millor marquen un canvi al voltant del capítol 382: en *molt*, amb valor positiu, augmenta la proporció després del canvi, mentre que en *jo, si, és, no*, amb valor negatiu, disminueixen després del canvi. Cal observar que aquestes paraules no són de les més freqüents i que algunes paraules molt més freqüents i que presenten un punt de canvi molt clar al voltant del capítol 382, *e, de, la*, es troben en el gràfic molt més a prop de l'origen.

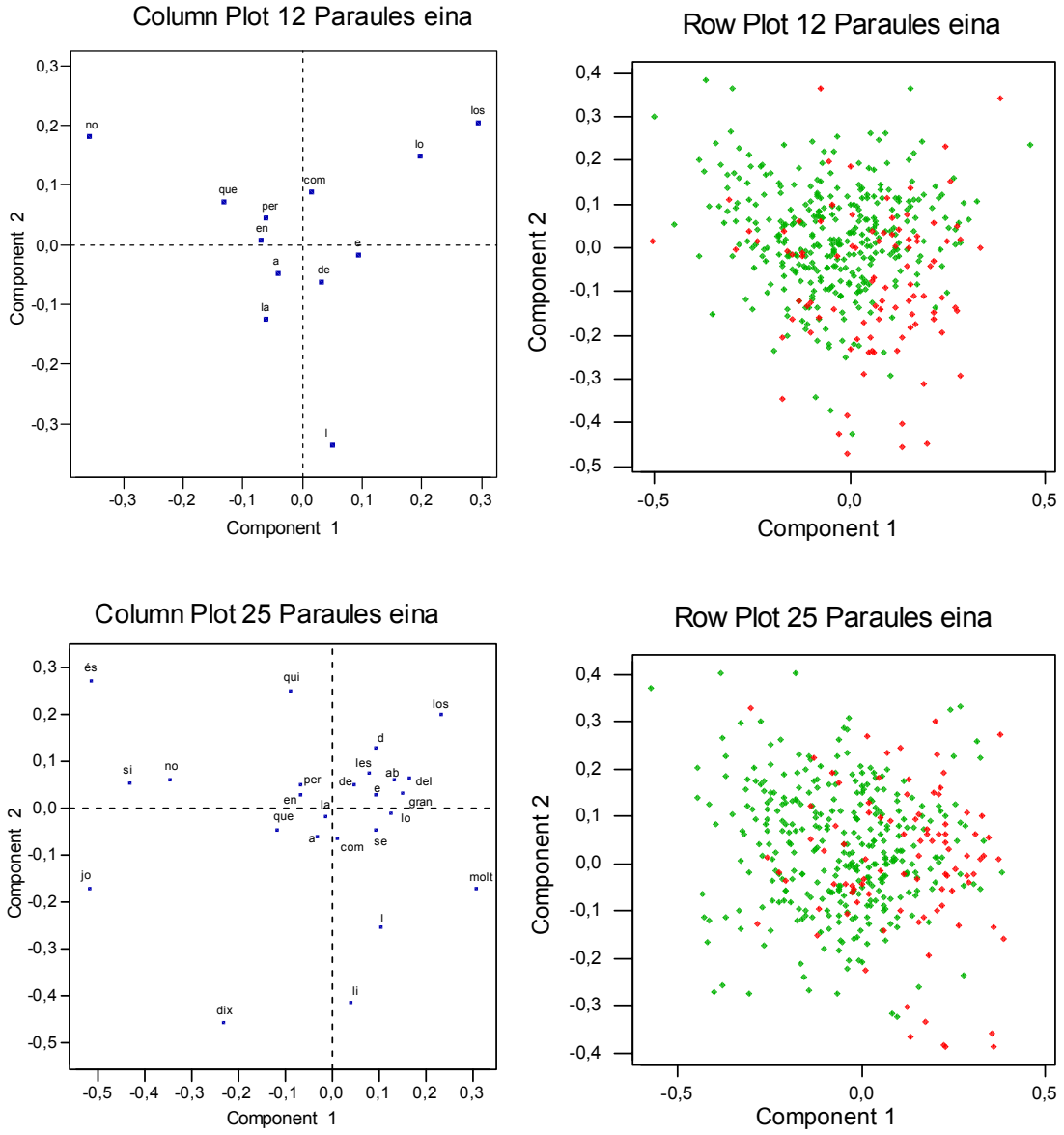

Figura 12.6: Gràfic simètric per columnes a l'esquerra i files a la dreta de l'anàlisi de correspondències per a 12 (a d'alt) i per a 25 (a baix) paraules eina. S'han representat en color verd els primers 382 capítols, i en vermell els capítols 383-487.

L'evolució temporal de les dues primeres components per files, representades en els gràfics de la figura 12.7, mostren un comportament força diferent si es consideren *12* o

25 paraules eina. Per *12* paraules eina s'observa com el canvi entre els capítols anteriors i els posteriors al *382*, marcat amb un canvi de color, és molt més evident per a la segona component, en la que els capítols posteriors semblen tenir valor esperat més petits. Per la primera component el canvi també és visible, gairebé tots els capítols posteriors al *382* prenen valors positius per a la primera component, mentre que els capítols anteriors prenen valors tant positius com negatius. Quan s'analitzen 25 paraules eina el canvi en el capítol 382 és molt més evident per la primera component, per la que s'observa un augment en el nivell, prenent la majoria dels capítols posteriors al 382 valors positius. La segona component no sembla bona per a detectar el possible canvi d'estil.

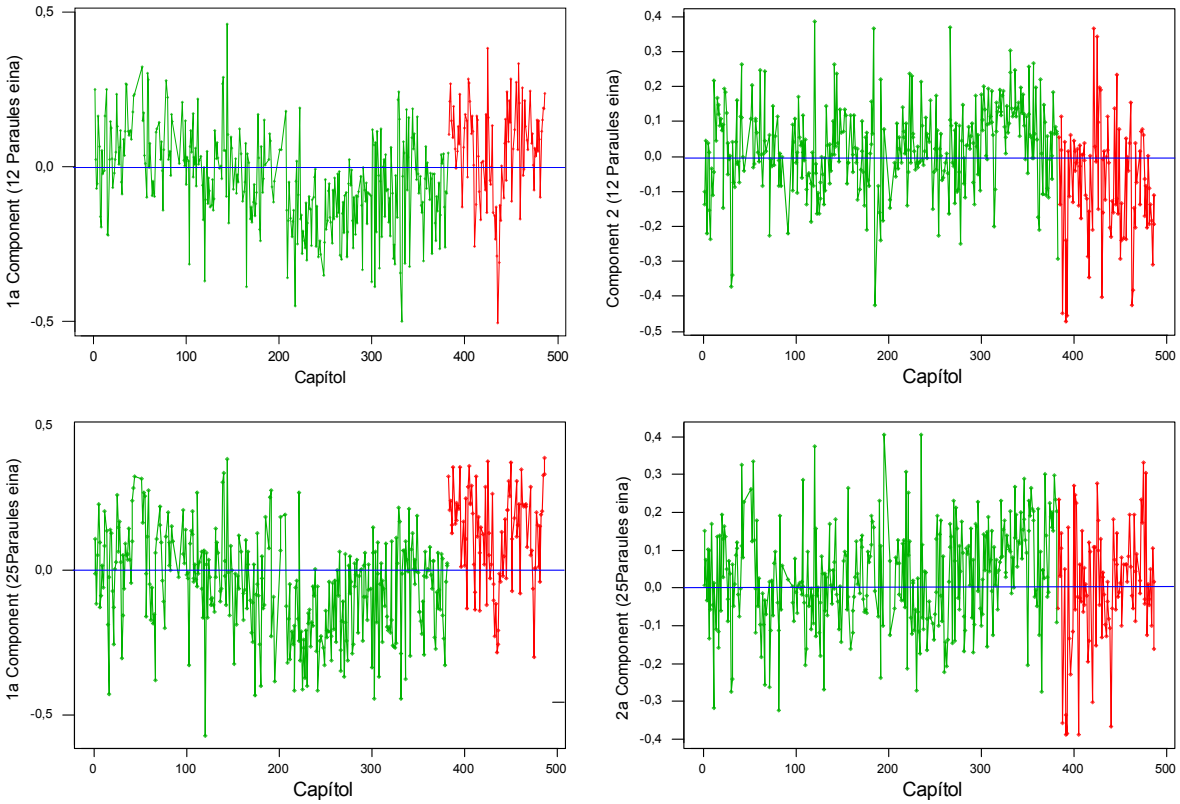

Figura 12.7: Evolució temporal de les dues primeres components de l'anàlisi de correspondències: els gràfics de l'esquerra són per la primera component i els de l'esquerra per la segona, quan s'analitzen 12 (a d'alt) i 25 (a baix) *paraules eina*. El color verd correspon als capítols 1-382 i el color vermell als capítols posteriors.

12.2.5 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 dels valor de les dues primeres components de l'anàlisi de correspondències, en ambdós casos tant per 12 com per 25 *paraules eina*. A continuació, mitjançant l'aproximació al problema basada en models logístics, s'estudien les 25 seqüències de paraules eina per separat i les seqüències de proporcions de paraules diferents de les 12 o 25 paraules eina considerades, és a dir, la proporció que suposa la variable *Resta*, i s'obtenen vint-i-cinc més dues estimacions del

punt de canvi. A continuació, s'estima el punt de canvi en la seqüència de dades multinomials contingudes en la taula 12.1 via l'ajust de models politòmics.

12.2.5.1 Punt de canvi en seqüències Normals

S'ha estimat el punt de canvi per la seqüència de valors de les dues primeres components de l'anàlisi de correspondències fent servir les tècniques proposades per a dades Normals descrites a la secció 6.2. Si anomenem pc_{1i} i pc_{2i} als valors de les dues primeres components en el capítol i-èssim per l'anàlisi fet amb dotze paraules, sota les hipòtesi de independència i Normalitat, es distribueixen:

$$pc_{1i} \sim N(\mu_{1i} = \beta_{10}^{(r)} + \beta_{11}^{(r)} Ind_{1i}^{(r)}, \sigma_{1i}^2 = \frac{\sigma_1^2}{N_i}),$$

i

$$pc_{2i} \sim N(\mu_{2i} = \beta_{20}^{(r)} + \beta_{21}^{(r)} Ind_{1i}^{(r)}, \sigma_{2i}^2 = \frac{\sigma_2^2}{N_i}),$$

per *r=1,2,...,n-1*, 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*. Per vint-i-cinc paraules els models són anàlegs, 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 $w_i=N_i$.

Quan s'analitza la seqüència de *12* paraules eina, la primera component, que explica prop del 22% de la inèrcia total, té el màxim de F_r a *r=110*, tot i que també es troben màxims locals per a la *r=382 i r=445*, tal i com es pot observar en el gràfic de l'esquerra de la figura 12.8. La segona component, que explica prop del 16% de la inèrcia, senyala el punt de canvi a *r=387*, el que confirma el que s'havia vist de l'anàlisi del gràfic de control. El valor màxim de F_r més gran per a la segona component que per la primera, el que indica que el canvi és més accentuat.

Quan s'analitza la seqüència de 25 paraules eina, el màxim és molt més marcat per a la primera component, marcant el punt de canvi en el capítol *382*. La primera component explica el 18% de la inèrcia total, la segona el 10%.

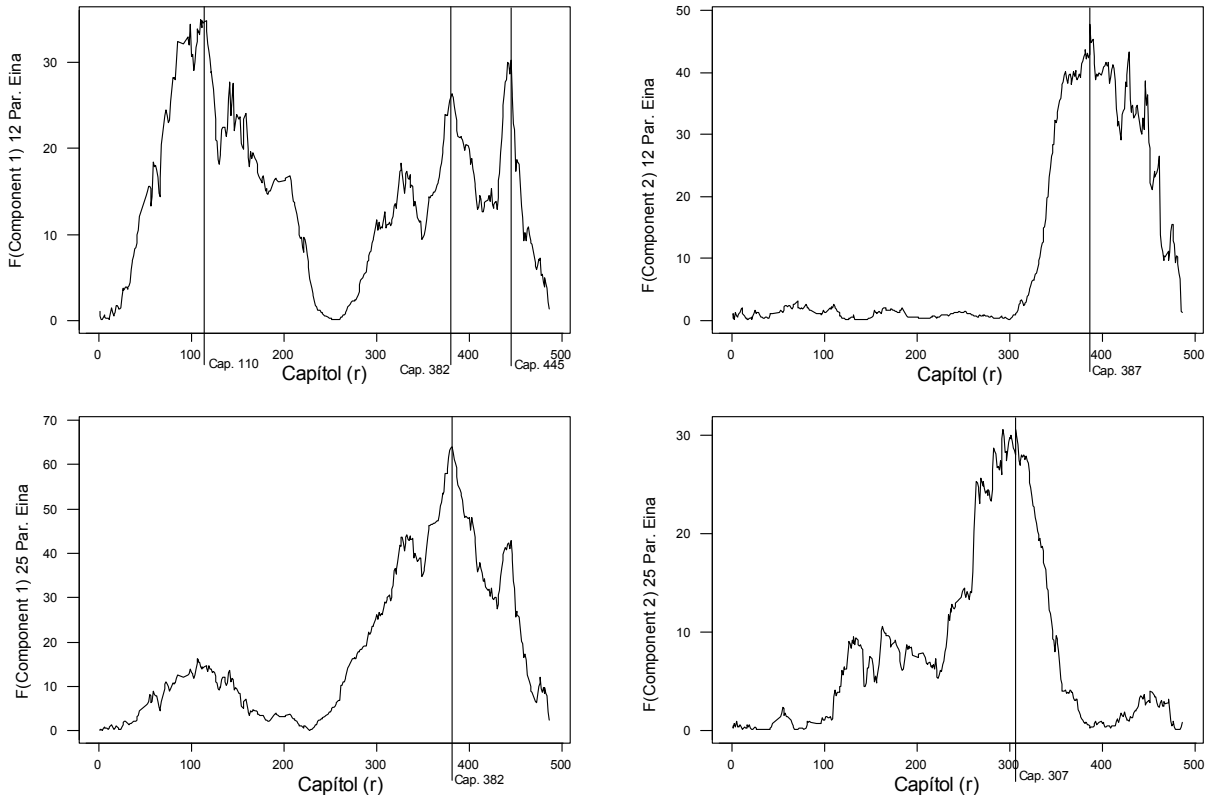

Figura 12.8: Evolució de F_r en funció de r per a la seqüència de valors de les dues primeres components de l'anàlisi de correspondències. Els gràfics de dalt corresponen a l'estudi de 12 paraules eina, els de baix a 25.

12.2.5.2 Punt de canvi en seqüències binomials

En aquest apartat s'estudien per separat les seqüències univariants de les *l=12* i *l=25 paraules eina* separat i les seqüències de proporcions de paraules diferents de les 12 o 25 paraules eina considerades. Aquest procediment dóna *l+2* estimacions del punt de canvi, possiblement diferents.

Suposant que existeix un punt de canvi a r, per una categoria j fixada $j=1,2,...l$, y_{ji} té una distribució $binomial(N_i,\pi_{ja})$ per $i=1,...,r$ i $binomial(N_i,\pi_{jd})$ per $i=r+1,...,n$. Per cadascuna de les l categories s'ajusta el model logístic 6.12:

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

La figura 12.9 conté els gràfics de el màxim del logaritme de la versemblança, $L_j\left(r, \hat{\beta}^{(r)}\right)$, en funció de r per a les 25 paraules.

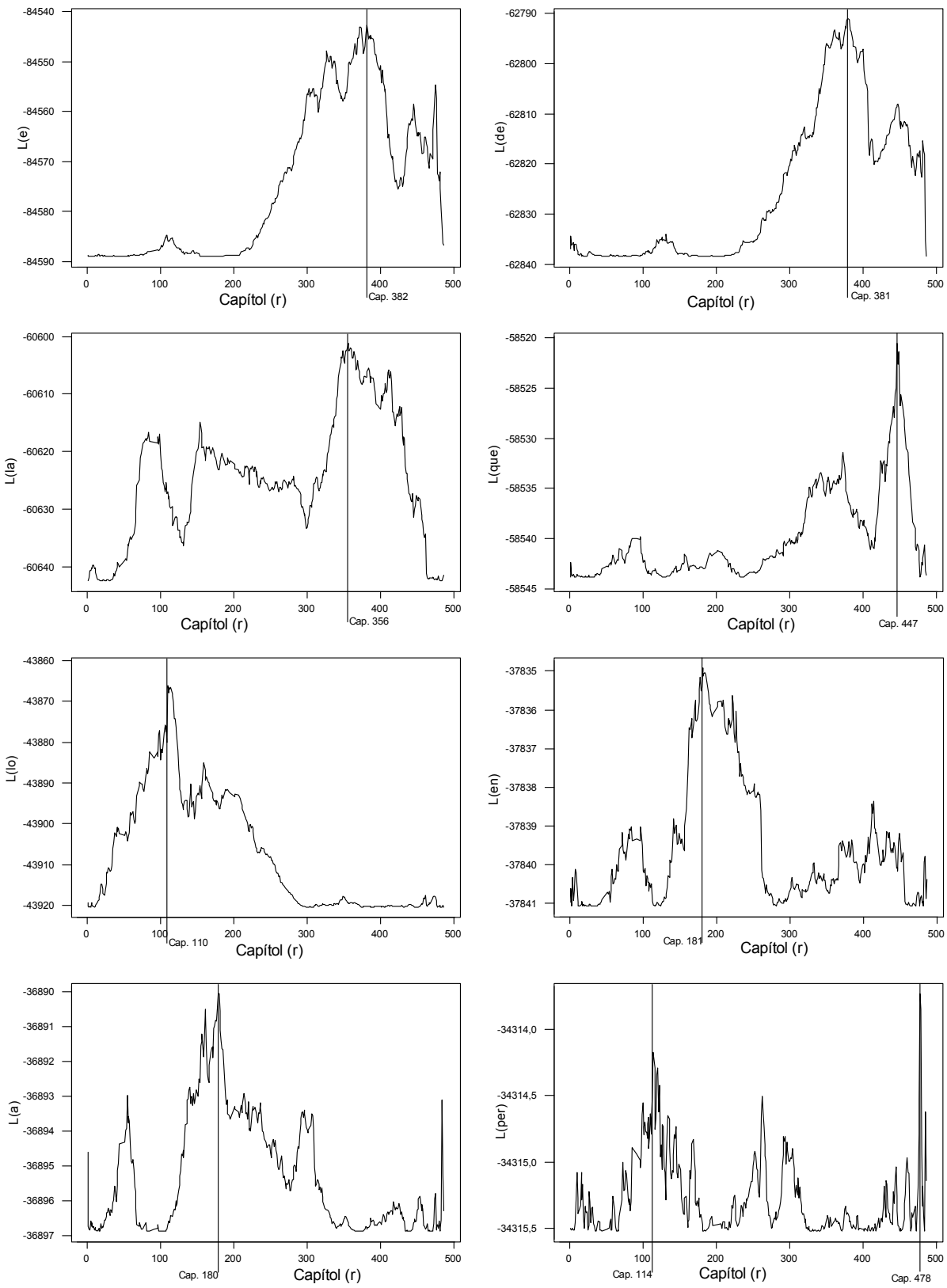

Figura 12.9a: Relació de $L_j\left(r, \hat{\beta}^{(r)}\right)$ en funció de *r* per als models amb resposta la proporció d'aparicions de les paraules eina: *e, de, la, que, lo, en, a, per* per capítols.

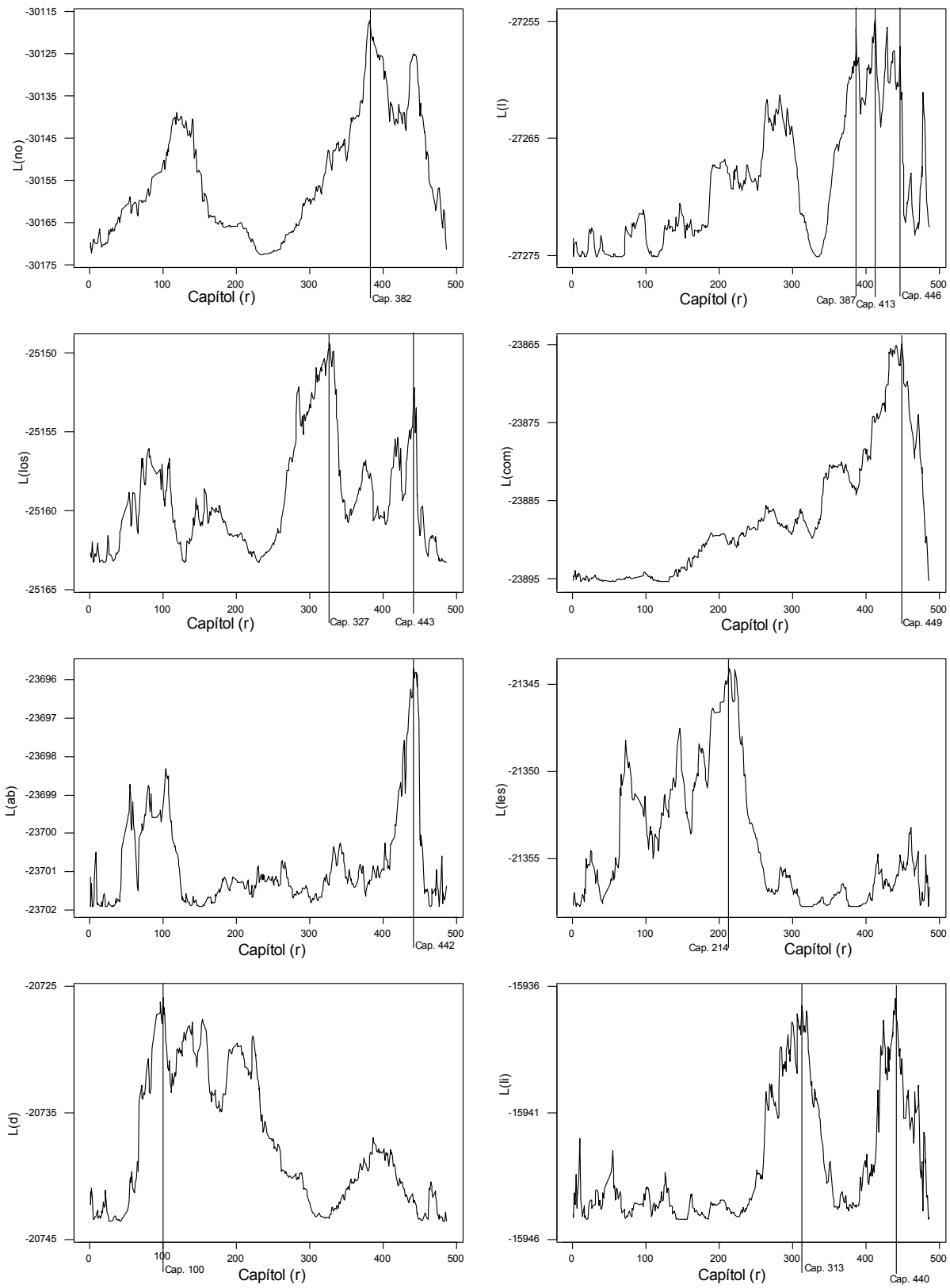

Figura 12.9b: Relació de $L_j\left(r, \hat{\beta}^{(r)}\right)$ en funció de *r* per als models amb resposta la proporció d'aparicions de les paraules eina: *no, l, los, com, ab, les, d, li* per capítols.

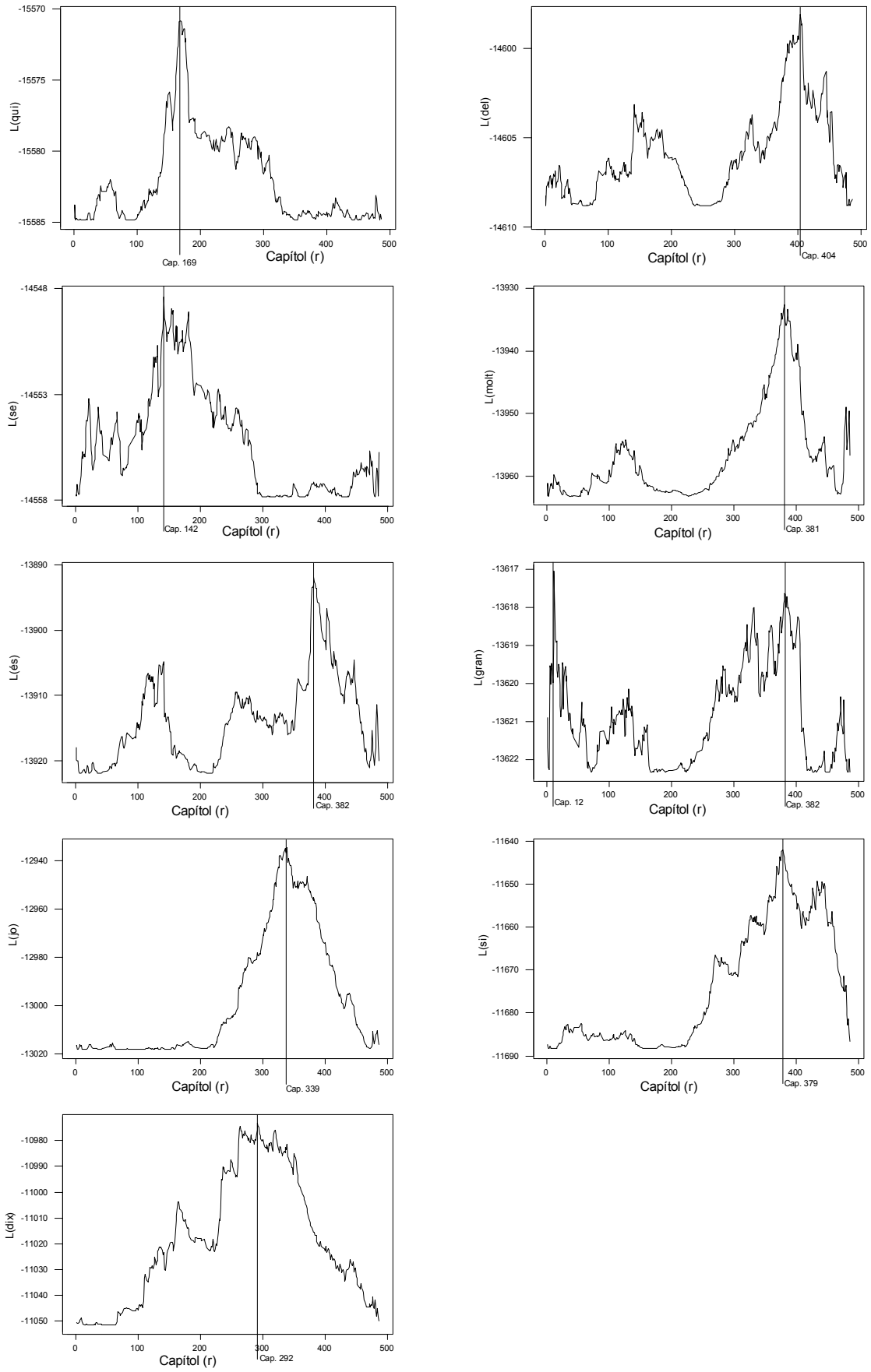

Figura 12.9c: Relació de $L_j\left(r, \hat{\beta}^{(r)}\right)$ en funció de *r* per als models amb resposta la proporció d'aparicions de les paraules eina: *qui, del, se, molt, és, gran, jo, si, dix* per capítols.

Observem com el màxim del logaritme de la versemblança és més gran a mesura que $\hat{\pi}_j$ són més petites. Això vol dir que si volem resumir les *l* estimacions del punt de canvi en un sol valor, adaptant els estadístics proposats per Wolfe i Chen (1990) com hem descrit al capítol 6, donarem més pes en l'estimació a les paraules amb menor probabilitat d'aparició.

S'ha estimat el punt de canvi per la seqüència de proporcions de paraules diferents de les 12 o 25 paraules eina considerades, fent servir les tècniques proposades per a seqüències de binomials.

Els gràfics de la figura *12.10* mostren l'evolució de $L_j\left(r,\hat{\beta}^{(r)}\right)$ en funció de *r* quan estudiem *12* i *25* paraules eina. El màxim de $L_j\left(r,\hat{\beta}^{(r)}\right)$ i, per tant, la millor estimació del punt de canvi per a les seqüències de $Resta_i$ pels capítols per *l=12* és $\hat{r}_L=382$ i per *l= 25* a $\hat{r}_L=381$. El gràfic mostra com els màxims en $\hat{r}_L$ estan clarament diferenciats.

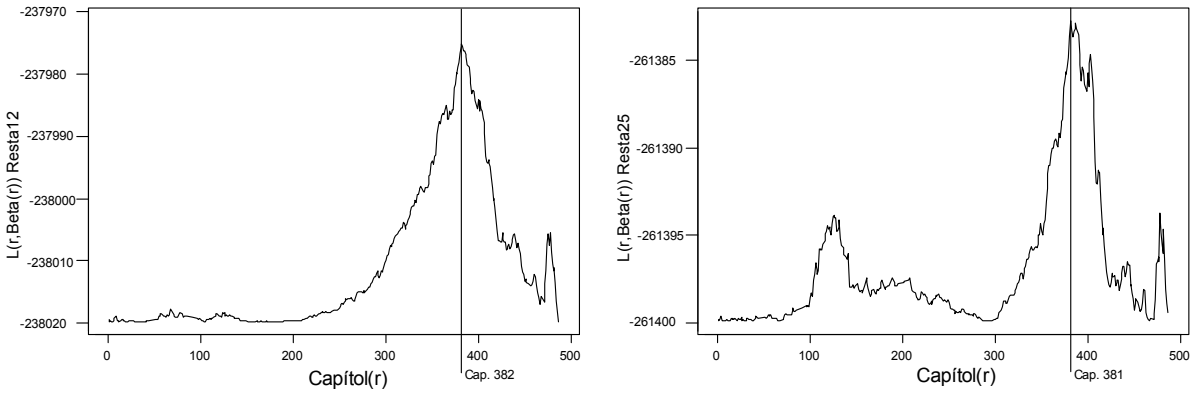

Figura 12.10: Gràfic que representa l'evolució de $L_j\left(r,\hat{\beta}^{(r)}\right)$ en funció de *r* per a l'estimació del punt de canvi en les seqüències de les proporcions de paraules diferents a les *paraules eina* considerades.

12.2.5.3 Punt de canvi en seqüències multinomials; model politòmic

El problema d'estimar el punt de canvi en la seqüència multinomial formada per les *l* paraules, s'ha abordat mitjançant l'ajust de models. Si y_{ji} és el nombre d'ocurrències de la paraula *j* en el capítol *i*, per *j=1,2,...,l,* $\mathbf{y_i}=(y_{1i},y_{2i},....,y_{li})$ està distribuït $Mult(N_i,\pi_i)$, on N_i és la llargada del capítol i $\pi_i=(\pi_{1i},\pi_{2i},...,\pi_{li})$ és el vector amb les probabilitats que una paraula 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, per $l=12,\ 25$. El gràfic de la figura 12.11 mostra l'evolució de $L\left(\hat{\beta}^{(r)}\right)$ en funció de r. L'estimació del punt de canvi és, tant per $l=12$ com per $l=25$, $\hat{r}_M=382$.

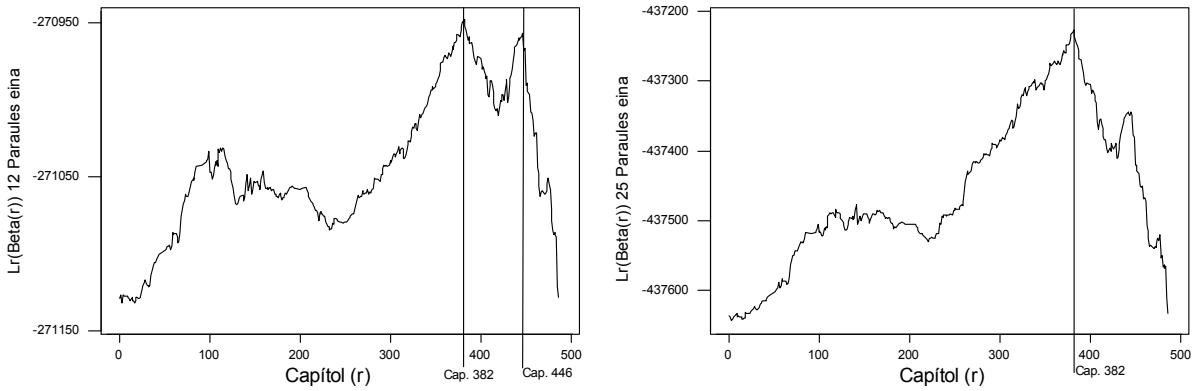

Figura 12.11: Gràfic de l'evolució de $L_r\left(\hat{\beta}^{(r)}\right)$ en funció de r per a les seqüències multinomials de 12 (esquerra) i 25 (dreta) *paraules eina*, per capítols.

Estimar r en una seqüència multinomial via l'ajust de models de regressió politòmica és equivalent a ajustar-lo via màxim versemblant.

12.2.5.4 Paraules discriminants

Podem intuir, a partir dels gràfics de la figura 12.9, quines paraules són bones per discriminar entre els dos estils separats pel punt de canvi. La comparació d'aquests gràfics és complicada, molt laboriosa i poc eficient. Per a determinar les paraules marcadores d'estil, fem servir l'aproximació de la binomial a la Normal per ajustar, per cadascuna de les *paraules eina*, el model de regressió lineal amb pesos:

$$\hat{\pi}_{ji}=\frac{y_{ji}}{N_i}\sim N(\beta_{j0}+\beta_{j1}Ind_{1i}^{(r=382)},\sigma_{ji}^{2}=\frac{\sigma_j^{2}}{N_i}),$$

on y_{ji} és el nombre d'ocurrències de la paraula j-èssima en el capítol i-èssim, $Ind_{1i}^{(r=382)}$ és una variable indicadora que pren valor *0* pels primers 382 capítols, i valor *1* pels capítols del 383 al final. Ponderem les observacions amb un pes proporcional a N_i. Aquest procediment és equivalent a fer un test de comparació de les mitjanes (amb pesos) dels dos conjunts de capítols: 1-382 i 383-final.

Per cada paraula examinada s'observa el valor de l'estadístic:

$$t_r=\frac{b_1}{s_{b_1}},$$

i es consideraran bones discriminants de l'estil totes aquelles paraules per les que el valor absolut de l'estadístic t_r sigui superior a 3. L'annex A12.2 mostra, per cadascuna

de les 100 paraules més freqüents, el valor absolut de l'estadístic *t*. Podem concloure que entre les 25 més freqüents, n'hi ha 12 que poden ser considerades com a bones discriminants: *e, de, la, no, l, com, del, molt, és, jo, si, dix*. Observem com *e* i *de* són les dues paraules més freqüents i les més discriminants. Les paraules subratllades són aquelles que la seva proporció disminueix després del capítol 382. A banda d'aquestes 12 paraules, entre les 100 més nombroses en trobem altres 16 amb $|t_r|>3$: *era, un, féu, hi, sua, aquell, bé, ho, ni, molta, foren, tal, o, sinó, qual, dir*.

No s'han considerat com a possibles paraules discriminants aquelles que són sensibles al context: *Tirant, rei, senyor, emperador, cavaller, ciutat, princesa, amor, armes, gent, terra, gent, terra, senyora, honor* i *mort,* és a dir, tots els substantius que apareixen en la llista de les 100 paraules més abundants tret de *paraules*, un mot que es fa servir molt sovint per concloure un capítol amb expressions del tipus: "*fer-li resposta en estil de semblants paraules*".

12.2.5.5 Conclusions

La conclusió d'aquesta anàlisi és que es constata l'existència d'un punt en el que l'estil, quantificat a través de l'ús de paraules eina per capítols, canvia. L'estimador del punt de canvi de la seqüència multinomial és r=382.

La taula 12.2 mostra un quadre resum amb les estimacions del punt de canvi obtingudes en l'estudi de l'ús de les paraules més freqüents pels capítols de més de 200 paraules.

Distribució	**Seqüència**	**Pt. canvi**		**Seqüència**	**Pt. canvi**
Normal	*1ª Comp. A. corr. l=12*	$\hat{r}_N$ =110		*1ª Comp. A. corr. l=25*	$\hat{r}_N$ =382
	2º Comp. A. corr. l=12	$\hat{r}_N$ =387		*2ª Comp. A. corr. l=25*	$\hat{r}_N$ =307
Binomial	*e*	$\hat{r}_L$ =382		*les*	$\hat{r}_L$ =214
	de	$\hat{r}_L$ =381		*d*	$\hat{r}_L$ =100
	la	$\hat{r}_L$ =356		*li*	$\hat{r}_L$ =440
	que	$\hat{r}_L$ =447		*qui*	$\hat{r}_L$ =169
	lo	$\hat{r}_L$ =110		*del*	$\hat{r}_L$ =404
	en	$\hat{r}_L$ =181		*se*	$\hat{r}_L$ =142
	a	$\hat{r}_L$ =180		*molt*	$\hat{r}_L$ =381
	per	$\hat{r}_L$ =478		*és*	$\hat{r}_L$ =382
	no	$\hat{r}_L$ =382		*gran*	$\hat{r}_L$ =382
	l	$\hat{r}_L$ =413		*jo*	$\hat{r}_L$ =334
	los	$\hat{r}_L$ =327		*si*	$\hat{r}_L$ =379
	com	$\hat{r}_L$ =449		*dix*	$\hat{r}_L$ =292
	ab	$\hat{r}_L$ =442			
	Resta l=12	$\hat{r}_L$ =382		*Resta l=25*	$\hat{r}_L$ =381
Multinomial	*Model Politòmic l=12*	$\hat{r}_M$ =382		*Model Politòmic l=25*	$\hat{r}_M$ =382

Taula 12.2: Quadre resum de les estimacions del punt de canvi obtingudes en l'anàlisi de la l'ús de les paraules més freqüents per als capítols de més de 200 paraules.

Donant per bo aquest punt de canvi, s'han identificat aquelles paraules, entre les 100 més abundants en el *Tirant*, que al voltant del capítol 382 presenten un canvi en la proporció d'ocurrències, i que, per tant, poden ser considerades, en el *Tirant*, bones discriminants de l'estil. Aquestes paraules són: *e, de, la, no, l, com*, *del, molt, és, jo, si, dix*, *era, un, féu, hi, sua, aquell, bé, ho, ni, molta, foren, tal, o, sinó, qual, dir*.

12.2.6 Anàlisi cluster de les files de la taula de contingència

De l'anàlisi gràfica de les seqüències de paraules eina s'ha pogut apreciar com alguns capítols posteriors al punt de canvi prenen valors que els fan més semblants als anteriors a $\hat{r}$ que als posteriors, per tant hi ha la possibilitat que en la part final del llibre hi hagi barreja d'estils. Aquesta situació ja s'ha detectat en l'anàlisi de la llargada de paraula.

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 per a l'aplicació d'aquestes tècniques són les dues taules de contingència, per *l=12* i *l=25 paraules eina*, on les files són els 425 capítols i les columnes representen el nombre d'ocurrències per capítol de cadascuna de les *l* paraules considerades.

12.2.6.1 Anàlisi cluster basada en la distància χ^2

Tant per *l=12* com per *l=25* s'ha executat més de 1000 vegades l'algorisme cluster basada en la distància χ^2, amb assignacions inicials aleatòries de capítols a un dels dos grups. El criteri escollit per a decidir l'agrupació en clusters òptima és el de maximitzar χ^2_B, és a dir, la distància *entre clusters*.

Quan es busquen dos clusters en funció de les *l=12* paraules eina més abundants, l'algorisme convergeix a dues solucions, una d'elles millor en ser χ^2_B màxima i χ^2_W mínima. L'algorisme convergeix més del 77% de les vegades a la solució òptima. Per aquesta, els dos clusters contenen, respectivament, 239 i 186 capítols. L'estadístic χ^2 per la taula de contingència completa val $\chi^2=10305.1$, mentre que el valor entre clusters obtingut és de $\chi^2_B=1382.6$, el que dóna una relació $\chi^2_B/\chi^2 = 0{,}134$, és a dir l'agrupació en aquests dos clusters explica el 13,3% de la "no-homogeneïtat" entre tots els capítols en funció de l'ús de les dotze paraules eina més freqüents. Aquest valor, tal i com s'ha avançat en el capítol 7, és menor que la proporció de inèrcia explicada per la primera component de l'anàlisi de correspondències, que és del 21%.

Quan s'analitza la taula amb 25 paraules eina, l'algorisme convergeix també a dues úniques solucions. L'agrupació per la que el valor de χ^2_B és màxim agrupa els blocs en clusters de 216 i 209 capítols. L'algoritme ha convergit més del 70% de les vegades a aquesta solució. L'altra solució, a la que ha convergit prop del 30% de les iteracions que, tot i tenir un valor de χ^2_B menor, té χ^2_W, és a dir *dintre dels clusters*, menor al de la solució triada, el que vol dir que tot i que els dos clusters que s'han obtingut estan menys separats entre ell, però són més homogenis. Aquesta solució agrupa els capítols en clusters de 219 i 206 capítols.

L'estadístic χ^2 per la taula de contingència val $\chi^2=21555{,}2$, mentre que el valor entre clusters és de $\chi^2_B=2456.2$, el que dóna una relació $\chi^2_B/\chi^2 = 0{,}114$, és a dir la divisió en aquests dos clusters explica el 11,4% de la "no-homogeneïtat" entre tots els capítols en funció de l'ús de les 25 paraules eina més freqüents.

A l'annex A12.3 hi ha les assignacions dels capítols a cadascun dels 2 clusters tant per *l=12* com per *l=25* paraules eina. Per *l=12* en total, 121 capítols anteriors al 382 i 30 de posteriors han canviat d'assignació. Per *l=25* canvien d'assignació inicial 147 capítols anteriors al 382 i 23 de posteriors.

Els clusters obtinguts de l'aplicació de l'algorisme, tant per dotze com per vint-i-cinc paraules eina, separen els capítols del *Tirant* en funció de la primera component de l'anàlisi de correspondències, com es pot observar del gràfic de la figura 12.12.

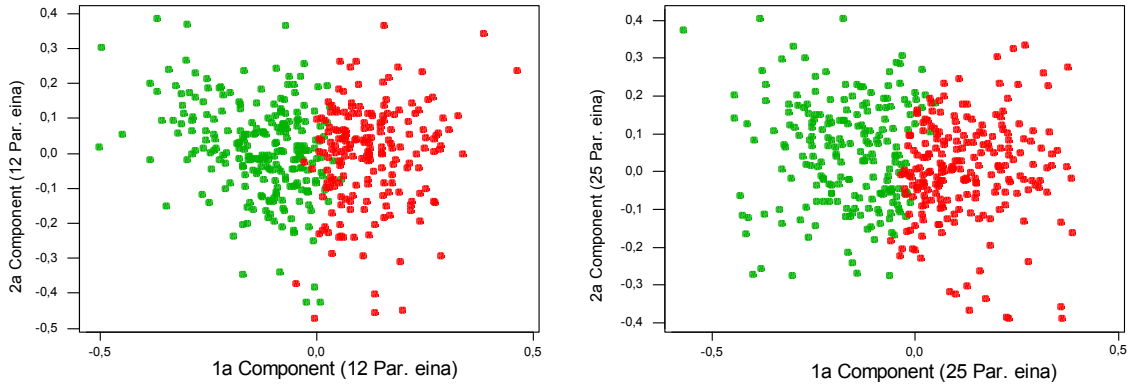

Figura 12.12: Projecció dels perfils fila en el pla format per les dues primeres components de l'anàlisi de correspondències. Els punts en vermell corresponen als capítols assignats al cluster *0*, i els punts en vermell als capítols assignats al cluster *1*, de l'anàlisi cluster basat en la distància χ^2 per 12 (esquerra) i 25 (dreta) paraules eina

12.2.6.2 Anàlisi cluster basada en la deviança

L'execució de l'algorisme cluster basat la deviança dels models de regressió politòmica porta a resultats molt semblants als obtinguts amb l'algorisme basat en la distància χ^2. Quan s'analitza la taula amb *l=12* paraules eina, hi ha un únic capítol que l'algorisme basat en la distància χ^2 agrupa en el cluster *1*, el que conté majoritàriament els capítols del final del llibre, que l'algorisme basat la deviança assigna al cluster *0*, el 219. El guany que s'obté amb aquest mètode respecte a l'aplicació de l'algorisme basat en la distància χ^2, en termes del logaritme de la versemblança, és insignificant.

Quan s'analitza la taula amb *l=25* paraules eina, hi ha 37 capítols anteriors al punt de canvi i 4 posteriors que l'algorisme basat en la distància χ^2classifica en el cluster *1* que l'algorisme basat la deviança assigna al cluster *0*. Els capítols posteriors al punt de canvi són: 428, 476, 477, 480. De nou, el guany que s'obté amb aquest mètode respecte a l'aplicació de l'algorisme basat en la distància χ^2, en termes del logaritme de la versemblança, és extraordinàriament petit.

12.2.6.3 Validació dels resultats obtinguts en l'anàlisi cluster

Per validar els resultats obtinguts en l'anàlisi cluster, s'ha ajustat, tant per *l=12* com per *l=25* categories (*paraules eina*) el model politòmic:

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

on $Ind_{1i}^{(c)}$ és una variable indicadora que pren valor *0* pels capítols assignats al cluster *0* i pren valor *1* pels capítols assignats al cluster *1* pel l'algorisme basat en la distància χ^2, i s'ha calculat la màxima versemblança del model, $L_c\left(\hat{\beta}^{(c)}\right)$.

La comparació dels resultats obtinguts en l'estimació del punt de canvi i en les anàlisi cluster basades tant en la distància χ^2 com en la deviança dels models politòmics es mostren el la taula 12.3. En ella es pot observar com l'anàlisi cluster millora, tot i que de forma poc notable, els estadístics anàlegs utilitzats per a estimar el punt de canvi, pel que es pot assegurar que l'estimació del punt de canvi proposa un bona separació en grups.

Estadístic	***Nº Par. Eina***	***Punt de Canvi***	***Cluster χ^2***	***Cluster Deviança***
$L(\hat{\beta})$	*l=12*	*-272033,8*	*–271496,4*	*-270469,5*
$L(\hat{\beta})$	*l=25*	*-437227*	*-436465*	*-436456*

Taula 12.3: Comparació del valor màxim del logaritme de la versemblança pels models de regressió politòmica ajustats, per les seqüència de *l=12* i *l=25* paraules eina, amb una variable indicadora com a explicativa, que separa els capítols en dos grups en tres situacions: separació donada pel punt de canvi i pels algorismes cluster basats en la distància χ^2 i en la deviança.

12.3 Estudi de les paraules eina per blocs

12.3.1 Notació i forma de les dades

La taula 12.4 és l'anàloga per blocs a la taula 12.1 per capítols. Mostra el tros corresponent als blocs 302-350 de la taula que es farà servir per a l'anàlisi de l'ús de les paraules eina pels blocs, quan s'analitzen les dotze paraules més freqüents. La taula sencera té 396 files, que corresponen als blocs, i 15 columnes: la primera indica el número de bloc, la segona el nombre total de paraules, N_i, que és sempre igual a 1000, les dotze columnes centrals formen una taula contingència que conté el nombre d'ocurrències per cadascuna de les dotze paraules eina per cada bloc, i la darrera columna conté el nombre de vegades que apareixen totes les altres paraules en el bloc, de manera que la suma del contingut d'aquesta columna més les altres dotze és N_i. Aquesta columna s'ha etiquetat com a *Resta*. Per exemple, a la taula es pot llegir com en el bloc 302 la paraula *e* apareix 70 vegades, la paraula *de* 33 vegades, 20 vegades *la*, 33 *que*, ..., 9 vegades *com*. La suma d'aquestes dotze cel·les és 274, pel que *Resta=726* altres paraules en el bloc. La taula que es farà servir per a 25 paraules eina és anàloga, amb 25 columnes centrals i valor per a la columna *Resta* diferent.

Denotem per y_{ji} el nombre d'ocurrències de la paraula *j*-èssima en el bloc i-èssim, per $\hat{\pi}_{ji} = y_{ji}/N_i$ la proporció d'ocurrències i per π_{ji} el seu valor esperat, de manera que:

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

Bloc	**Ni**	***e***	***de***	***la***	***que***	***lo***	***En***	***a***	***per***	***no***	***l***	***los***	***com***	**Resta**
302	*1000*	*70*	*33*	*20*	*33*	*21*	*12*	*18*	*22*	*11*	*8*	*17*	*9*	*726*
303	1000	61	29	40	36	23	17	12	16	12	9	17	11	717
304	1000	43	30	39	32	26	13	21	13	20	6	20	14	723
305	1000	61	28	28	30	35	24	21	15	14	5	21	15	703
306	1000	62	31	33	40	18	21	11	18	15	8	20	16	707
307	1000	73	34	23	34	28	17	16	23	9	6	16	19	702
308	1000	65	30	30	36	26	19	17	16	9	8	26	14	704
309	1000	49	48	26	34	17	19	23	17	18	5	10	18	716
310	1000	66	27	25	36	29	18	20	11	15	3	21	19	710
311	1000	61	29	33	32	39	20	20	11	13	11	15	10	706
312	1000	65	27	26	29	33	22	15	20	16	7	14	13	713
313	1000	61	32	33	22	23	15	22	18	18	5	12	7	732
314	1000	66	25	35	26	33	19	20	10	11	5	8	6	736
315	1000	56	37	27	30	28	16	18	12	8	5	28	15	720
316	1000	52	42	46	42	21	21	24	24	1	7	10	13	697
317	1000	58	31	40	45	22	14	23	13	20	7	19	16	692
318	1000	47	42	27	40	13	16	21	17	20	9	13	10	725
319	1000	56	36	34	26	10	26	18	21	26	8	7	7	725
320	1000	46	46	43	23	18	24	11	11	20	9	7	15	727
321	1000	41	33	31	38	21	16	16	25	22	4	10	7	736
322	1000	53	37	42	33	23	23	14	17	15	8	14	12	709
323	1000	52	29	40	42	27	17	17	12	12	13	5	8	726
324	1000	35	46	43	35	26	23	14	19	16	9	9	12	713
325	1000	54	41	55	32	15	28	17	16	16	12	6	12	696
326	1000	35	45	28	35	17	22	17	25	27	11	8	4	726
327	1000	53	41	49	45	20	19	21	16	11	4	9	14	698
328	1000	54	43	47	43	15	22	19	13	13	13	7	10	701
329	1000	52	40	34	37	7	15	17	15	21	8	10	12	732
330	1000	77	28	47	20	9	16	22	9	24	6	10	6	726
331	1000	76	32	31	16	17	17	14	17	24	14	14	8	720
332	1000	47	31	32	24	16	21	9	19	32	10	17	5	737
333	1000	42	50	40	25	25	15	19	20	15	7	12	8	722
334	1000	67	50	36	21	26	25	17	23	3	16	17	10	689
335	1000	54	50	50	28	28	16	9	20	9	8	16	7	705
336	1000	59	31	32	44	31	13	20	14	18	7	24	11	696
337	1000	73	47	34	29	26	22	21	17	4	30	14	12	671
338	1000	72	55	49	36	16	12	24	7	9	34	5	21	660
339	1000	60	36	49	21	19	15	22	27	11	13	19	12	696
340	1000	79	33	43	34	21	26	13	18	12	14	11	19	677
341	1000	69	51	28	32	31	18	20	17	4	2	11	10	707
342	1000	62	52	41	26	21	20	18	21	14	8	11	7	699
343	1000	58	42	30	36	26	20	10	21	10	13	17	9	708
344	1000	74	54	34	21	42	29	26	15	3	17	8	8	669
345	1000	72	76	40	26	36	20	19	15	6	15	7	18	650
346	1000	67	64	40	33	33	13	16	14	6	12	10	17	675
347	1000	72	54	35	28	15	24	15	26	7	8	7	13	696
348	1000	73	26	29	31	28	25	18	16	17	14	4	7	712
349	1000	62	31	33	40	25	22	21	21	19	8	7	12	699
350	1000	82	47	40	28	33	25	20	12	9	14	10	8	672

Taula 12.4: Part corresponent als blocs 302-350 de les dades utilitzades en l'estudi de l'ús de paraules eina. Cada fila correspon a un bloc, la primera columna dóna el número de bloc, la segona el nombre de paraules per bloc Ni, les dotze següents el nombre de vegades que apareixen les 12 primeres paraules eina per bloc, i la darrera la suma de totes les altres paraules.

Sota la hipòtesi de independència, y_{ji} té distribució Binomial ($N_i=1000$, π_{ji}). Per tant:

$$E(y_{ji})=N_i\pi_{ji}=1000\pi_{ji},$$

$$Var(y_{ji})=N_i\pi_{ji}(1-\pi_{ji})=1000\pi_{ji}(1-\pi_{ji}).$$

Observem com els valors esperats i les variances de y_{ji}, per $j=1,2,...,l$, no depenen de N_i, en ser constant, i com tant el valor esperat com la variança de y_{ji} depenen del capítol, i.

Cadascuna de les $N_i=1000$ formes gràfiques (paraules) d'un bloc pot ser classificada en una de les $l+1$ categories, amb $l=12$ *o* 25 paraules eina. La categoria addicional és la formada per la variable etiquetada com a *Resta*. El vector de comptatges $\boldsymbol{y_i}=(y_{1i},y_{2i,...},y_{l+1i})$ satisfà

$$\sum_{j=1}^{l+1} y_{ji} = N_i = 1000\,,$$

i es distribueix $Mult(N_i,\pi_i=(\pi_{1i},\pi_{2i},...,\pi_{l+1i}))$, on π_i és el vector amb les probabilitats de que una paraula presa a l'atzar en el bloc i-èssim sigui la j-èrssima, per a $i=1,...,396$.

12.3.2 Anàlisi descriptiva univariant de les dades

La figura 12.13 conté els histogrames de la variable Resta per 12, esquerra, i 25, dreta, paraules eina. Recordem que anomenem

$$Resta_i = N_i - \sum_{j=1}^{l} y_{ji}\,,$$

el total de paraules diferents a les l paraules eina considerades.

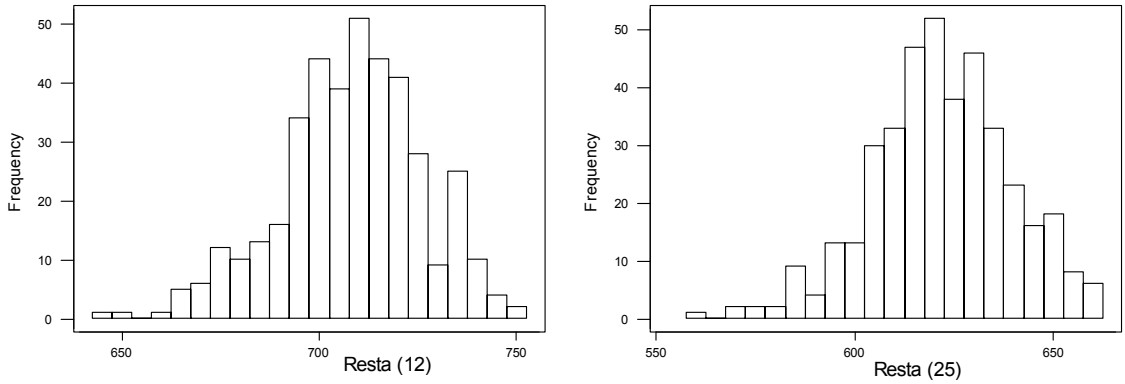

Figura 12.13: histogrames del nombre de paraules diferents a les 12 (esquerra) i 25 (dreta) *paraules eina* considerades (variable *Resta*).

El nombre de paraules per bloc que no corresponen a una de les l paraules eina és una variable discreta i, per tant, no Normal. De totes formes, si l'estil fos el mateix al llarg de tot el llibre, pel Teorema Central del Límit, la distribució de la variable *Resta* podria aproximar-se per la Normal. Els histogrames de les variables *Resta* per $l=12$ i per $l=25$ de la figura 12.13 mostren distribucions no Normals, el que porta a pensar que podem trobar-nos davant una mescla de distribucions, que possiblement impliqui la presència de més d'un estil.

12.3.3 Gràfics de Control per les proporcions de paraules

De manera anàloga a com s'ha fet per capítols, representem gràficament l'evolució de les proporcions d'ús de les paraules eina, y_{ji}/N_i, al llarg del *Tirant* per observar si les π_{ji} són constants i, per tant, de si l'estil és homogeni o be hi ha algun punt en el que canvi. Si no hi ha canvi d'estil, el valor esperat de y_{ji}/N_i serà constant al llarg del llibre. Si no hi ha canvi d'estil, com que $N_i=1000$ per a tots els blocs, tant el valor esperat com la variança de y_{ji} també seran constants al llarg del *Tirant*.

La figura 12.14 mostra l'evolució al llarg de tot el llibre del nombre d'ocurrències de cadascuna de les 25 paraules eina més freqüents, mitjançant un gràfic *NP*. Els límits de control s'han obtingut a partir de les dades en tots els blocs, tal i com s'ha explicat en el capítol 5. La informació que es pot extreure de l'anàlisi de cadascun dels gràfics per separat és, bàsicament, la mateixa que obteníem pels capítols:

- *e*: en el bloc 330, aproximadament, s'observa un augment en el nivell,
- *de*: en el bloc 330, aproximadament, s'observa un augment en el nivell,
- *la*: en el bloc 321, aproximadament, s'observa un augment en el nivell,
- *que*: en el bloc 370, aproximadament, s'observa una disminució en el nivell,
- *lo*: en el bloc 90, aproximadament, s'observa una disminució en el nivell,
- *en*: no s'observa cap canvi significatiu en el nivell,
- *a:* no s'observa cap canvi significatiu en el nivell,
- *per:* no s'observa cap canvi significatiu en el nivell,
- *no:* s'hi observen dos possibles punts de canvi, el primer cap al bloc 100 i el segon cap el 332. Abans del primer punt de canvi el nivell està lleugerament per sota de la mitjana global, entre els dos punts de canvi el nivell està clarament per sobre de la mitjana global, mentre que després del segon canvi el nivell està per sota tant de la mitjana global com del nivell d'abans del primer punt de canvi,
- *l:* entre els blocs 275 i 325 hi ha un notable descens en el nivell. Després del bloc 325 el nivell és semblant al dels blocs anteriors al 275, tot i que la variabilitat sembla augmentar,
- *los:* no s'observa cap canvi significatiu en el nivell,
- *com:* en el bloc 365, aproximadament, s'observa una disminució en el nivell,
- *ab*: no s'observa cap canvi significatiu en el nivell,
- *les*: en el bloc 200, aproximadament, s'observa una disminució en el nivell,
- *d*: no s'observa cap canvi significatiu en el nivell,
- *li*: en el bloc 365, aproximadament, s'observa una disminució en el nivell,
- *qui*: no s'observa cap canvi significatiu en el nivell,
- *del*: en el bloc 345, aproximadament, s'observa un augment en el nivell,
- *se:* no s'observa cap canvi significatiu en el nivell,
- *molt:* en el bloc 333, aproximadament, s'observa un augment en el nivell,
- *és:* no s'observa cap canvi significatiu en el nivell,
- *gran:* no s'observa cap canvi significatiu en el nivell,
- *jo:* en el bloc 310, aproximadament, s'observa una disminució en el nivell,
- *si:* en el bloc 332, aproximadament, s'observa una disminució en el nivell,

- *dix:* després del bloc 270, aproximadament, hi ha una disminució en el nivell.

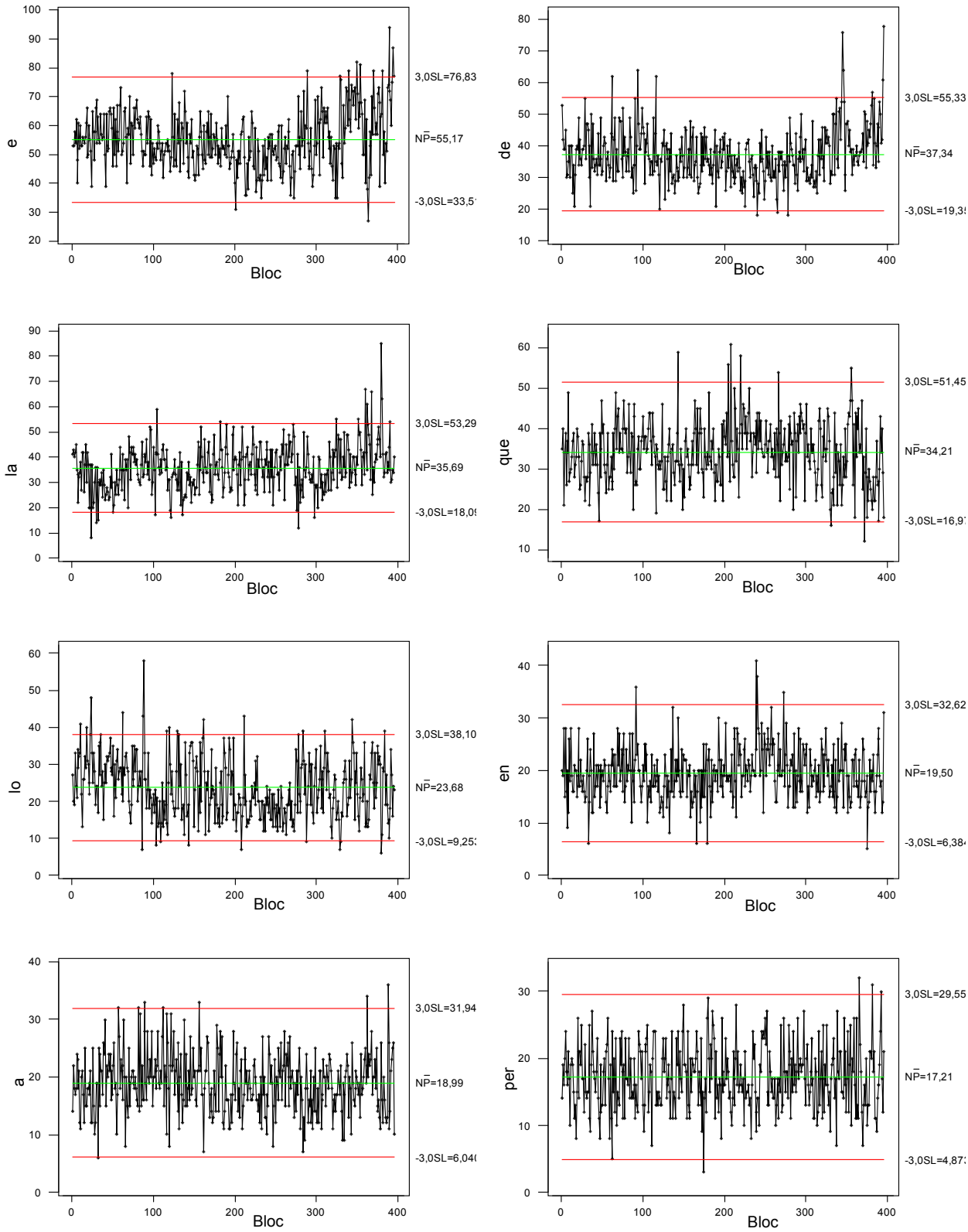

Figura 12.14a: Evolució temporal del nombre d'ocurrències de les paraules eina: *e, de, la, que, lo, en, a, per* en cada bloc.

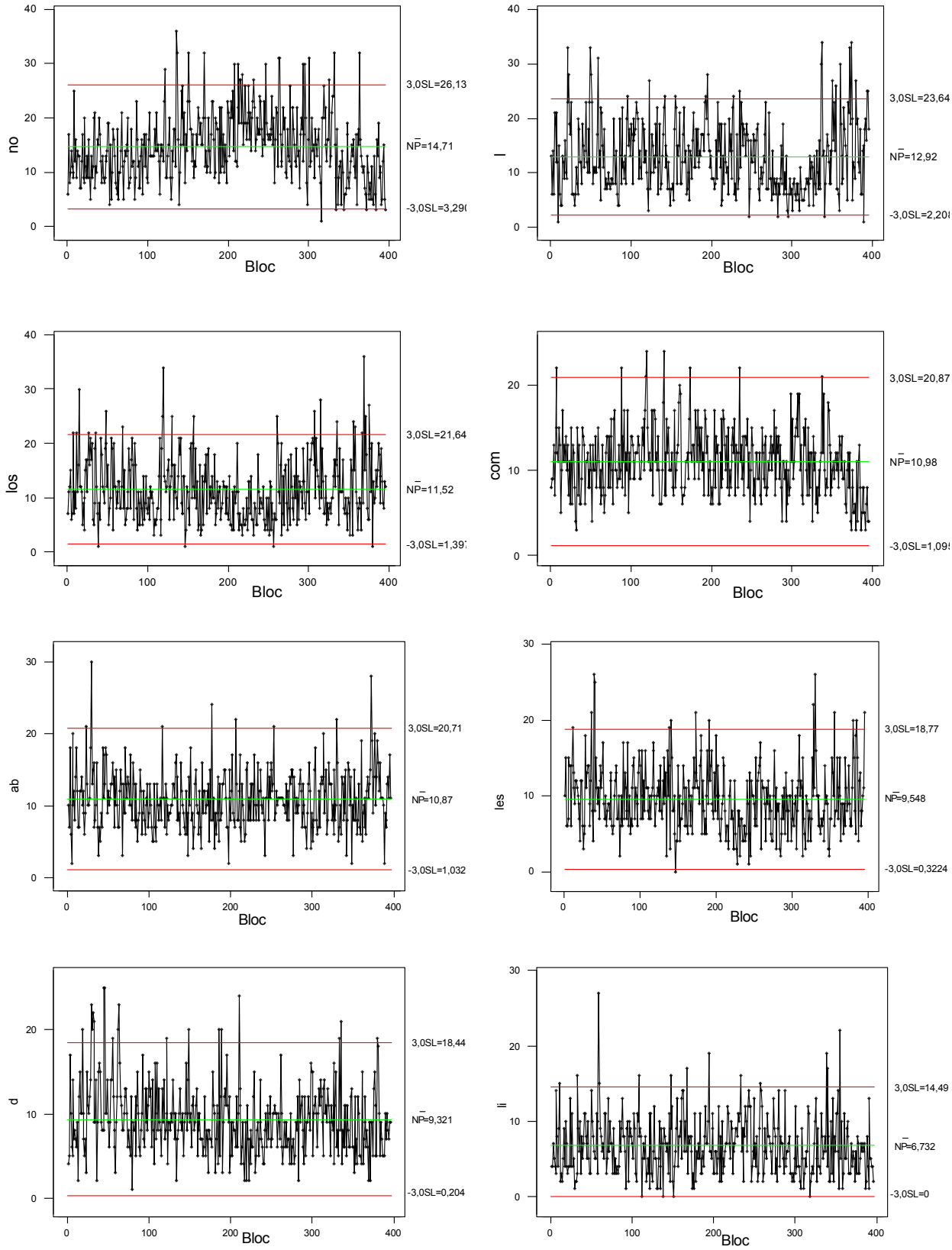

Figura 12.14b: Evolució temporal del nombre d'ocurrències de les paraules eina *no, l, los, com, ab, les, d, li* en cada bloc.

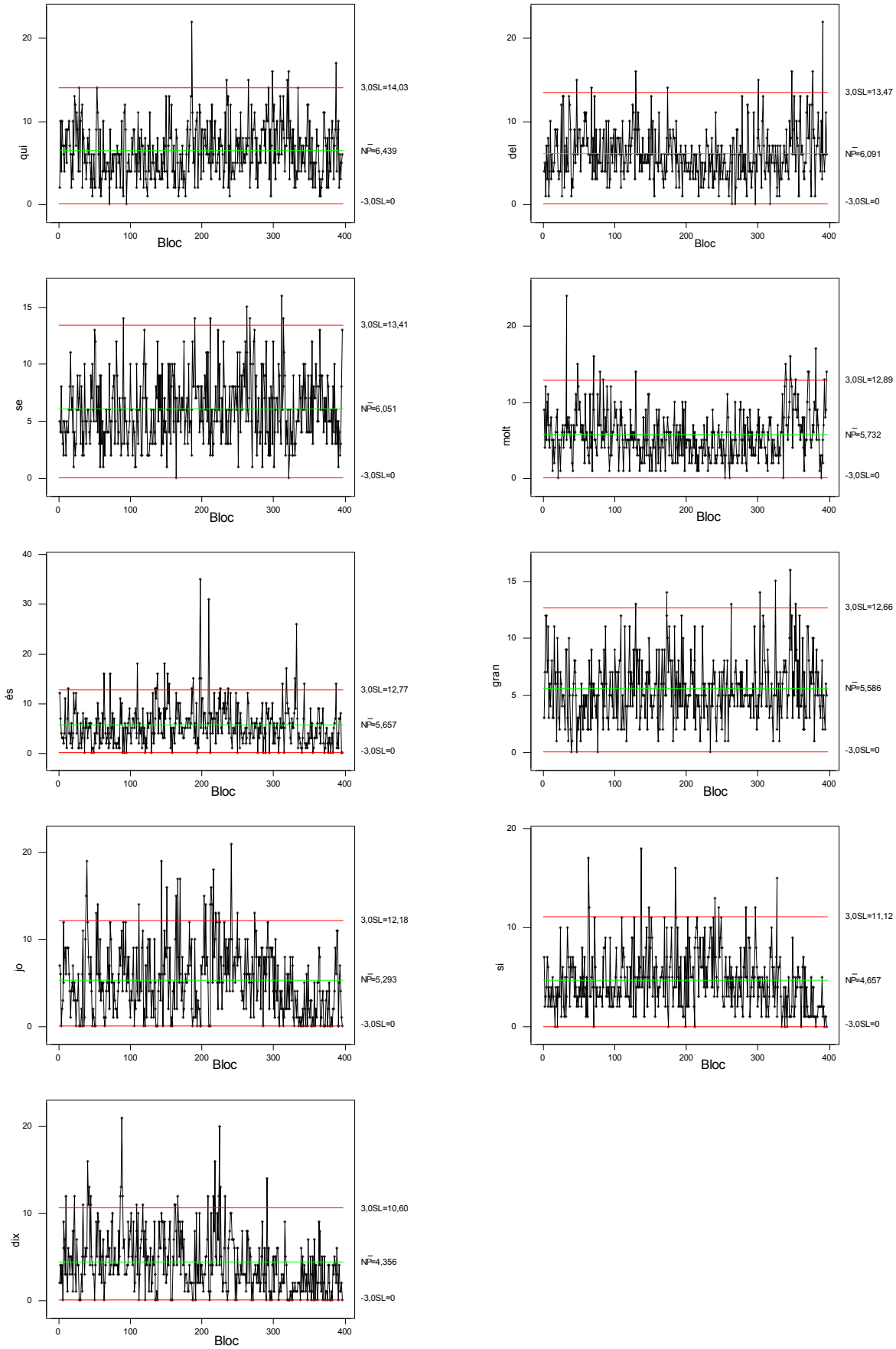

Figura 12.14c: Evolució temporal del nombre d'ocurrències de les paraules eina: *qui, del, se, molt, és, gran, jo, si, dix* en cada bloc.

La figura 12.15 mostra l'evolució, al llarg del T*irant*, del nombre de paraules diferents a les 12 i 25 *paraules eina* considerades: les seqüències etiquetades com a *Resta*. La informació continguda explica si l'estil afavoreix l'aparició de les paraules que són més freqüents en el llibre o, pel contrari, aquestes paraules són menys usades. En totes dues seqüències es pot apreciar una forta disminució en el nivell en el bloc 333, que correspon, aproximadament, al capítol 383, que s'ha senyalat amb un canvi de color.

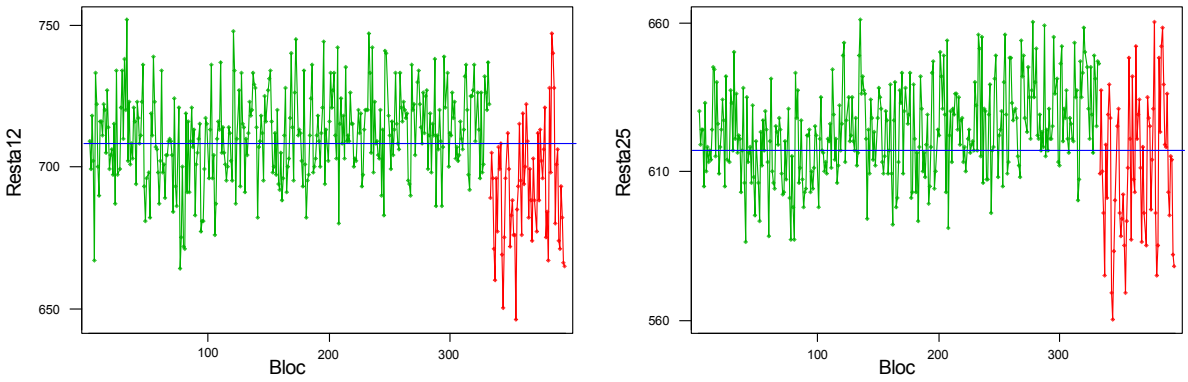

Figura 12.15: Evolució temporal del nombre de paraules diferents a les 12 (esquerra) i 25 (dreta) *paraules eina* considerades. S'han representat en verd els blocs 1-333, i en vermell els blocs 334-final

La figura 12.16 mostra l'evolució temporal de les dues paraules que, tot i estar entre les més freqüents, no s'han considerat en l'estudi per dependre del context. Com ja observàvem pels capítols, la paraula *Tirant* no surt fins al bloc 20, i hi ha parts del llibre on apareix freqüentment la paraula *rei*, mentre n'hi ha d'altres en les que apareix de forma molt més escassa, en funció de l'argument.

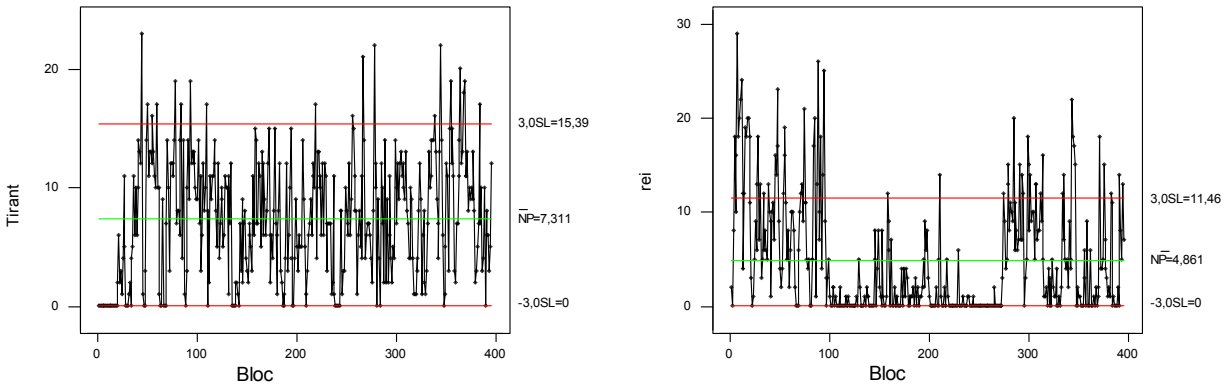

Figura 12.16: Evolució temporal de les paraules descartades de l'estudi entre les més freqüents (*Tirant* i *rei*), en considerar-les depenents del context.

La Figura 12.17 mostra el gràfic CUSUM per la variable *Resta* per *l=12* i *l=25*. En el capítol 5 s'ha descrit com aquests gràfics representen l'evolució temporal de:

$$S_m = \sum_{i=i}^{m} \left(\mathrm{Re}\, sta_i - \mu_0 \right) = \sum_{i=i}^{m} \left(N_i - \sum_{j=1}^{l} y_{ji} + \mu_0 \right),$$

per *m=1,2,..., 396*, i on S_m rep el nom de suma acumulada fins al bloc *m*. S'ha triat, en cada cas, com a μ_0 la mitjana de la variable *Resta* per a tots els blocs. Tots dos gràfics mostren com en els primers blocs, fins al 100 aproximadament, la suma acumulada disminueix. Aquesta disminució és molt més accentuada per 25 paraules eina. A partir

del mínim S_m té una clara tendència a augmentar fins al bloc 333, on hi ha el màxim per, a partir d'aquest punt, disminuir fins a acabar prenent valor *0* per al darrer bloc. El fet que S_m tingui un màxim en el bloc 333 indica que en aquest punt hi ha un canvi en el valor esperat de la variable *Resta*, i que l'ús de les *l* paraules més freqüents pels blocs anteriors al màxim és inferior a la mitjana global, mentre que pels blocs posteriors al màxim és més gran.

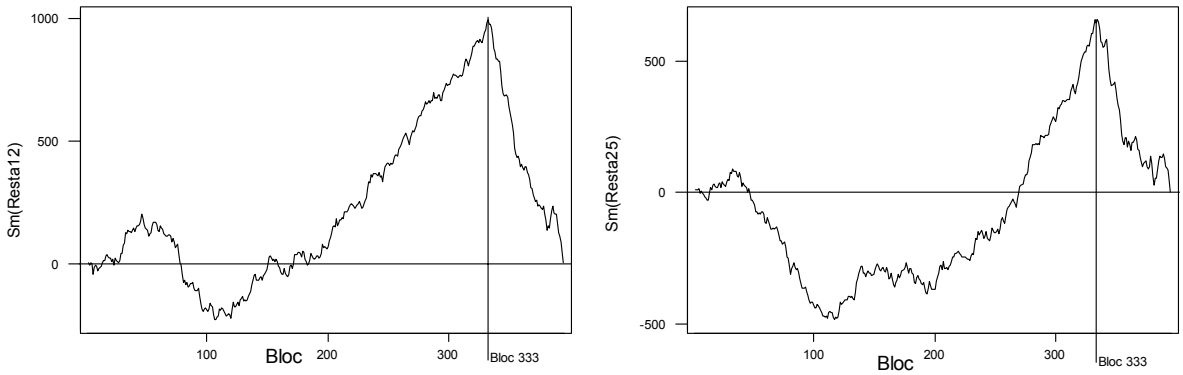

Figura 12.17: Gràfic CUSUM per nombre de paraules diferents a les12 (esquerra) i 25 (dreta) *paraules eina* considerades, *Resta*, pels blocs. S'ha pres com a μ_0 la mitjana de la variable. El màxim de la suma acumulada es té pel bloc 333 en ambdós casos.

Es pot representar la seqüència de dades multinomials en un gràfic de control Chi-quadrat tal i com s'ha explicat en el capítol 5 i tal i com ja s'ha fet per la llargada de paraula i pels capítols. Els gràfics de la figura 12.18, per *l=12* i *l=25* s'han obtingut prenent com a estimació de les π_j, $\hat{\pi}_j$, les mitjanes de les proporcions de la paraula *j*:

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

i monitoritzant 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=12* o *l=25*. En els gràfics es pot observar com el punt de canvi sembla coincidir amb el canvi de color, en el bloc 333. S'observa com després del punt de canvi no només varia el nivell si no que, a més a més, augmenta la variabilitat.

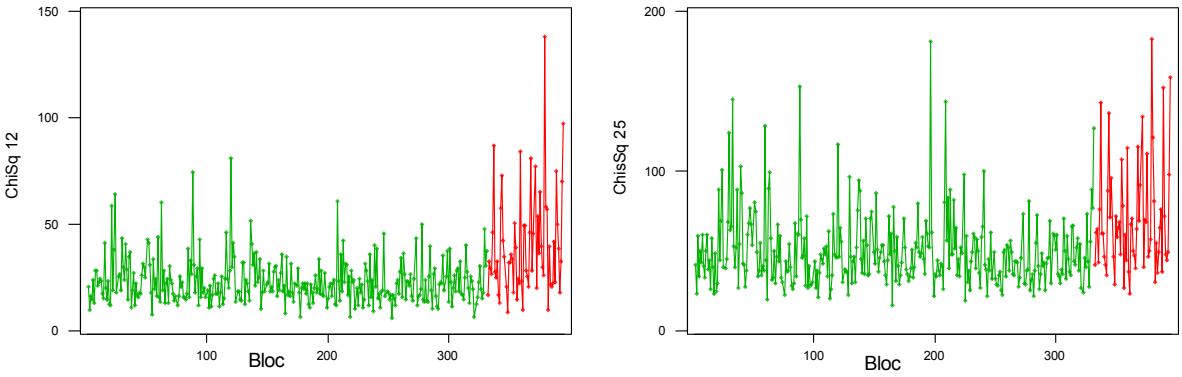

Figura 12.18: Gràfic Chi-quadrat per les seqüències multinomials formades per 12 (esquerra) i 25 (dreta) paraules eina. Hem representat en verd els blocs 1-333 i en vermell els blocs 333-final.

12.3.4 Anàlisi de Correspondències l'ús de paraules

Quan s'analitzen l'ús de 12 paraules eina, el punt de partida per a l'anàlisi de correspondències està format per la taula 12.4. Per 25 paraules eina la taula és anàloga. La figura 12.19 conté els gràfics simètrics tant per files, blocs, com per columnes, paraules, i tant per *l=12* com per *l=25*.

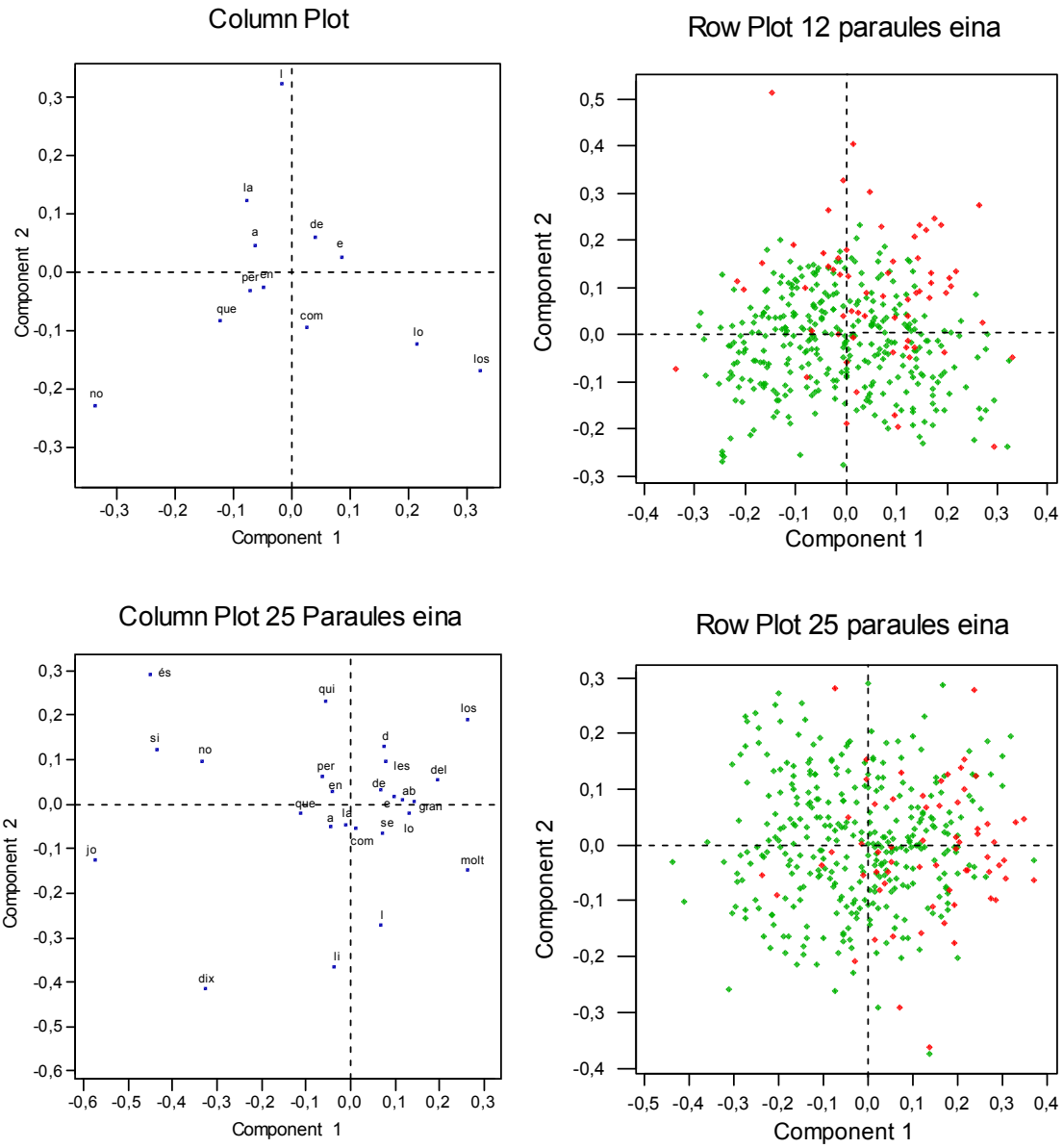

Figura 12.19: Gràfics simètric per columnes a l'esquerra i files a la dreta de l'anàlisi de correspondències per 12 (a d'alt) i per 25 (a baix) paraules eina. S'han representat en color verd els primers 332 blocs i en vermell els blocs 333-396.

Els gràfics que s'obtenen tant per dotze com per vint-i-cinc paraules eina són molt semblants als obtinguts en l'anàlisi dels capítols i, per tant, valen les mateixes consideracions. La única diferència és que per 12 paraules eina la segona component té el sentit invers al que tenia en l'estudi dels capítols. Així, la majoria de punts que corresponen a blocs posteriors al 332 tenen valors de la segona component positius en el gràfic de les files. Pel que fa a la primera component no sembla discriminar entre els

blocs anteriors i els posteriors al 332, tot i que la majoria dels blocs posteriors al 332, en vermell a la figura, prenen valors positius.

Observem en el gràfic simètric per columnes, a l'esquerra en el gràfic, com per *l=12* les paraules que estan en el primer quadrant (*e* i *de*) són les que tenen el canvi més acusat al voltant del bloc 332 i, a més a més, després del canvi la proporció d'aquestes paraules augmenta. De les paraules que es troben en el tercer quadrant: *no*, que és la que es troba més allunyada de l'origen, té el canvi molt marcat en el bloc 333 però en sentit contrari: la proporció de *no* disminueix després del canvi; *que* té el canvi menys marcat i també disminueix després del bloc 333. Les altres dues paraules que es troben en el tercer quadrant, *per* i *en*, es troben molt més a prop de l'origen i en la seqüència temporal no s'aprecia cap canvi en el nivell. Pel que fa a les paraules que es troben en el segon quadrant, *l*, la més allunyada de l'origen, té un canvi de nivell proper al bloc 333, augmentant la proporció després del punt de canvi, *la* té un comportament semblant, mentre que en *a*, molt més propera a l'origen, no s'observa un canvi de nivell. De les paraules que es troben en el quart quadrant, *los, lo* i *com*, en cap d'elles s'observa un canvi de nivell al voltant del bloc 333. Per *l=25* el mostra com les paraules amb valor absolut de la primera component més elevat són algunes de les que millor marquen un canvi al voltant del bloc 333: en *molt*, amb valor positiu, augmenta la proporció després del canvi, mentre que en *jo, si, és, no*, amb valor negatiu, disminueixen després del canvi. Cal observar, com passava en els capítols, que aquestes paraules no són de les més freqüents i que algunes paraules molt més freqüents i que presenten un punt de canvi molt clar al voltant del bloc 333, *e, de, la*, es troben en el gràfic molt més a prop de l'origen.

Els biplots que mostren en el mateix gràfic files i columnes, tant el gràfic simètric com els asimètrics per files i per columnes, són poc útils. El fet de tenir prop de 400 files i, per tant, prop de 400 punts representats en molt poc espai en dificulta extraordinàriament la seva interpretació.

L'evolució temporal de les dues primeres components per files, representades en els gràfics de control per observacions individuals de la figura 12.20, mostra com el comportament és força diferent si es consideren 12 o 25 *paraules eina*. Per 12 paraules s'observa com el canvi entre els blocs anteriors i els posteriors al 332 és molt més evident per a la segona component, i com els blocs posteriors semblen tenir valor esperat més gran que els anteriors. Per la primera component gairebé tots els blocs posteriors al 332 prenen valors positius, comportament que ja es dóna en els primers 100 blocs del llibre, mentre que els blocs entre el 100 i el 332 prenen valors tant positius com negatius. Quan s'analitzen 25 paraules eina s'observa un augment en el nivell de la primera component a partir del bloc 332, mentre que la segona component no sembla indicar cap canvi rellevant. Aquest comportament és el mateix que havíem detectat pels capítols.

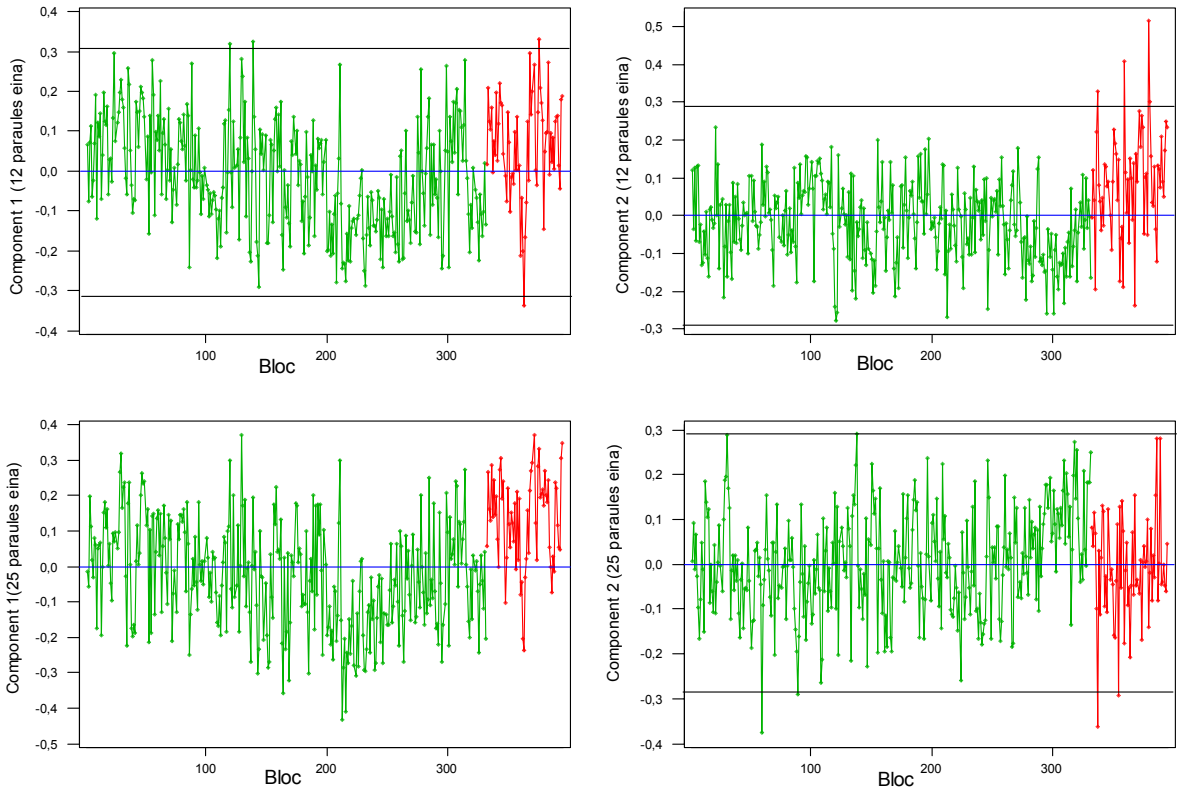

Figura 12.20: Evolució temporal de les dues primeres components de l'anàlisi de correspondències: en els gràfics de l'esquerra hi trobem la primera component i en els de l'esquerra la segona, quan s'analitzen 12 (a d'alt) i 25 (a baix) *paraules eina*. El color verd correspon als blocs 1-332 i el color vermell als blocs posteriors.

12.3.5 Estimació del punt de canvi

Per a l'estimació del punt de canvi es segueix el mateix procés que s'ha fet servir pels capítols: en primer lloc es fa servir el mètode proposat per seqüències de Normals, i s'aplica a les seqüències dels valor de les dues primeres components de l'anàlisi de correspondències, en ambdós casos tant per 12 com per 25 *paraules eina*. A continuació, mitjançant l'aproximació al problema basada en models logístics, s'estudien les 25 seqüències de paraules eina per separat i les seqüències de proporcions de paraules diferents de les 12 o 25 paraules eina considerades, és a dir, la proporció que suposa la variable *Resta*. A continuació, s'estima el punt de canvi en la seqüència de dades multinomials contingudes en la taula 12.4 via l'ajust de models politòmics.

12.3.5.1 Punt de canvi en seqüències Normals

S'ha estimat el punt de canvi per la seqüència de valors de les dues primeres components de l'anàlisi de correspondències fent servir les tècniques proposades per a dades Normals descrites a la secció 6.2. Si anomenem pc_{1i} i pc_{2i} als valors de les dues primeres components en el bloc i-èssim per l'anàlisi fet amb dotze paraules, sota les hipòtesi de independència i Normalitat, es distribueixen:

$$pc_{1i} \sim N(\mu_{1i} = \beta_{10}^{(r)} + \beta_{11}^{(r)} Ind_{1i}^{(r)}, \sigma_1^2),$$

i

$$pc_{2i} \sim N(\mu_{2i} = \beta_{20}^{(r)} + \beta_{21}^{(r)} Ind_{1i}^{(r)}, \sigma_2^2),$$

per *r=1,2,...,n-1*, 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*. Per vint-i-cinc paraules els models són anàlegs. 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 variable indicadora $Ind_{1i}^{(r)}$ ajustat fent servir el criteri dels mínims quadrats.

El gràfic de la figura 12.21 mostra l'evolució de F_r en funció de *r* per a les quatre seqüències. La millor estimació del punt de canvi per a la seqüència del valor de la primera component per 12 paraules eina és $\hat{r}_N$=85, tot i que també hi ha màxims local per a *r=298, 332 i 366*. En el gràfic de F_r en funció de *r* per la segona component observem un punt de canvi per a *r=336*, el que confirma el que s'havia vist de l'anàlisi del gràfic de control. El valor de F_r per al màxim és força més gran per a la segona component, el que indica que el canvi és més accentuat.

Quan s'analitza la seqüència de 25 paraules eina, el màxim és molt més accentuat per a la primera component, marcant el punt de canvi en el bloc 332.

En ambdós casos, l'estimació del punt de canvi corrobora el que s'ha vist en l'estudi gràfic de les seqüències de valors de les dues primeres components i el que havíem vist en l'estudi dels capítols.

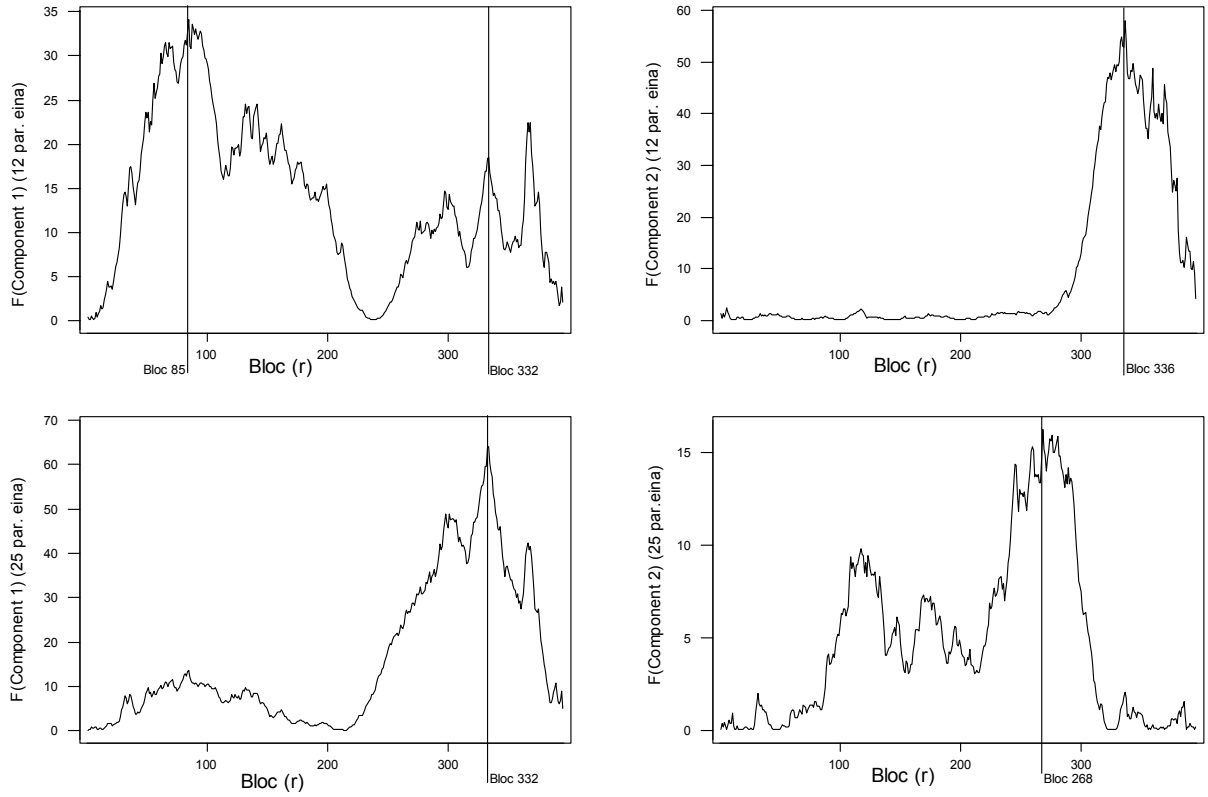

Figura 12.21: Evolució de F_r en funció de r per a la seqüència de valors de les dues primeres components de l'anàlisi de correspondències per 12 i 25 paraules eina. Els gràfics de dalt corresponen a l'estudi de 12 paraules, els de baix a 25.

12.3.5.2 Punt de canvi en seqüències binomials

En aquest apartat s'estudien per separat les seqüències univariants de les *l=12* i *l=25 paraules eina* separat i les seqüències de proporcions de paraules diferents de les dotze o vint-i-cinc paraules eina considerades. Aquest procediment dóna *l+2* estimacions del punt de canvi, possiblement diferents.

Suposant que existeix un punt de canvi a r, per una categoria j fixada $j=1,2,...l$, y_{ji} té una distribució $binomial(N_i=1000,\pi_{ja})$ per $i=1,...,r$ i $binomial(N_i=1000,\pi_{jd})$ per $i=r+1,...,n$. Per cadascuna de les l categories s'ajusta el model logístic 6.12:

$$y_{ji} \sim Binomial\left(N = 1000, \pi_{ji} = \frac{e^{\beta_0^{(r)} + \beta_1^{(r)} Ind_{1i}^{(r)}}}{1 + e^{\beta_0^{(r)} + \beta_1^{(r)} Ind_{1i}^{(r)}}} \right),$$

i es representa el màxim del logaritme de la versemblança, $L_j\left(r, \hat{\beta}^{(r)}\right)$, en funció de r. La figura 12.22 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 25 paraules estudiades.

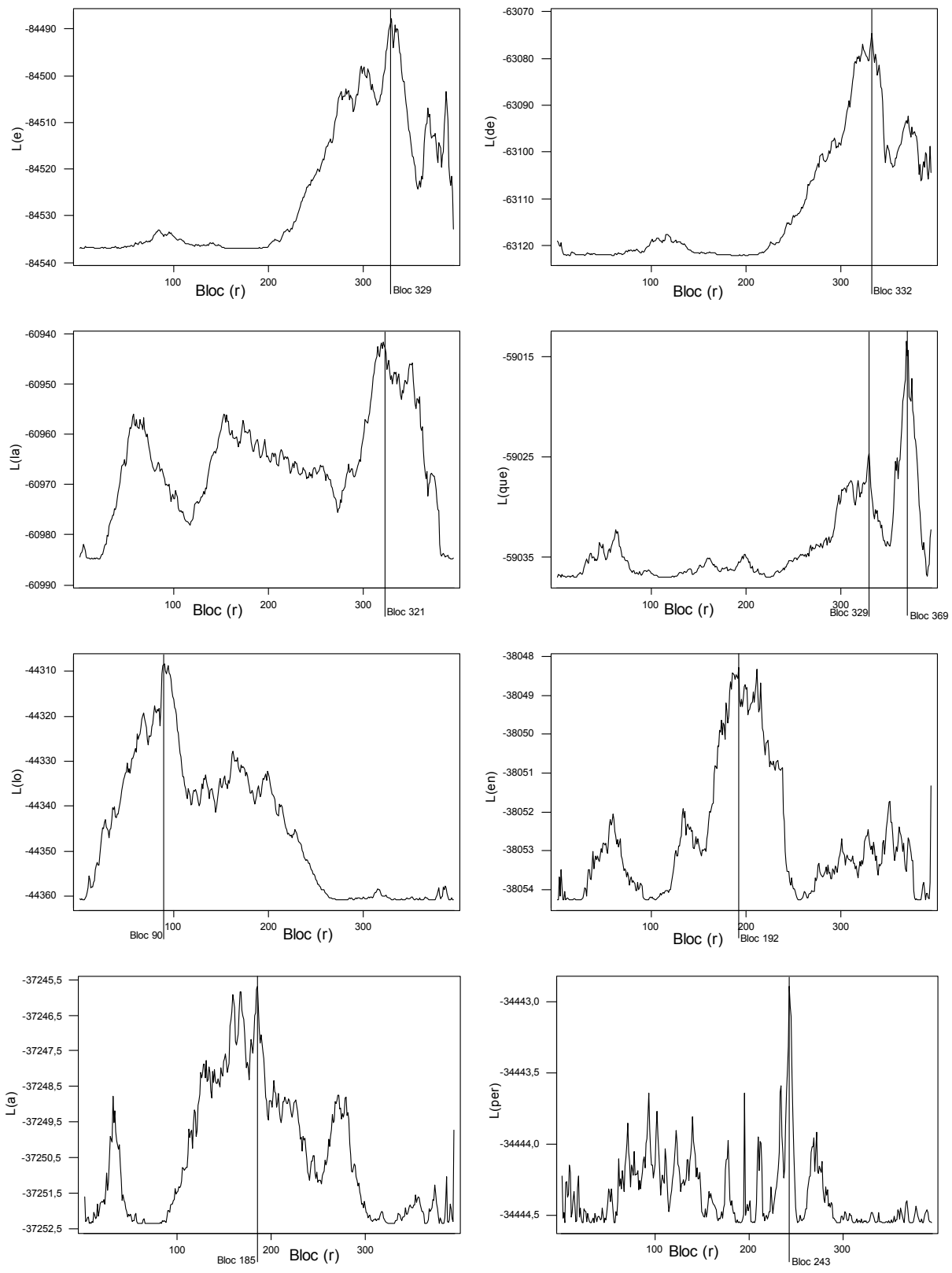

Figura 12.22a: Relació de $L_j\left(r, \hat{\beta}^{(r)}\right)$ en funció de *r* per als models amb resposta el nombre d'aparicions de les paraules eina: *e, de, la, que, lo, en, a, per* per blocs.

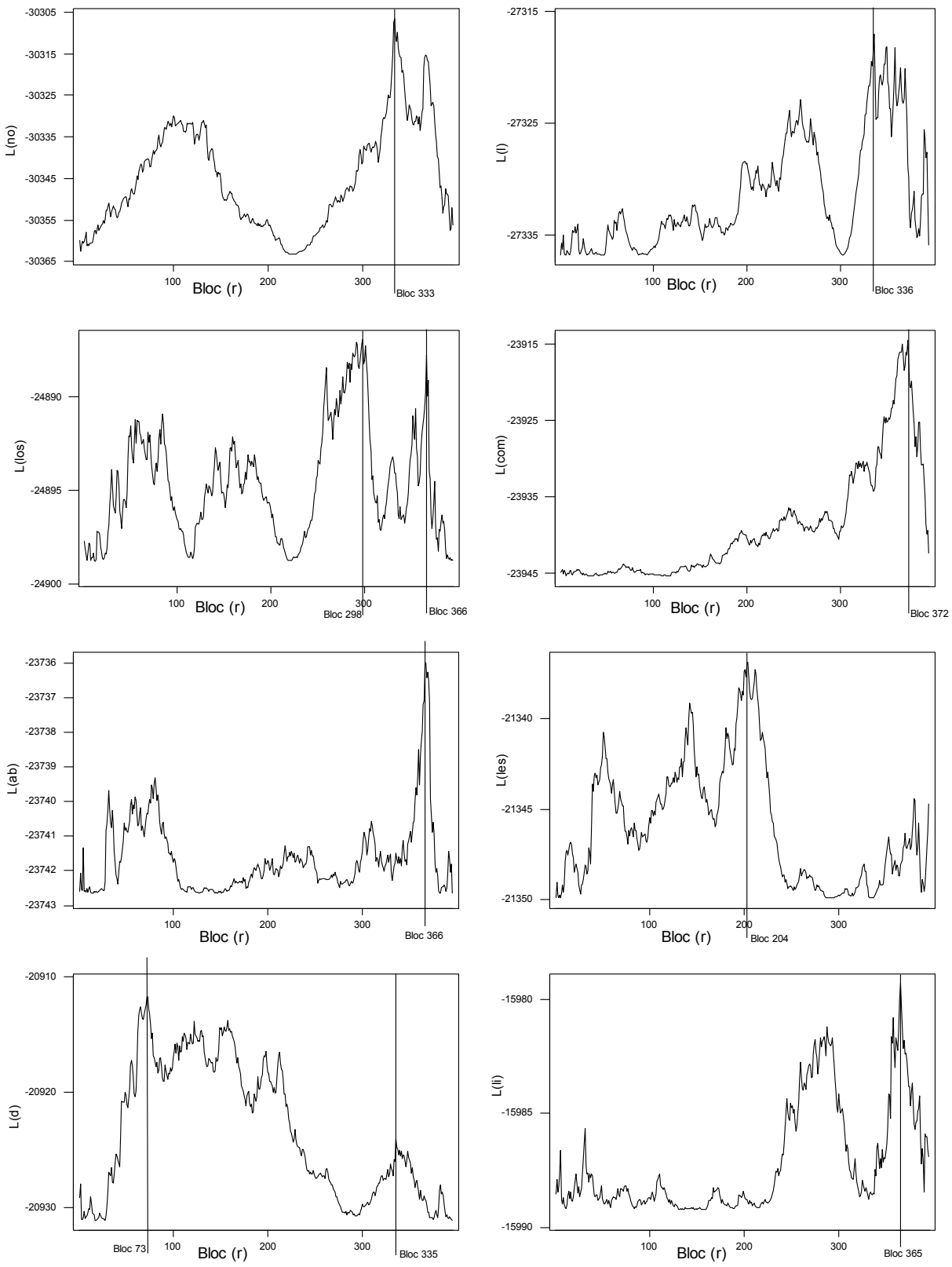

Figura 12.22b: Relació de $L_j\left(r,\hat{\beta}^{(r)}\right)$ en funció de *r* per als models amb resposta el nombre d'aparicions de les paraules eina: *no, l, los, com, ab, les, d, li* per blocs.

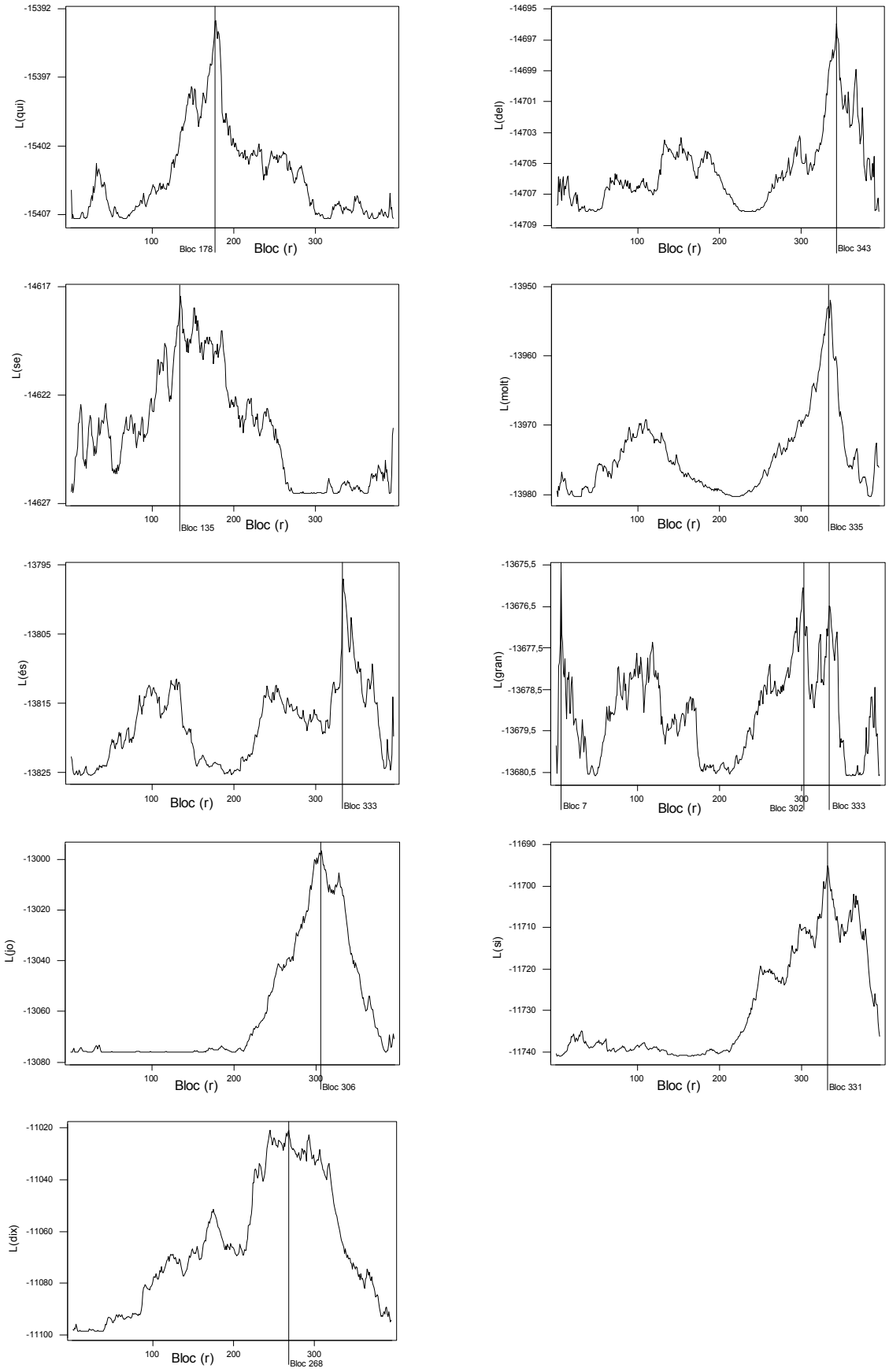

Figura 12.22c: Relació de $L_j\left(r,\hat{\beta}^{(r)}\right)$ en funció de *r* per als models amb resposta el nombre d'aparicions de les paraules eina: *qui, del, se, molt, és, gran, jo, si, dix* per blocs.

Observem com el màxim del logaritme de la versemblança és més gran a mesura que $\hat{\pi}_j$ són més petites. Això vol dir que si volem resumir les *l* estimacions del punt de canvi en un sol valor, adaptant els estadístics proposats per Wolfe i Chen (1990) com hem descrit al capítol 6, donarem més pes en l'estimació a les paraules amb menor probabilitat d'aparició.

S'ha estimat el punt de canvi per la seqüència de proporcions de paraules diferents de les dotze o vint-i-cinc paraules eina considerades, $Resta_i$, fent servir les tècniques proposades per a seqüències de binomials.

Els gràfics de la figura 12.23 mostren l'evolució de $L_j\left(r,\hat{\beta}^{(r)}\right)$ en funció de *r* quan estudiem dotze i vint-i-cinc paraules eina. El màxim de $L_j\left(r,\hat{\beta}^{(r)}\right)$ i, per tant, la millor estimació del punt de canvi per a les seqüències de $Resta_i$ pels capítols per *l=12* és $\hat{r}_L$ =332 i per *l= 25* és $\hat{r}_L$ =335. El gràfic mostra com els màxims en $\hat{r}_L$ estan clarament diferenciats.

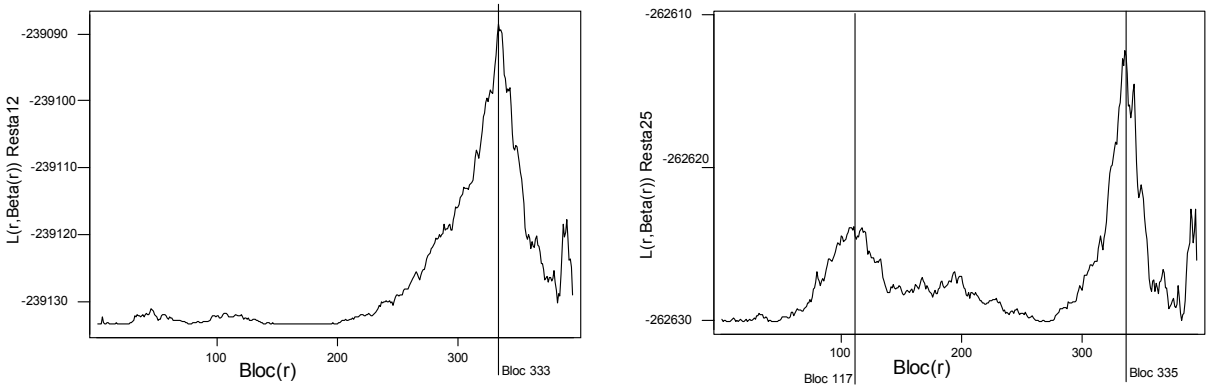

Figura 12.23: Gràfic que representa l'evolució de $L_j\left(r,\hat{\beta}^{(r)}\right)$ en funció de *r* per a l'estimació del punt de canvi en les seqüències del nombre de paraules diferents a les *paraules eina* considerades, *Resta*..

12.3.5.3 Punt de canvi en seqüències multinomials; model politòmic

El problema d'estimar el punt de canvi en la seqüència multinomial formada per les *l* paraules, s'ha abordat mitjançant l'ajust de models politòmics. Si y_{ji} és el nombre d'ocurrències de la paraula *j* en el bloc *i*, per *j=1,2,...,l,* $\mathbf{y_i}=(y_{1i},y_{2i},....,y_{li})$ està distribuït $Mult(N_i,\pi_i)$, on N_i és la llargada del capítol i $\pi_i=(\pi_{1i},\pi_{2i},...,\pi_{li})$ és el vector amb les probabilitats que una paraula 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, $L_j\left(r,\hat{\beta}^{(r)}\right)$, és màxim, per *l=12, 25*. El gràfic de la figura 12.24 mostra l'evolució del màxim del logaritme de la versemblança en funció de *r*. L'estimació del punt de canvi és, tant per *l=12* com per *l=25,* $\hat{r}_M = 333$, que correspon, aproximadament, al capítol 383.

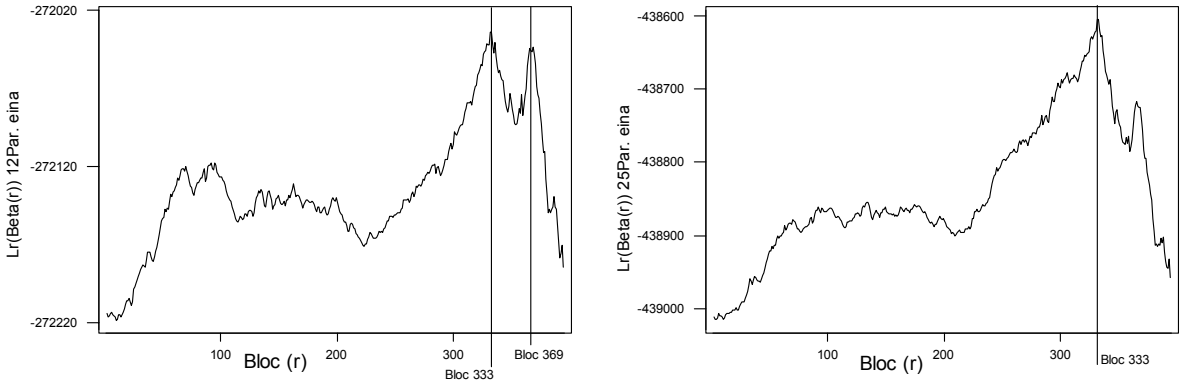

Figura 12.24: Gràfic de l'evolució de $L_j\left(r,\hat{\beta}^{(r)}\right)$ en funció de *r* per a les seqüències multinomials de 12 (esquerra) i 25 (dreta) paraules eina, per blocs. El màxim en ambdós casos es té pel bloc 333, que correspon, aproximadament, al capítol 383.

12.3.5.4 Paraules discriminants

Podem intuir, a partir dels gràfics de la figura 12.22, quines paraules són bones per discriminar entre els dos estils separats pel punt de canvi. La comparació d'aquests gràfics és complicada, molt laboriosa i poc eficient. Per a determinar les paraules marcadores d'estil, fem servir l'aproximació de la binomial a la Normal per ajustar, per cadascuna de les *paraules eina*, el model de regressió lineal amb pesos:

$$\hat{\pi}_{ji} = \frac{y_{ji}}{N_i} \sim N(\beta_{j0} + \beta_{j1} Ind_{1i}^{(r=333)}, \sigma_j^2),$$

on y_{ji} és el nombre d'ocurrències de la paraula *j*-èssima en el bloc i-èssim, $Ind_{1i}^{(r=333)}$ és una variable indicadora que pren valor *0* pels primers 333 blocs, i valor *1* pels blocs del 334 al final. Aquest procediment és equivalent a fer un test de comparació de les mitjanes dels dos conjunts de blocs: 1-333 i 334-final.

Per cada paraula examinada s'observa el valor de l'estadístic:

$$t_r = \frac{b_1}{s_{b_1}},$$

i es consideraran bones discriminants de l'estil totes aquelles paraules per les que el valor absolut de l'estadístic t_r sigui superior a 3. L'annex A12.2 mostra, per cadascuna de les 100 paraules més freqüents, el valor absolut de l'estadístic t_r. Entre les 25 més freqüents n'hi ha 12 que poden ser considerades com a bones discriminants: *e, de, la, no, l, com, del, molt, és, jo, si, dix*. A més a més, *e* i *de* són les dues paraules més freqüents i les més discriminants. Les paraules subratllades són aquelles que la seva proporció disminueix a partir del bloc 333. A banda d'aquestes 12 paraules, entre les 100 més nombroses en trobem altres 16 amb $|t_r|>3$: *era, un, féu, hi, sua, aquell, bé, ho, ni, foren, tal, o, sinó, qual, dir, tu*. La única diferència respecte al resultats trobats en els capítols és la paraula *tu*, que ara té un $|t_r|$ lleugerament superior a *3,* mentre que per capítols era lleugerament inferior a *3*.

Tal i com ja s'havia fet en l'estudi per capítols, no s'han considerat com a possibles paraules discriminants els substantius, en ser sensibles al context.

12.3.5.5 Conclusions

La conclusió d'aquesta anàlisi és que es constata l'existència d'un punt en el que l'estil, quantificat a través de l'ús de paraules eina per blocs, canvia. Els estimadors del punt de canvi de la seqüència multinomial no són tots coincidents, però de forma força consistent el localitzem en el bloc 333 que correspon, aproximadament al capítol 383.

Donant per bo aquest punt de canvi, s'han identificat aquelles paraules que presenten també un canvi en el nombre de vegades que apareixen per capítol al voltant del bloc 333 i que, per tant, poden ser considerades, en el *Tirant*, bones discriminants de l'estil i que es troben en l'Annex A12.2.

La taula 12.5 mostra un quadre resum amb les estimacions del punt de canvi obtingudes en l'estudi de l'ús de les paraules més freqüents pels capítols de més de 200 paraules.

Distribució	**Seqüència**	**Pt. canvi**		**Seqüència**	**Pt. canvi**
Normal	*1ª Comp. A. corr. l=12*	$\hat{r}_N$ =85		*1ª Comp. A. corr. l=25*	$\hat{r}_N$ =332
	2ª Comp. A. corr. l=12	$\hat{r}_N$ =336		*2ª Comp. A. corr. l=25*	$\hat{r}_N$ =268
Binomial	*e*	$\hat{r}_L$ =329		*les*	$\hat{r}_L$ =204
	de	$\hat{r}_L$ =321		*d*	$\hat{r}_L$ =73
	la	$\hat{r}_L$ =356		*li*	$\hat{r}_L$ =365
	que	$\hat{r}_L$ =369		*qui*	$\hat{r}_L$ =178
	lo	$\hat{r}_L$ =90		*del*	$\hat{r}_L$ =343
	en	$\hat{r}_L$ =192		*se*	$\hat{r}_L$ =135
	a	$\hat{r}_L$ =185		*molt*	$\hat{r}_L$ =335
	per	$\hat{r}_L$ =243		*és*	$\hat{r}_L$ =333
	no	$\hat{r}_L$ =333		*gran*	$\hat{r}_L$ =302
	l	$\hat{r}_L$ =336		*jo*	$\hat{r}_L$ =306
	los	$\hat{r}_L$ =298		*si*	$\hat{r}_L$ =331
	com	$\hat{r}_L$ =372		*dix*	$\hat{r}_L$ =268
	ab	$\hat{r}_L$ =366			
	Resta l=12	$\hat{r}_L$ =332		*Resta l=25*	$\hat{r}_L$ =335
Multinomial	*Model Politòmic l=12*	$\hat{r}_M$ =333		*Model Politòmic l=25*	$\hat{r}_M$ =333

Taula 12.5: Quadre resum de les estimacions del punt de canvi obtingudes en l'anàlisi de la l'ús de les paraules més freqüents per als blocs de 1000 paraules.

12.3.6 Anàlisi cluster de les files de la taula de contingència

De l'anàlisi gràfica de les seqüències de paraules eina s'ha pogut apreciar com alguns blocs posteriors al punt de canvi prenen valors que els fan més semblants als anteriors a

$\hat{r}$ que als posteriors, per tant hi ha la possibilitat que en la part final del llibre hi hagi barreja d'estils. Aquesta situació ja s'ha detectat en l'anàlisi de la llargada de paraula i en l'estudi dels capítols.

S'ha decidit fer una anàlisi cluster, per veure si és possible recol·locar algun bloc 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 per a l'aplicació d'aquestes tècniques són les dues taules de contingència, per *l=12* i *l=25 paraules eina*, on les files són els 396 blocs de 1000 paraules i les columnes representen el nombre d'ocurrències per bloc de cadascuna de les *l* paraules considerades.

12.3.6.1 Anàlisi cluster basada en la distància χ^2

Tant per *l=12* com per *l=25* s'ha executat més de 1000 vegades l'algorisme cluster basat en la distància χ^2 amb assignacions inicials aleatòries de capítols a un dels dos grups. El criteri escollit per a decidir l'agrupació en clusters òptima és el de maximitzar χ^2_B, és a dir, la distància *entre clusters*.

Els resultats no han convergit sempre a la mateixa solució. Quan es busquen els dos clusters en funció de les *l=12* paraules eina més abundants, l'agrupació que per la que el valor de χ^2_B és màxim, dóna agrupacions de 205 i 191 blocs. L'algoritme ha convergit prop del 30% de les vegades a aquesta solució. Hi ha una altra solució, a la que ha convergit prop del 5% de les iteracions que, tot i tenir un valor de χ^2_B menor al del màxim, té χ^2_W, és a dir *dintre dels clusters*, menor al de la solució triada, el que vol dir que tot i que els dos clusters que s'han obtingut estan menys separats, són més homogenis.

L'estadístic χ^2 per la taula de contingència completa val χ^2=10059.5, mentre que el valor entre clusters és de χ^2_B =1497.9, el que dóna una relació χ^2_B / χ^2 = 0,149. És a dir, les agrupacions de blocs obtingudes expliquen el 14,9% de la "no-homogeneïtat" entre tots els blocs en l'ús de les 12 paraules eina més freqüents. 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 22%.

Quan s'analitza la taula amb 25 paraules eina, l'algorisme convergeix a dues úniques solucions, una d'elles millor en ser χ^2_B màxim i χ^2_W mínim. Més del 85% de les execucions s'arriba a la solució òptima, en la que els dos clusters contenen, respectivament, 210 i 186 blocs. L'estadístic χ^2 per la taula de contingència val χ^2=20936, mentre que el valor entre clusters és de χ^2_B =2629.6, el que dóna una relació χ^2_B / χ^2 = 0,126.

A la Annex A12.4 hi ha les assignacions dels blocs a cadascun dels 2 clusters tant per *l=12* com per *l=25* paraules eina. Per *l=12*, 143 blocs anteriors al 333 i 15 de posteriors han canviat d'assignació. Per *l=25* canvien d'assignació inicial 155 blocs anteriors al 333 i 9 de posteriors.

Els clusters obtinguts de l'aplicació de l'algorisme, tant per 12 com per 25 paraules eina, separen els blocs del *Tirant* en funció de la primera component de l'anàlisi de correspondències, com es pot observar en el gràfic de la figura 12.25, en el que hi ha les projeccions dels perfils fila dels blocs en el pla generat per les dues primeres components.

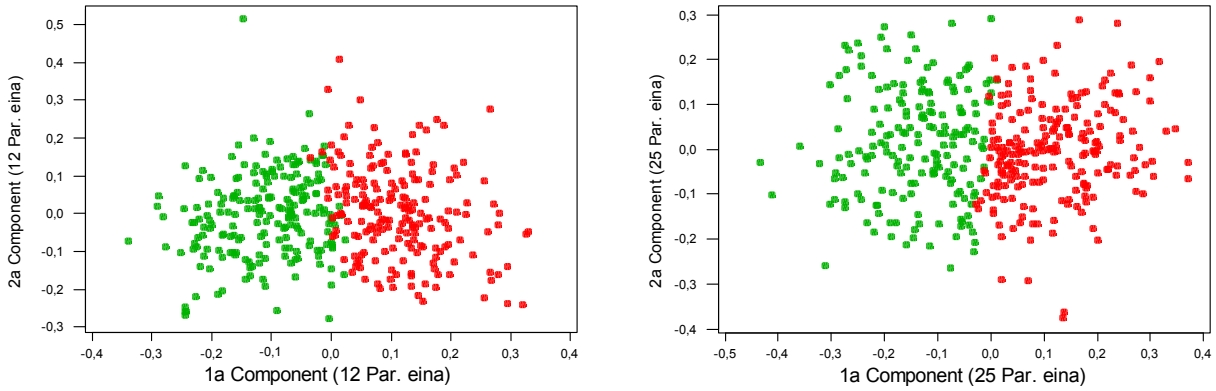

Figura 12.25: Projecció dels perfils fila en el pla format per les dues primeres components de l'anàlisi de correspondències. Els punts en vermell corresponen als blocs assignats al cluster *0*, i els punts en vermell als blocs assignats al cluster *1*, de l'anàlisi cluster basat en la distància χ^2 per 12 (esquerra) i 25 (dreta) paraules eina

12.3.6.2 Anàlisi cluster basada en la deviança

L'execució de l'algorisme cluster basat la deviança dels models de regressió politòmica porta a resultats molt semblants als obtinguts amb l'algorisme basat en la distància χ^2. Quan s'analitza la taula amb *l=12* paraules eina, hi ha quatre blocs que l'algorisme basat en la distància χ^2 agrupa en el cluster *1*, el que conté majoritàriament els capítols del final del llibre, que l'algorisme basat la deviança assigna al cluster *0*. Aquests blocs són: 97, 124, 272, 333, és a dir, que el punt de canvi s'endarrereix un bloc. Es pot observar com cap d'ells és posterior a l'estimació del punt de canvi. El guany que s'obté amb aquest mètode respecte a l'aplicació de l'algorisme basat en la distància χ^2, en termes del logaritme de la versemblança, és insignificant, i en termes de l'estadístic χ^2_B la disminució també es poc rellevant

Quan s'analitza la taula amb *l=25* paraules eina, hi ha hi ha 5 blocs posteriors al punt de canvi que l'algorisme basat en la distància χ^2classifica en el cluster *1,* el que conté majoritàriament els blocs del final del llibre, que l'algorisme basat la deviança assigna al cluster *0*. Aquests blocs són: *343, 357, 359, 374, 388.* De nou, el guany que s'obté amb aquest mètode respecte a l'aplicació de l'algorisme basat en la distància χ^2, en termes del logaritme de la versemblança, és extraordinàriament petit.

12.3.6.3 Validació dels resultats obtinguts en l'anàlisi cluster

Per validar els resultats obtinguts en l'anàlisi cluster, s'ha ajustat, tant per *l=12* com per *l=25* categories (*paraules eina*) el model politòmic:

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

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 pel mètode basat en la distància χ^2, i s'ha calculat la màxima versemblança del model, $L_c\left(\hat{\beta}^{(c)}\right)$.

La comparació dels resultats obtinguts en l'estimació del punt de canvi i en les anàlisi cluster basades tant en la distància χ^2 com en la deviança dels models politòmics es mostren el la taula 12.6. En ella es pot observar com l'anàlisi cluster millora, tot i que de

forma poc notable, els estadístics anàlegs utilitzats per a estimar el punt de canvi, pel que es pot assegurar que l'estimació del punt de canvi proposa un bona separació en grups.

Estadístic	***Nº Par. Eina***	***Punt de Canvi***	***Cluster χ^2***	***Cluster Deviança***
$L(\hat{\beta})$	*l=12*	-272033,8	-271496,4	-271485
$L(\hat{\beta})$	*l=25*	-438605	-437747	-437747

Taula 12.6: Comparació del valor màxim del logaritme de la versemblança pels models politòmics ajustats per les seqüència de *l=12* i *l=25* paraules eina i amb una variable indicadora com a explicativa, que separa els blocs en dos grups en tres situacions: separació donada pel punt de canvi i pels algorismes cluster basats en la distància χ^2 i en l'ajust de models politòmics.