CAPÍTOL 9

LLARGADA DE FRASE

Índex Capítol

9.1 Estudi de la llargada de frase per Capítol 125
9.1.1 Notació i forma de les dades 126
9.2 Estudi de la relació de la llargada de frase amb N_i 128
9.3 Anàlisi descriptiva de les dades 129
9.3.1 Gràfics de Control per la Llargada de Frase 130
9.3.2 Anàlisi de Correspondències 133
9.4 Estimació del punt de canvi 136
9.4.1 Punt de canvi en seqüències Normals 136
9.4.2 Punt de canvi en seqüència multinomial 138
9.5 Anàlisi cluster del les files de la taula de contingència 140
9.5.1 Anàlisi cluster basada en la distància χ^2 140
9.5.2 Anàlisi cluster basada en la deviança 141
9.6 Conclusions 142

CAPÍTOL 9

LLARGADA DE FRASE

Al capítol 2 de la tesi hem dedicat un apartat a llistar exemples en els que s'ha fet servir la llargada de frase per a determinar l'autoria de textos. En general ha sigut una mesura que no ha tingut èxit a l'hora de caracteritzar l'estil literari i de discriminar entre autors. Tot i així abordem l'estudi de l'homogeneïtat de l'estil a través de la llargada de frase en el *Tirant*. L'anàlisi es farà només per capítols, perquè comencen amb l'inici d'una frase i acaben amb el final d'una frase, mentre que els blocs comencen i acaben en punts que no necessàriament coincideixen amb inici o final de frase, raó per la qual es fa difícil comptabilitzar-les. Si eliminéssim aquestes porcions de frase, tallades per l'inici o final de bloc, ens quedaríem amb trossos de text de llargada no constant, pel que els blocs perdrien bona part de la seva utilitat.

9.1 Estudi de la llargada de frase per Capítol

Tal i com s'ha indicat en el capítol 4, mesurem la llargada de frase en nombre de paraules per frase, i considerem com a frase tot el que acaba en un punt, un signe d'interrogació o bé un signe d'exclamació. Observem que aquesta variable depèn de la puntuació i, per tant, queda sota el control conscient de l'autor. A més a més, en els texts medievals, com el *Tirant*, la puntuació no s'introdueix fins més tard, i és obra de l'editor. Tot i això, emprendrem la seva anàlisi per estudiar la (possible manca de) homogeneïtat en l'estil del *Tirant*.

Les frases s'han codificat en 10 categories en funció de la seva llargada. La codificació que s'ha fet servir és la següent:

Llargada Frase	nº de frases	nº de frases (N_i>200)
1-9	1192	1160
10-15	1897	1876
16-20	1657	1642
21-25	1463	1440
26-30	1306	1298
31-40	1974	1949
41-50	1276	1264
51-60	742	737
61-70	500	491
>70	712	706
	Tot =12719	**Tot =12563**

Taula 9.1: Categories en les que hem classificat les frases en funció de la seva llargada en el *Tirant*, nombre de frases en cada categoria en tot el llibre, i nombre de frases en la cada categoria en els capítols amb més de *200* paraules.

La primera columna indica les categories en que s'han subdividit les frases en funció de la seva llargada, la segona el nombre de frases en cada categoria al llarg de tot el *Tirant*, i la tercera el nombre de frases en cada categoria que seran considerades en l'estudi, en ser les que es troben en els capítols de més de 200 paraules.

Les variables que s'han obtingut de la quantificació de la llargada de frase són les següents:

- El nombre de frases en cadascuna de les 10 categories esmentades, per cadascun dels capítols de més de 200 paraules
- la llargada mitjana de frase per capítol, mesurada en paraules per frase.

9.1.1 Notació i forma de les dades

La taula 9.2 mostra el tros corresponent als capítols 352-400 de la taula que es farà servir per a l'estudi de la llargada de frase en el *Tirant*. La taula sencera té 425 files, que corresponen als capítols amb $N_i>200$, i 14 columnes: la primera indica el número de capítol, la segona la seva llargada, N_i, la tercera el nombre total de frases en el capítol f_i, les deu columnes següents formen una taula de contingència que conté el nombre d'ocurrències en cadascuna de les $l=10$ categories en que s'han agrupat les frases en funció de la seva llargada, per cada capítol, i la darrera columna conté la llargada mitjana de frase en el capítol:

$$\overline{llf}_i = \frac{N_i}{f_i}$$

Per exemple, a la taula es pot llegir com en el bloc 355 hi ha un total de 45 frases, de les quals 6 tenen una llargada entre 1 i 9 paraules, 3 frases tenen una llargada entre 10 i 15 paraules, ... i 3 frases tenen una llargada superior a les 70 paraules. La llargada mitjana de frase en el capítol és de 34,58 paraules.

Anomenem *llf* a la variable aleatòria llargada de una frase, i y_{ji} al nombre de frases en la categroria *j*-èssima, per $j=1,2,...,10$, en el capítol i-èssim. $\mathbf{y_i}=(y_{1i},y_{2i,...},y_{10i})$ es distribueix

Mult(N_i, $\boldsymbol{\pi_i}$), on $\boldsymbol{\pi_i}=(\pi_{1i},\pi_{2i},...,\pi_{10i})$ és el vector tal π_{ji} és la probabilitat que una frase presa a l'atzar en el capítol i-èssim pertanyi a la categoria *j*-èssima.

Cap.	N_i	f_i	**1-9**	**10-15**	**16-20**	**21-25**	**26-30**	**31-40**	**41-50**	**51-60**	**61-70**	**+70**	$\overline{llf}$
352	217	7	1	0	2	0	1	0	2	1	0	0	31,00
353	738	26	4	2	4	4	3	3	3	0	1	2	28,38
354	852	23	2	3	2	0	1	4	6	2	2	1	37,04
355	1556	45	6	3	6	1	7	7	6	2	4	3	34,58
356	428	9	1	0	1	1	0	1	2	0	0	3	47,56
357	1002	27	5	1	2	0	4	6	2	2	2	3	37,11
358	279	11	2	0	3	1	1	2	1	1	0	0	25,36
359	311	8	0	1	0	1	2	0	2	1	1	0	38,88
360	416	13	1	2	2	2	1	0	2	1	2	0	32,00
362	446	12	1	0	3	3	0	0	2	1	1	1	37,17
363	244	9	2	2	0	1	1	2	0	0	0	1	27,11
364	349	9	0	1	0	2	2	1	0	1	1	1	38,78
365	261	7	1	0	0	1	0	3	1	0	0	1	37,29
366	579	21	2	2	5	2	3	2	2	3	0	0	27,57
367	579	21	2	2	5	2	3	2	2	3	0	0	27,57
368	331	8	0	0	1	0	1	3	1	1	0	1	41,38
369	463	13	0	1	2	2	2	2	2	0	1	1	35,62
370	505	12	0	2	1	0	0	2	3	1	2	1	42,08
371	750	20	3	2	1	2	2	3	3	1	0	3	37,50
372	892	38	4	7	7	7	4	4	5	0	0	0	23,47
373	388	9	0	2	0	0	2	0	0	3	1	1	43,11
374	1441	39	0	4	3	7	3	10	4	3	3	2	36,95
375	357	14	0	0	3	7	1	1	2	0	0	0	25,50
376	752	22	1	3	0	3	3	7	1	2	1	1	34,18
377	303	7	0	2	0	0	0	1	1	1	1	1	43,29
378	523	20	3	3	1	6	2	2	1	1	0	1	26,15
380	414	12	0	2	0	0	3	4	1	2	0	0	34,50
381	272	7	0	0	1	1	1	0	3	0	1	0	38,86
382	330	11	0	4	0	2	1	0	3	0	1	0	30,00
383	309	10	0	3	3	0	1	1	1	0	0	1	30,90
384	512	12	0	1	1	0	1	1	6	1	0	1	42,67
385	287	5	0	1	0	1	0	0	1	0	1	1	57,40
386	588	19	0	2	5	4	3	2	1	0	1	1	30,95
387	1639	42	1	1	5	3	8	8	6	4	3	3	39,02
388	357	13	1	2	1	3	1	3	1	0	1	0	27,46
390	454	9	1	0	0	1	1	1	0	2	1	2	50,44
391	333	10	0	3	2	1	0	0	1	1	2	0	33,30
392	287	8	0	0	2	0	1	2	1	1	1	0	35,88
393	327	8	0	1	0	0	3	1	1	0	0	2	40,88
394	1059	25	0	1	1	3	3	5	5	4	1	2	42,36
395	257	7	0	1	0	0	0	4	1	0	1	0	36,71
397	356	6	0	2	0	1	0	0	0	0	0	3	59,33
399	203	4	0	2	0	0	0	0	0	0	0	2	50,75
400	362	13	0	5	1	3	2	0	1	0	0	1	27,85

Taula 9.2: Part corresponent als capítols 352-400 de les dades utilitzades en l'estudi de llargada de frase. Cada fila correspon a un capítol, la primera columna dóna el número de capítol, la segona la llargada del capítol, N_i, la tercera el nombre de frases del capítol f_i, les deu següents el nombre frases en cadascuna de les deu categories en que s'ha classificat segons la llargada, i la darrera, la llargada mitjana de frase.

S'han realitzat quatre tipus d'anàlisi de l'homogeneïtat d'estil en el *Tirant* fent servir la llargada de frase: en primer lloc s'estudia la relació entre la llargada de frase i la llargada de capítol, a continuació es fa una anàlisi descriptiva de les dades on s'analitza l'evolució temporal de la llargada de frase i una anàlisi de correspondències on les files de la taula de contingència són els capítols de més de 200 paraules i les columnes són les deu categories en que es classifiquen les frases en funció de la seva llargada, en el tercer estudi s'estima un possible punt de canvi en les seqüències univariants i multinomials, i per últim es fa una anàlisi cluster basada en la distància de χ^2.

9.2 Estudi de la relació de la llargada de frase amb N_i

El valor esperat de la llargada mitjana de frase no depèn de la llargada del capítol, N_i, mentre que la seva variança és proporcional a $1/N_i$, és a dir és més gran com més curt és el capítol. Aquest comportament es pot observar en la figura 9.1.

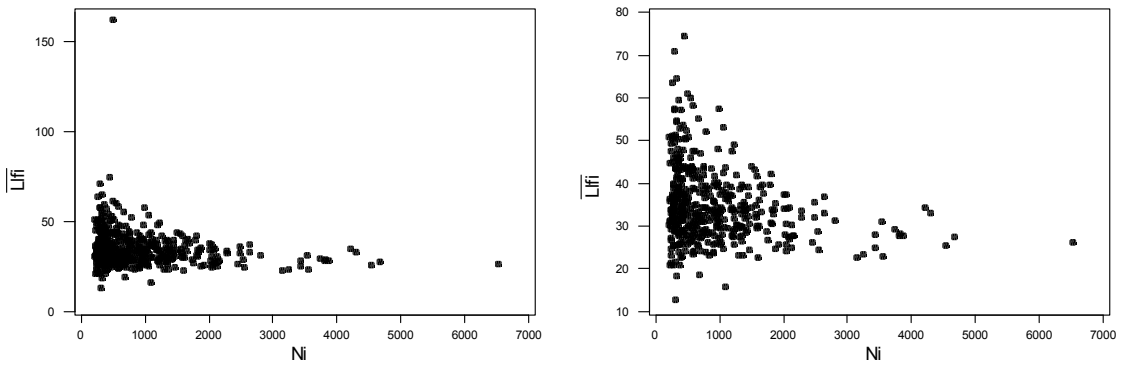

Figura 9.1: Relació entre la llargada mitjana de frase i la llargada de capítol N_i. En el gràfic de l'esquerra es pot observar una clara anomalia que en el de la dreta ha sigut eliminada.

La relació entre el nombre de frases, o el nombre de frases d'una determinada llargada (categoria), i la llargada del capítol, N_i, és força lineal, amb gran correlació positiva. Per exemple, la correlació entre N_i i el nombre de frases és del 94,7%, i entre N_i i el nombre de frases de llargada compresa entre les 31 i les 40 paraules és del 90,7%. En la figura 9.2 es poden observa aquestes dues relacions.

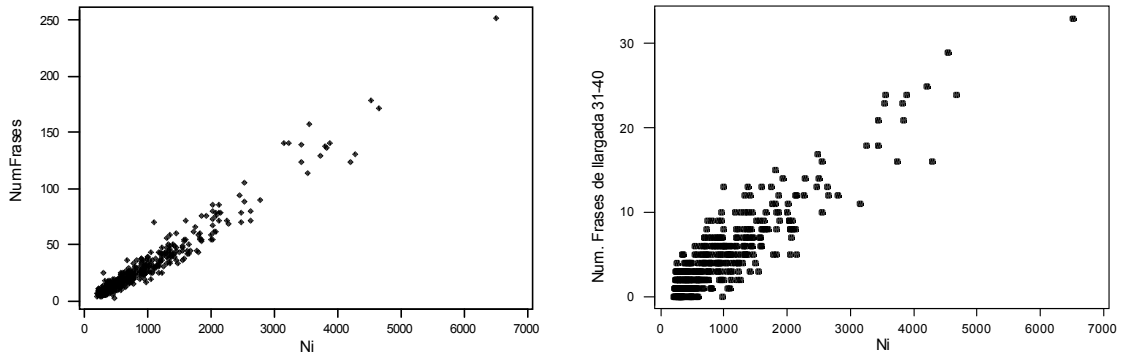

Figura 9.2: Relació entre el nombre de frases i la llargada de capítol N_i (esquerra) i entre el nombre de frases de llargada entre 31 i 40 paraules i la llargada de capítol (dreta).

Tant el valor esperat com la variança de la llargada mitjana de frase com del nombre de frases en cada categoria depenen del capítol, *i*.

9.3 Anàlisi descriptiva de les dades

L'histograma de la llargada de frase mostra una distribució no simètrica, amb la cua de la dreta clarament més llarga. Les frases, de fet, poden tenir una llargada que va des d'una paraula fins a tan llargues com vulguem.

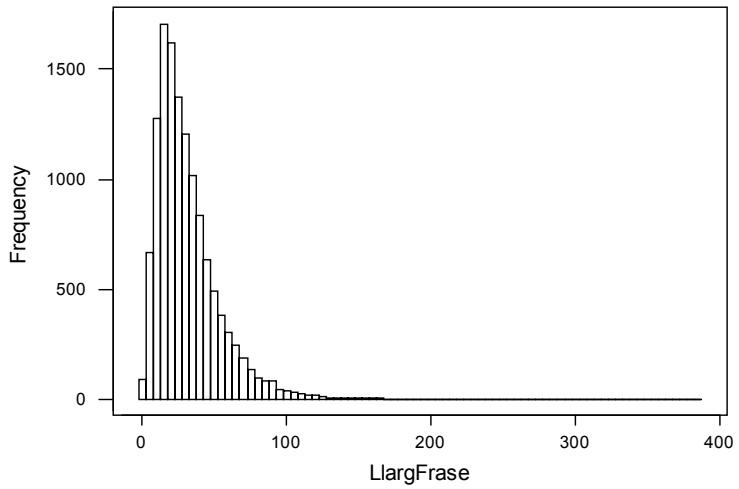

Figura 9.3: Histograma de la llargada de frase per a totes les frases del *Tirant.*

L'histograma de la llargada mitjana de frase pels capítols de més de 200 paraules mostra una claríssima anomalia en el capítol 15, un capítol relativament curt (485 paraules), amb només tres frases, dues de les quals es troben en la categoria de més de 70 paraules, i una d'elles és la més llarga de tot el Tirant, amb 384 paraules, molt més llarga que la següent de 256 paraules. D'aquesta manera la llargada mitjana de frase d'aquest capítol és de 161,67 paraules, mentre que la mitjana i desviació tipus de totes les altres llargades mitjanes de frase valen, respectivament, 34,725 i 8,72 paraules. Aquesta anomalia coincideix amb la que s'havia detectat en l'estudi de la relació entre la llargada de frase i la llargada de capítol.

Com ja s'ha esmentat en el capítol 2 de la tesi, Williams (1940) va atribuir a la distribució de freqüències dels logaritmes del nombre de paraules per frase una aproximació a una Normal per a cada autor. Holmes (1995), però, assegura que el model lognormal no funciona, ja que la majoria de les distribucions de freqüència de la llargada de frase, després de la transformació logarítmica, són negativament asimètriques. Observem que el comportament descrit per Holmes (1995) es dóna per la distribució de llargades de frase en el *Tirant*, tal i com pot observar-se en el gràfic de la figura 9.4 i en els test de Normalitat de la figura 9.5.

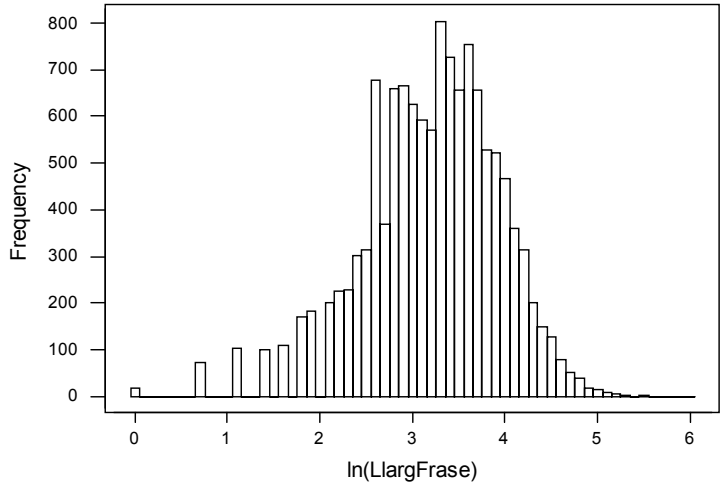

Figura 9.4: Histograma del logaritme de la llargada de Frase, per a totes les frases del *Tirant*.

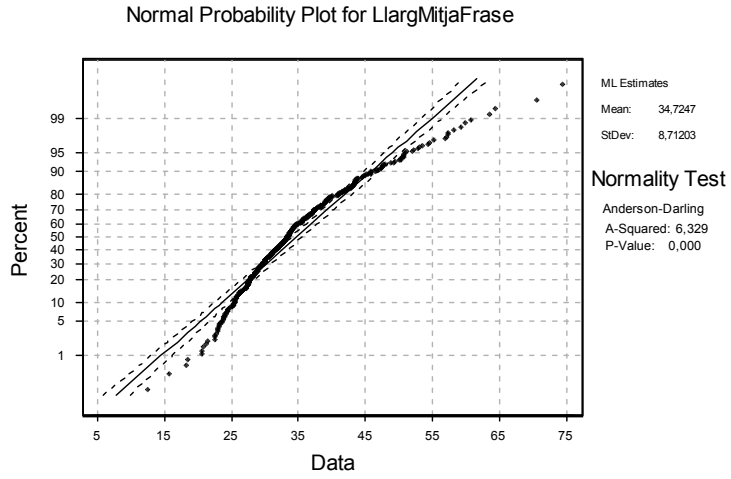

Figura 9.5: Test de Normalitat per al logaritme de la llargada de frase, per a totes les frases del *Tirant*. Es pot rebutjar la hipòtesi de Normalitat.

9.3.1 Gràfics de Control per la Llargada de Frase

S'ha representat gràficament, a la figura 9.6, l'evolució de la llargada mitjana de frase al llarg del *Tirant*, L'anàlisi gràfica no mostra cap característica rellevant, i no s'observa cap canvi en el nivell o en la variabilitat, en la seqüència de llargades mitjanes de frase que senyali una possible manca d'homogeneïtat. Com ja van observar Mosteller i Wallace (1984) entre d'altres, la llargada de frase no sembla ser una bona mesura per a discriminar entre estils literaris.

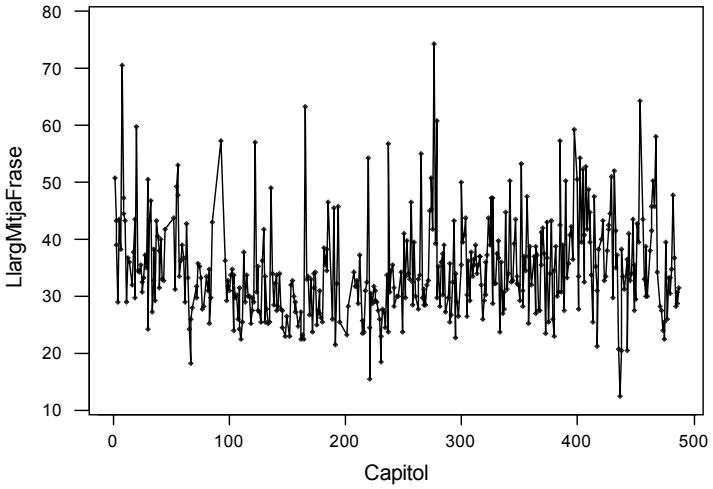

Figura 9.6: Evolució de la llargada mitjana de frase pels capítols de més de 200 paraules al llarg del *Tirant*. El capítol 15, que suposa una anomalia, ha sigut eliminat del gràfic.

Anomenem y_{ji} i $\hat{\pi}_{ji}$ al nombre i a la proporció de frases en la categoria j-èssima del capítol i-èssim, amb:

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

La figura 9.7 conté l'evolució, al llarg del *Tirant*, de les proporcions de frases en cadascuna de les deu categories. No s'observa en cap gràfic la presència d'un canvi en el nivell. El que es veu és que, a partir d'aproximadament el capítol 250, augmenta en la majoria d'ells la proporció de capítols amb *0* frases en cada categoria. Això pot ser degut a que també augmenta la proporció de capítols curts i, per tant, amb poques frases.

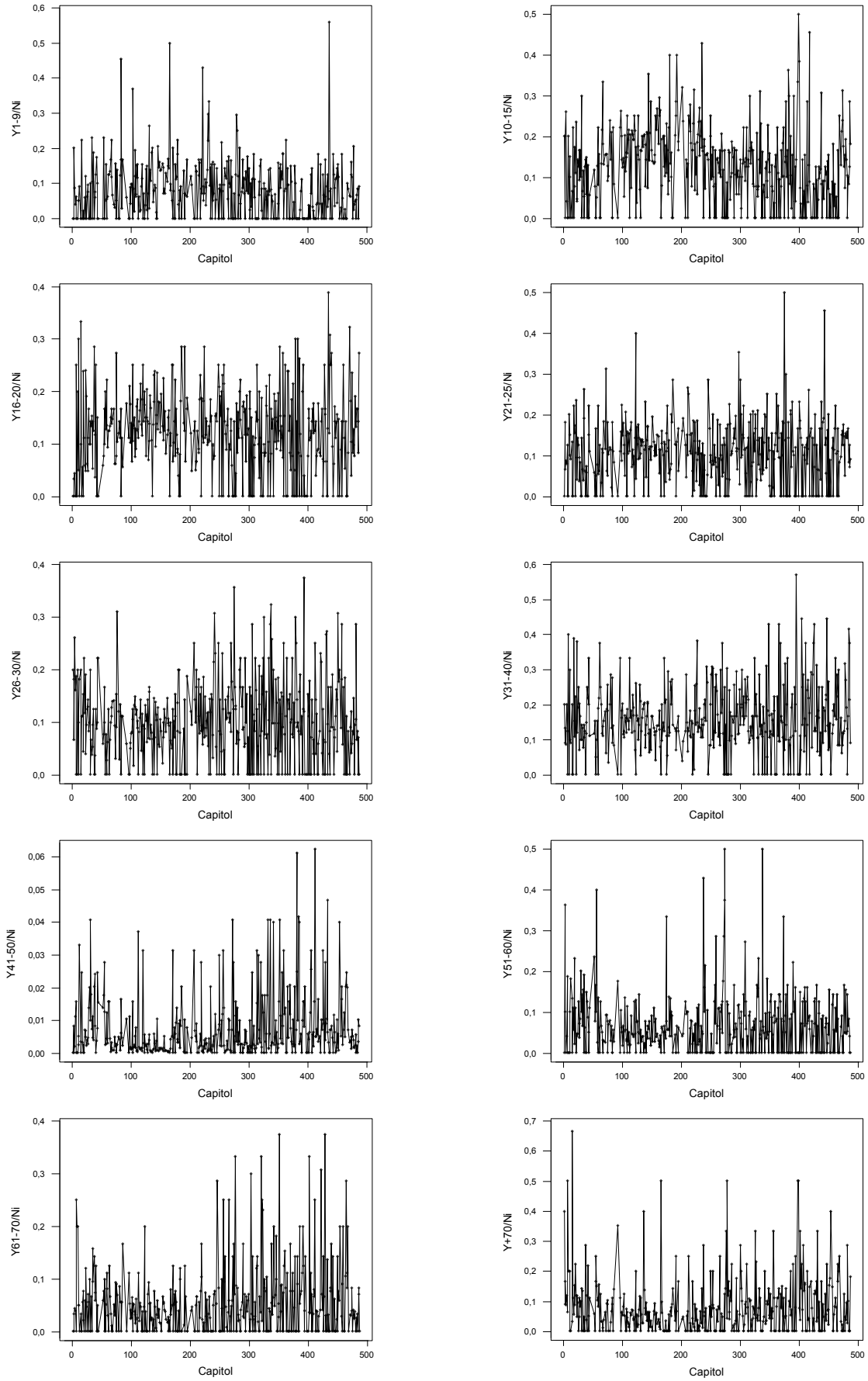

Figura 9.7: Evolució al llarg del *Tirant* de la proporció de frases en les 10 categories en que s'han classificat en funció de la llargada.

La seqüència del vector multinomial de nombres de frases en cadascuna de les deu categories, $\boldsymbol{y_i}=(y_{1i},y_{2i},y_{3i},...,y_{10i})$, es pot representar en un gràfic de control Chi-quadrat, com el de la figura 9.8, que representa l'evolució temporal de l'estadístic:

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

amb $l=10$ categories. Prenem com a valors esperats de les proporcions, $\hat{\pi}_j$, les proporcions promig de frases en la categoria j-èssima, per $j=1,2,...,10$:

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

amb $n=425$ capítols de més de 200 paraules. El gràfic, que monitoritza les distàncies Chi-quadrat entre els perfils fila respecte del perfil fila promig, tampoc mostra un canvi clar en el nivell que pugui ser interpretat com un canvi en l'estil.

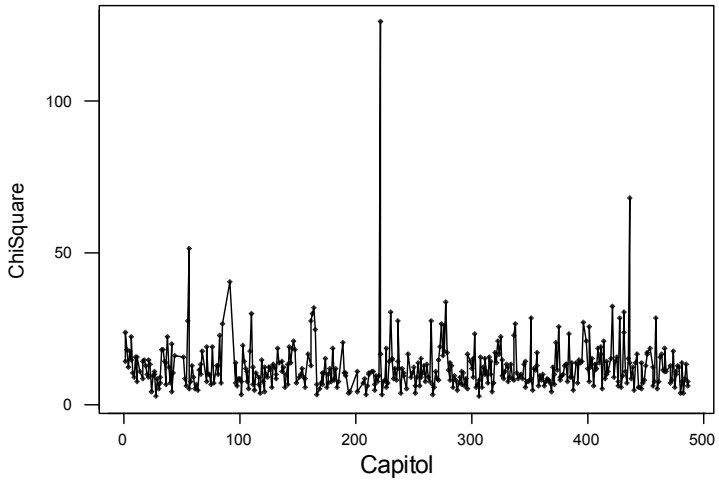

Figura 9.8: Gràfic Chi-quadrat per a la seqüència multinomial formada pel nombre de frases en cadascuna de les $l=10$ categories en que s'han subdividit en funció de la llargada.

9.3.2 Anàlisi de Correspondències

Disposem de les distribucions de llargada de paraula tal i com apareixen a la taula de contingència 9.2. Els detalls sobre l'anàlisi de correspondències s'han descrit en el capítol 5 de la tesi.

El gràfic simètric per columnes de la figura 9.9 mostra com la primera component representa la llargada de frase: les categories estan ordenades per llargada, les frases curtes es troben a l'esquerra de l'eix i les llargues en la part dreta. La primera component representa el 24% de la inèrcia total i la segona el 14,3%.

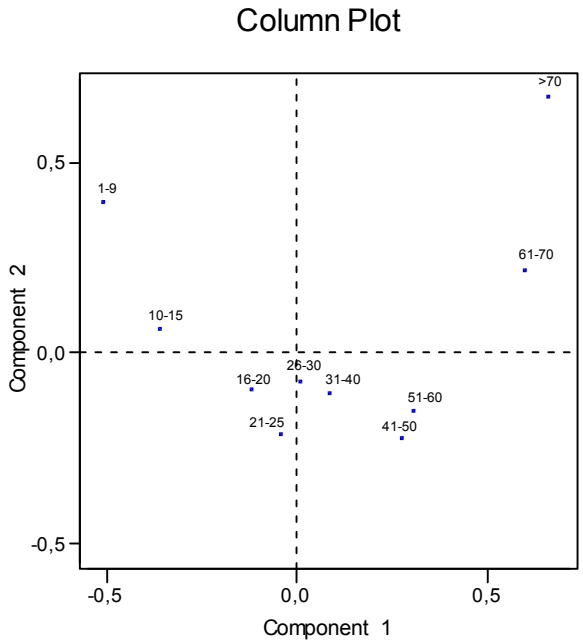

Figura 9.9: Gràfic simètric per columnes per a l'anàlisi de correspondències de la llargada de frase pels capítols

El gràfic 9.10 mostra la forta correlació lineal entre la llargada mitjana de frase i el valor de la primera component de l'anàlisi de correspondències.

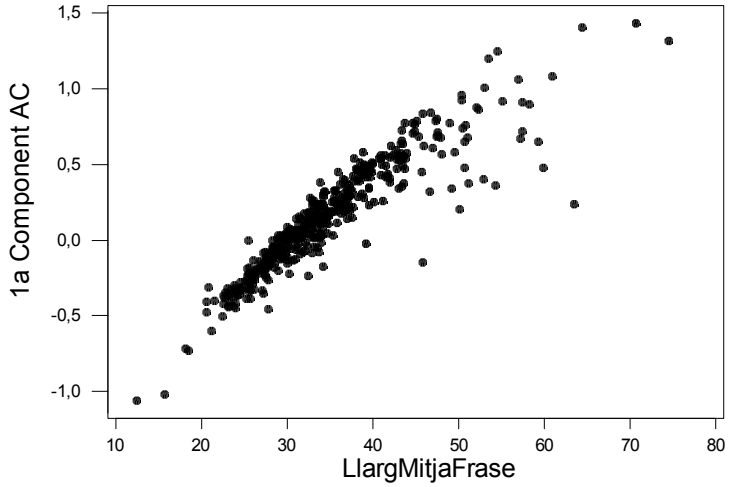

Figura 9.10: Relació entre la llargada mitjana de frase i el valor de la primera component de l'anàlisi de correspondències.

El gràfic asimètric per files de la figura 9.11 és de difícil interpretació, en haver-hi representats 425 punts molt concentrats. Els capítols que apareixen amb valor més gran de la primera component són aquells on es troben proporcions més elevades de frases llargues i proporcions menors de frases curtes, els que queden prop del centre són els que tenen un perfil més semblant al perfil promig, mentre que els que són negatius amb un valor absolut més elevat tenen proporcions més elevades de frases curtes i proporcions més petites de frases llargues.

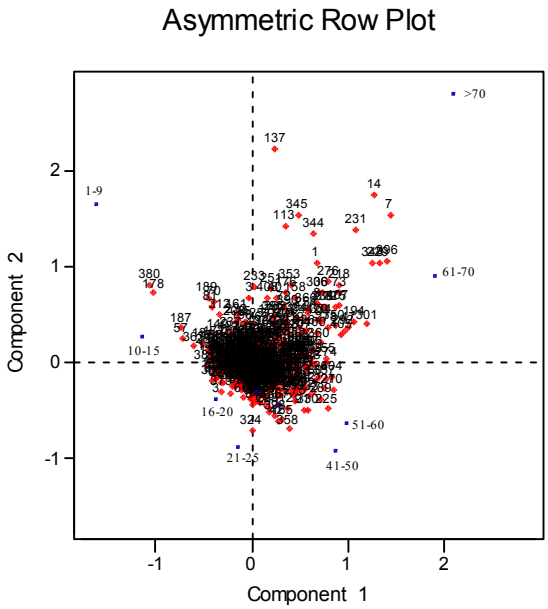

Figura 9.11: Gràfic asimètric per files de l'anàlisi de correspondències de la llargada de frase en els capítols de més de 200 paraules del *Tirant*.

El valor esperat de la primera component de l'anàlisi de correspondències no depèn de la llargada del capítol N_i, però, com en el cas de la llargada mitjana de paraula, la seva variança sí que en depèn, tal i com es pot veure en la figura 9.12.

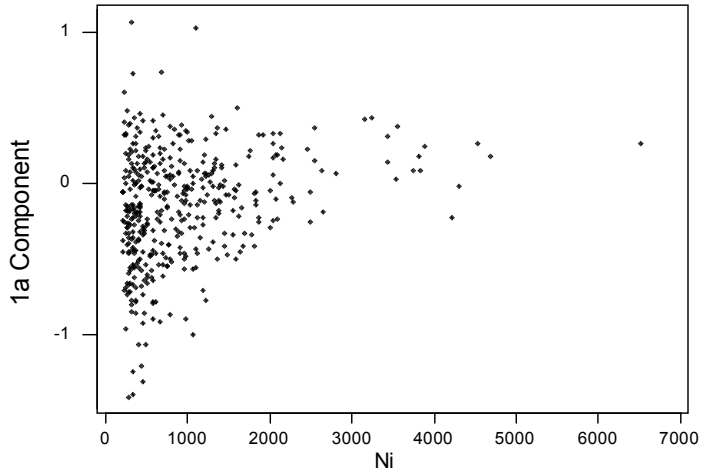

Figura 9.12: Relació entre la primera component de l'Anàlisi de Correspondències per la llargada de frase amb la llargada de capítol.

L'evolució temporal de la primera component, representada en el gràfic de la figura 9.13, mostra una lleugera tendència a augmentar fins al capítol 236, després s'estabilitza fins al 300, després la tendència és a disminuir fins al 399 i a partir d'aquí la tendència torna a ser a créixer fins al final.

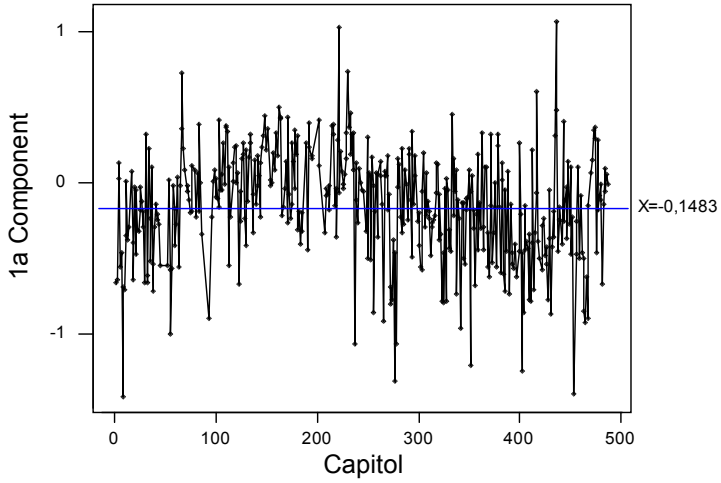

Figura 9.13: Evolució, al llarg del *Tirant*, del valor de la 1ª Component de l'anàlisi de correspondències per a la llargada de frase.

9.4 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 frase, del seu logaritme, i del valor de la primera component de l'anàlisi de correspondències. A continuació, s'estima el punt de canvi en la seqüència de dades multinomials contingudes en la taula 9.2 via l'ajust de models politòmics basat en la deviança.

9.4.1 Punt de canvi en seqüències Normals

S'ha estimat el punt de canvi per la llargada mitjana de frase fent servir les tècniques proposades per a dades Normals descrites a la secció 6.2. Tot i que la llargada de frase és una variable discreta, les hipòtesis en que es sustenten els models lineals s'acompleixen de forma aproximada gràcies al teorema central del límit. Si diem $\overline{llf}_i$ a la llargada mitjana de frase per al capítol i-èssim, assumint independència entre les observacions, s'han ajustat els *n-1* models:

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

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,...,425*. Estimem el punt de canvi com**:**

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

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

El gràfic de la figura 9.14 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 mitjanes de frase és $\hat{r}_N = 263$. Observem com el valor de F_r creix fins a $r=236$, a partir d'aquest capítol fins al *383* hi ha una zona en la que F_r té una lleugera tendència a decréixer, passant per una sèrie de màxims locals i del màxim global per, a partir del capítol *384,* decréixer de forma molt ràpida.

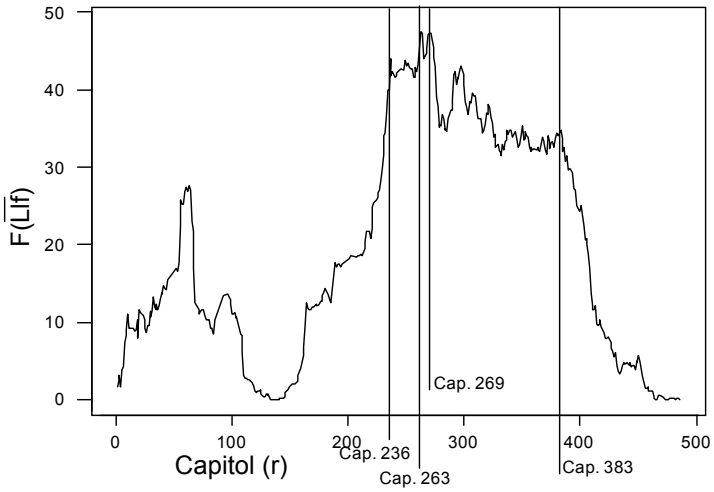

Figura 9.14: Evolució de F_r en funció del capítol r. El màxim de F_r és $r=263$.

S'obté un gràfic molt semblant i la mateixa estimació del punt de canvi fent servir el mateix procediment per la seqüència de $log(\overline{llf}_i)$.

Per la seqüència amb els valors de la primera component de l'anàlisi de correspondències, estimem el punt de canvi a $\hat{r}_N = 269$. El gràfic de la figura 9.15 mostra l'evolució de l'estadístic F_r en funció de r per a la primera component de l'anàlisi de correspondències.

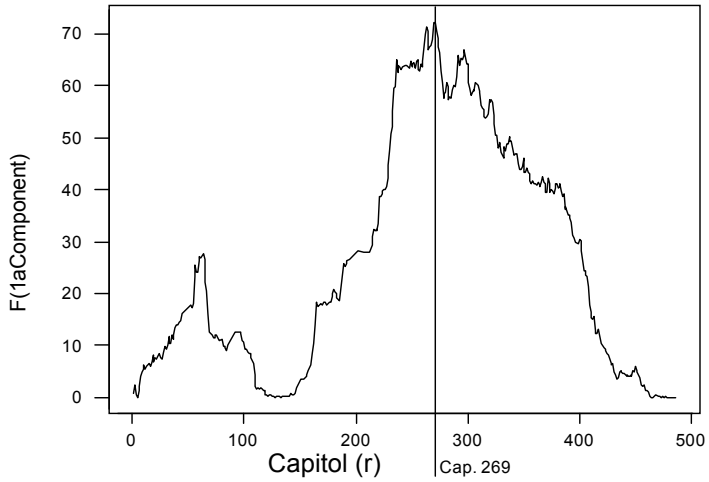

Figura 9.15: Evolució de F_r en funció de r per al valor de la primera component de l'anàlisi de correspondències obtingudes de l'anàlisi de la llargada de frase pels capítols de més de 200 paraules de *Tirant*.

9.4.2 Punt de canvi en seqüència multinomial

Model politòmic

Recordem que y_{ji} és el nombre de frases en la categoria j-èssima que comptem en el capítol i-èssim, i que el vector amb el nombre d'ocurrències en cada categoria, $\boldsymbol{y_i}=(y_{1i},y_{2i,...},y_{10i})$, es distribueix $Mult(N_i,\boldsymbol{\pi_i})$, on $\boldsymbol{\pi_i}=(\pi_{1i},\pi_{2i},...,\pi_{10i})$ és el vector de probabilitats. Obtenim l'estimació del punt de canvi ajustant models de regressió politòmica, seguint la metodologia exposada en el capítol 6.

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 $j=1,...,10$, 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 $j=2,...,l=10$, 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 9.16 mostra l'evolució de $L\left(r,\hat{\beta}^{(r)}\right)$ en funció de r. El màxim, i per tant l'estimació del punt de canvi, és $\hat{r}_M=269$ Observem que hi ha un una zona, que va del capítol 236 al 300, en la que el valor del màxim del logaritme de la versemblança és manté estable i on el gràfic és força pla.

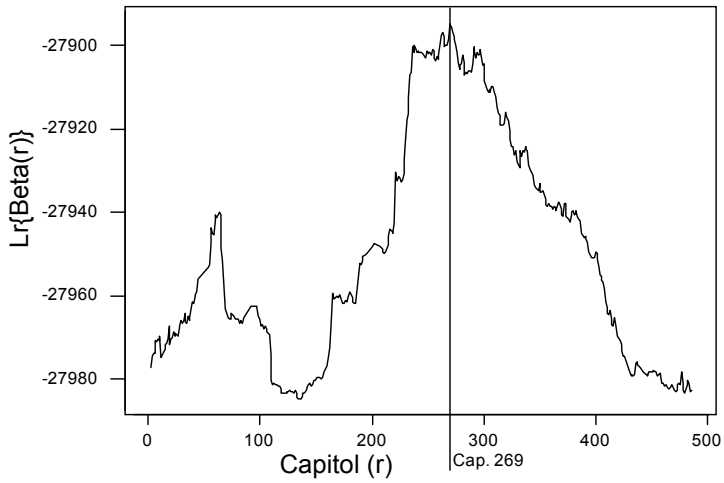

Figura 9.16: Evolució de $L\left(r,\hat{\beta}^{(r)}\right)$ en funció de r. El màxim és r=269, tot i que s'observa una zona força plana entre els capítol 236 i 300.

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 frases en la categoria *j*, per *j=1,2,...,10*. 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. Les tres estimacions del punt de canvi que s'obtenen no coincideixen, tot i que són bastant semblants: $\hat{r}_{ML1}$=263 i $\hat{r}_{ML2}=\hat{r}_{ML3}$=269.

Conclusions

La conclusió d'aquesta anàlisi és que l'existència d'un punt de canvi en l'estil, quan aquest es quantifica mitjançant la llargada de frase, no és molt clara. La millor estimació és $\hat{r}$=269, que no coincideixen amb els que s'havien trobat en l'estudi de la llargada de paraula.

Es fa difícil caracteritzar els capítols anteriors i posteriors al 269, pel que fa a la llargada de frase. No s'observen proporcions més grans o més petites de frases llargues o curtes en cap de les dues subseqüències en que queda dividit el *Tirant* pel capítol 269.

La taula 9.3 mostra un quadre resum amb les estimacions del punt de canvi obtingudes en l'estudi de la llargada de frase pels capítols de més de 200 paraules.

Distribució	**Seqüència**	**Punt de canvi**
Normal	*Llargada mitjana*	$\hat{r}_N$=263
	log(Llargada mitjana)	$\hat{r}_N$=263
	1ª Comp. A. Corresp.	$\hat{r}_N$=269
Multinomial	*Combinació de binom.* $\max_{1\le j\le 10} L_j\left(r,\hat{\beta}^{(r)}\right)$	$\hat{r}_{ML1}$=263
	Combinació de binom. $\sum_j L_j\left(r,\hat{\beta}^{(r)}\right)$	$\hat{r}_{ML2}$=269
	Combinació de binom. $\sum_j L_j^2\left(r,\hat{\beta}^{(r)}\right)$	$\hat{r}_{ML3}$=269
	Model Politòmic	$\hat{r}_M$=269

Taula 9.3: Quadre resum de les estimacions del punt de canvi obtingudes en l'anàlisi de la llargada de frase per als capítols de més de 200 paraules

9.5 Anàlisi cluster del les files de la taula de contingència

Es faran servir les tècniques exposades en el capítol 7 d'anàlisi cluster, basades en dos criteris d'ajust de models per dades politòmiques per agrupar els capítols en dos clusters, en funció de la distribució de la llargada de frase. El punt de partida de l'anàlisi és la taula de contingència 9.2, on les files són els 425 capítols de més de *200* paraules i les deu columnes representen els nombres d'ocurrències en cadascuna de les categories en que s'han classificat les frases en funció de la seva llargada.

9.5.1 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 agrupacions diferents. Les dues més freqüents, cadascuna de les quals representa entre el 40 i el 45% de les execucions de l'algorisme, porten a dos clusters de 242 i 183 capítols i 254 i 171 capítols, respectivament. Per aquesta darrera classificació és té el màxim de χ^2_B, però la diferència amb el valor de l'estadístic χ^2_B per l'altra solució és insignificant. Per a les dues classificacions coincideixen les assignacions finals de 242 i 171 capítols als dos clusters, essent diferent la classificació dels altres 12 capítols, que anomenarem de dubtosa classificació. Aquest capítols són de llargades variables, des de curts de poc més de 200 paraules a força llargs, de gairebé 2.500 paraules.

En l'agrupació en clusters, tal i com havia passat per la llargada de paraula, els capítols han estat assignats a un dels dos grups en funció del valor de la primera component de l'Anàlisi de Correspondències, i els capítols de dubtosa classificació tenen valors que els situen en la frontera. En la figura 9.17 observem aquest comportament.

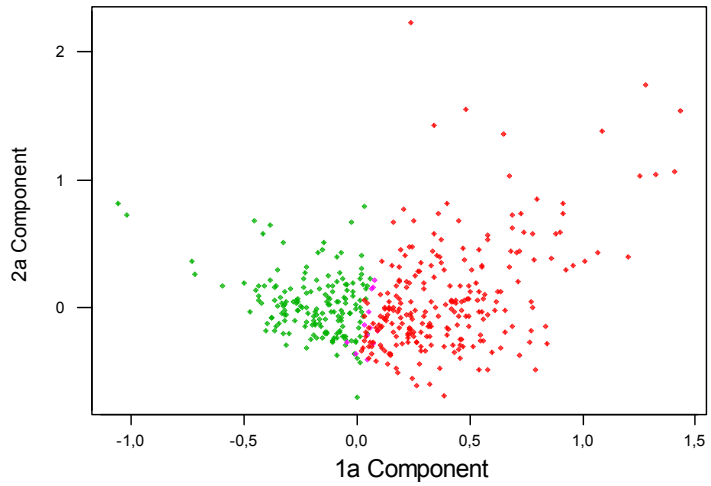

Figura 9.17: Gràfic de les files (capítols) en les dues primeres components de l'anàlisi de correspondències per la llargada de frase. El color verd correspon als capítols que han sigut assignats al cluster que conté majoritàriament els capítols del principi del llibre, mentre que els verds són aquells que han sigut assignats a l'altre cluster. En magenta hi tenim els de classificació dubtosa, aquells capítols que han estat col·locats en clusters diferents en les dues millors agrupacions en obtingudes de l'anàlisi cluster.

El gràfic de l'esquerra de la figura 9.18 mostra la relació entre la llargada mitjana de frase i la primera component de l'anàlisi de correspondències.

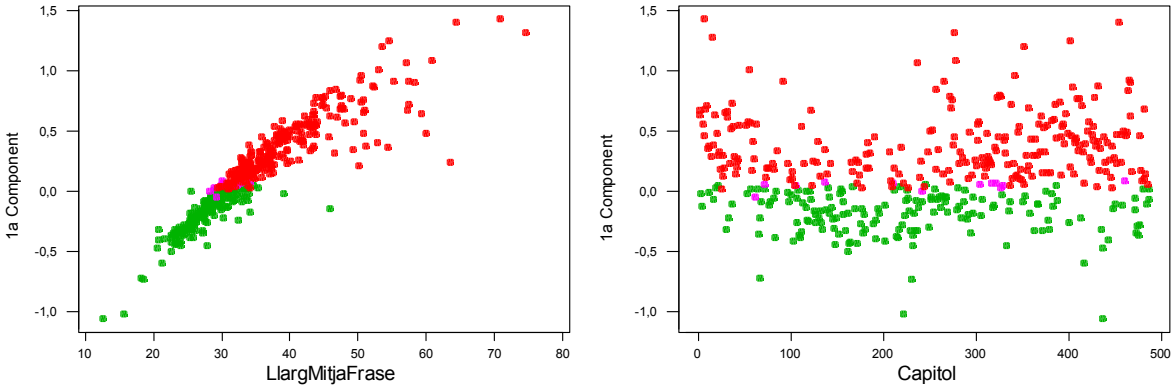

Figura 9.18: Gràfic de la relació de la primera component de l'anàlisi de correspondències amb la llargada mitjana de frase (esquerra), i de la seqüència temporal del valor de la primera component per la llargada de frase al llarg del *Tirant* (dreta). El color verd correspon als capítols que pertanyen al cluster que conté majoritàriament els capítols del principi del llibre, mentre que els verds són aquells que pertanyen a l'altre cluster. En magenta hi tenim els capítols de classificació dubtosa, aquells que han estat col·locats en clusters diferents en les dues millors agrupacions en obtingudes de l'anàlisi cluster.

En el gràfic de la dreta de la figura 9.18 s'observa com en la última part del llibre els punts vermells, que corresponen al capítols assignats al C*luster 1*, són més abundants, tot i que no s'observa una frontera clarament definida.

9.5.2 Anàlisi cluster basada en la deviança

El mètode per agrupar les files d'una taula de contingència en dos clusters, basat en la deviança de models politòmics, és el proposat en el capítol 7 de la tesi.

L'execució de l'algorisme porta a resultats molt semblants als obtinguts amb l'algorisme basat en la distància χ^2. Hi ha quatre capítols, el 72, 211, 375 i 386, que l'algorisme basat en la deviança classifica en el cluster que conté majoritàriament els capítols del final del llibre, mentre que les dues solucions millors per l'algorisme basat en la distància χ^2 classifiquen en el cluster que conté majoritàriament els capítols del principi del llibre. Per altres quatre capítols la situació és inversa: l'algorisme basat en la deviança classifica en el cluster que conté majoritàriament els capítols del principi del llibre que les dues solucions millors per l'algorisme basat en la distància χ^2 classifiquen en el que conté majoritàriament els capítols dels principi. Aquests capítols són: 104, 165, 209 i 468. Dels capítols que en l'apartat anterior havíem considerat com de dubtosa classificació, n'hi ha tres que l'algorisme basat en la deviança classifica en el cluster que conté majoritàriament els capítols del principi del llibre i n'hi ha nou que agrupa amb els del final.

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.

9.6 Conclusions

La conclusió de l'anàlisi de l'homogeneïtat d'estil al llarg del *Tirant* mitjançant la llargada de frase és que no es veu una frontera clara i que aquesta unitat d'estadística textual no és bona per a discriminar entre estils literaris diferents.