CAPÍTOL 4

DESCRIPCIÓ DE LA BASE DE DADES

Índex Capítol

CAPÍTOL 4

DESCRIPCIÓ DE LA BASE DE DADES

L'edició del *Tirant lo Blanc* que s'ha emprat en aquest estudi és la de Martí de Riquer (1983) de Edicions 62, dins de la col·lecció MOLC. L'edició reprodueix fidelment i íntegra la primera del *Tirant lo Blanc*, publicada a València per Nicolau Spindeler l'any 1490. La grafia ha estat regularitzada segons la normativa actual, però pel que fa al vocalisme àton l'editor ha mantingut fidelitat a l'original en alguns casos com *llauger, resplandor, sancer, veixella*, que el text presenta d'una forma sistemàtica. L'editor ha normalitzat també la puntuació i l'ús de majúscules. La versió digital del *Tirant* que s'ha fet servir és la que ha editat l'Institut Joan Lluís Vives dins de la biblioteca virtual, que es podia trobar a *http://www.lluisvives.com* i que correspon a l'edició en paper de Martí de Riquer (1983). Darrerament la versió digital ha estat retirada de la web, atès que estava basada en una edició que encara té drets d'autor

Aquesta edició del *Tirant* està dividida en cinc parts, basades en criteris argumentals. En la primera, del primer capítol al 97, l'acció es desenvolupa a Anglaterra, i Tirant es mostra com un excel·lent lluitador en batalles individuals i cortesanes. La segona part, que comprèn els capítols del 98 al 114, transcorre a Sicília i l'illa de Rodes, i Tirant es comporta com un gran almirall. A la tercera, que abraça els capítols entre el 98 i el 296, situada a Costantinoble i a l'Imperi de Grècia, o Bizantí, Tirant, cap de grans exercits de terra, lluita victoriosament contra moros i turcs, i apareix com un gran general. L'acció va acompanyada dels seus amors amb Carmesina, princesa de l'Imperi. A la quarta part, del capítol 297 al 407, Tirant, que ha naufragat en les costes de Tunis, es transforma en un cabdill cristià de forces nordafricanes i, per una banda, assoleix la conversió de extensos regnes moros i, per l'altra, domina i sotmet els qui persisteixen infidels. I a la cinquena, del capítol 408 al final, Tirant torna a Grècia, l'allibera totalment del perill turc i quan ha aconseguit la dignitat de Cèsar de l'Imperi i s'ha formalitzat el seu matrimoni amb la princesa Carmesina, hereva de la corona, mor a Andrianòpolis d'una pulmonia. Tot seguit moren també l'Emperador i Carmesina, i l'Imperi de Grècia, lliure completament de setges i d'atacs passa a l'Emperadriu viuda, que es casa amb Hipòlit, antic servent de Tirant.

4.1 Quantificació de l'estil del *Tirant* : passar del text a xifres

En la tesi s'ha quantificat l'estil a dos nivells, en funció del grau de complexitat a caracteritzar. En el nivell més simple es compten les freqüències d'ús d'unitats lingüístiques fàcils d'identificar i comptar. En un segon nivell s'ha caracteritzat la riquesa i diversitat de tot el vocabulari, tot fent un inventari de totes les paraules (tipus) emprades i comptabilitzant el nombre d'ocurrències de cada tipus.

4.1.1 Quantificació de l'estil

A l'hora de quantificar l'estil literari en el *Tirant*, hem eliminat del text totes les paraules que hi apareixen en cursiva, per tal de no considerar els títols dels capítols, ni les lletres de batalla ni les cites en llatí.

Per a la identificació de paraules, hem optat per considerar com a dues paraules diferents aquelles que tenen grafia diferent, és a dir, *home* i *homes* serien dues formes, i *menjo, menges, menja, mengem, mengeu, mengen* són formes diferents. També considerem com formes diferents tot el que va separat per guionets o apòstrofs. D'aquesta manera hem comptabilitzat *l'honor* com a dues paraules (*l* i *honor*) i *dix-li* com a dues paraules (*dix* i *li*),

Per a cada capítol tenim la llargada de totes les paraules, mesurada en nombre de lletres per paraula, la llargada de frase i la llargada de capítol mesurades en nombre de paraules, que utilitzarem, respectivament, en els capítols 8, 9 i 10. En particular, compararem els capítols a través de les distribucions de llargades, és a dir, mitjançant la proporció de paraules/frases/capítols de cadascuna de les llargades.

Per comptabilitzar frases hem considerat com a frase tot el que acaba en un punt, un signe d'interrogació o bé un signe d'exclamació. Com ja hem observat en el capítol 2, la frase depèn de la puntuació i, per tant, queda sota el control conscient de l'autor o de l'editor i que aquesta qüestió és molt important en el cas dels texts medievals, com el *Tirant*, perquè la puntuació no s'hi introdueix fins molt més tard, i és obra de l'editor.

Per a totes les paraules que apareixen en el *Tirant* tenim comptabilitzat el nombre total d'ocurrències al llarg del llibre, així com el nombre de vegades que apareixen per cada capítol. Aquestes freqüències ens seran d'utilitat per a les anàlisis que es faran en els capítols 12 i 13. A més a més, per cada capítol del *Tirant* hem comptat el nombre d'ocurrències per cadascuna de les lletres, i ho farem servir en el capítol 11.

4.1.2 El Tirant en Dades

En el *Tirant* hi ha un total de 415.293 paraules, de les quals 398.242 són les que s'empraran en l'estudi perquè no apareixen en cursiva. D'aquestes, hi ha un total de 13.828 paraules (formes) diferents. La paraula més abundant és la conjunció *e*, que apareix 22.114 vegades, i les que venen a continuació com a més abundants són *de*, *la* i *que* que hi surten, respectivament 14.890, 14.202 i 13.556 vegades. Hi ha 5.599 paraules que només apareixen una vegada en tot el text, 1.959 que surten dues vegades i 1.114 paraules que surten tres vegades. La paraula *Tirant* apareix 2.913 vegades, és la setzena més freqüent, i no apareix per primer cop fins al capítol 29.

Un cop eliminades totes les paraules en cursiva, comptem un total de 12.719 frases. La més llarga té 384 paraules i està al capítol 15, i les més curtes són frases d'una sola paraula, normalment una exclamació o una pregunta. En el *Tirant* hi ha 13 capítols formats per una sola frase, la majoria d'elles de menys de 50 paraules. El més llarg entre aquests capítols és el 45, que té una sola frase de 81 paraules. El capítol 189, el més llarg del *Tirant*, està format per 6.521 paraules no en cursiva i 251 frases. La suma de totes les frases que apareixen en capítols de més de 200 paraules és de 12.563.

4.2 Subdivisió del *Tirant* en trossos de text

La divisió "natural" del *Tirant* en trossos de text és per capítols que, a més a més, és la forma en la què ha estat editat.

Però, la distribució de molts dels estadístics que emprarem en les anàlisis d'aquesta tesi depenen de la llargada del tros de text analitzat, comptada en nombre de paraules. Així, per exemple, el valor esperat del nombre de paraules diferents d'un capítol augmenta en augmentar la seva llargada, i la variança de la mitjana per capítol del nombre de lletres per paraula disminueix en augmentar la llargada del capítol. Per evitar de confondre els efectes d'una possible no homogeneïtat en l'estil del *Tirant* amb qüestions relatives amb la llargada del capítol, hem decidit subdividir també el llibre en blocs de text de llargada constant.

La majoria de les anàlisis que realitzarem es faran en paral·lel, tant per capítols com per blocs de llargada constant.

4.2.1 Subdivisió del Tirant en capítols

El *Tirant* està format per 487 capítols, dels quals dos, el 71 i el 107, és subdivideixen en dos, que considerarem per separat i que denotarem per 71a, 71b, 107a i 107b. Per tant, a efectes de l'anàlisi que es farà, existeixen 489 capítols. En cap moment fem servir els texts de la dedicatòria, del pròleg ni del colofó.

Un cop eliminades totes les paraules en cursiva, hi ha 19 capítols editats íntegrament en cursiva, pels que la llargada és zero, i 470 capítols amb llargada més gran que zero, dels que el més extens té una llargada de 6.521 paraules. Sense eliminar cursives la llargada dels capítols va des de 24 fins a 6.567 paraules.

Pel fet ja esmentat de que la distribució de moltes de les unitats emprades en l'estudi depenen de la llargada del text, s'ha decidit no considerar per a les anàlisis els capítols de llargada inferior a les 200 paraules, un cop eliminades les que es troben en cursiva. L'única excepció és quan s'estudia la llargada dels capítols, en la que es consideren tots. En el *Tirant* hi ha 425 capítols amb llargada superior a les 200 paraules, que són els que farem servir sempre que la unitat de text sigui el capítol.

4.2.2 Subdivisió en blocs de 1000 paraules consecutives

Hem dividit tot el *Tirant* en 396 blocs de *N=1000* paraules consecutives, més un darrer bloc amb les 349 paraules últimes del llibre, que no s'utilitzaran en les anàlisis per blocs. Per fer la divisió en blocs s'han fet servir totes les paraules del Tirant, tret de les que apareixen en cursiva. Es consideraran, per tant, 396.349 paraules. El fet que aquesta xifra no coincideixi del tot amb el total de paraules del *Tirant* és deu a que, per a la subdivisió en blocs, s'ha fet servir una versió digital diferent, que és anterior i menys fiable que la versió que s'ha emprat pels capítols. Quan es va començar l'estudi la versió de l'Institut Joan Lluís Vives encara no existia. La versió que s'ha fet servir per a la subdivisió en blocs parteix de la que es pot trobar a l'adreça web: *http://www.fut.es/bdt/tirant*, després d'una neteja d'errors força obvis.

Triar blocs de 1000 paraules consecutives fa que hi hagi blocs que abracen més d'un capítol i que hi hagi capítols trencats en varis blocs. Les raons per subdividir en unitats de 1000 paraules són:

- Si suposem que cada bloc és una mostra de text representativa de l'estil del seu autor, els valors dels estadístics que fem servir per quantificar-ne l'estil intenten estimar els corresponents paràmetres poblacionals. L'elecció de la llargada de bloc suposa un compromís entre obtenir estimadors amb variança petita i tenir la possibilitat de detectar una possible frontera de forma precisa. Blocs llargs permeten estimadors més precisos, amb variances menors, que detecten diferències més petites, però no permeten determinar amb exactitud la frontera. Blocs curts donen estimadors menys precisos, disminuint la probabilitat de detectar diferències significatives, però permeten establir la frontera amb molta més exactitud.
- La llargada mitjana de capítol és de 847,4 paraules per capítol si els considerem tots, i 927 si eliminem els capítols curts, de menys de 200 paraules. La subdivisió del text en blocs de 1000 paraules dóna un nombre de blocs semblant al nombre de capítols i la llargada dels blocs és propera a la llargada mitjana dels capítols.
- L'elecció de 1000 paraules permet interpretacions molt directes d'algunes de les mesures emprades, com la llargada de paraula o la freqüència d'us de paraules eina i/o marcadors d'estil.
- 1000 paraules són fàcils de manipular.

4.3 Descripció de la base de dades

S'ha organitzat la base de dades en dos directoris: *Capitols*, on es troben les dades obtingudes dels capítols, i *Blocs*, on hi ha les dades per la subdivisió en blocs de 1000 paraules. Els dos directoris tenen la mateixa estructura, formada per tres subdirectoris en els que es troben, respectivament, les dades relatives a:

- Distribució de llargada de paraula, capítol i frase, distribució de lletres i paraules més freqüents, índexs de diversitat (*TesiCp* per capítols, *TesiBl* per blocs).
- Distribució de vocabulari (*DVocabCp* per capítols, *DVocabBl* per blocs).

- Freqüència d'ús de totes les paraules (*FrqTotCp* per capítols, *FrqTotBl* per blocs).

Les dades es troben en format MIINITAB.

4.3.1 Unitats d'estadística textual analitzades

Tal i com hem senyalat en el capítol 2, les unitats d'estadística textual, idealment, s'han d'escollir de manera que no estiguin sota el control conscient de l'autor, i han de ser fàcilment quantificables, freqüents i pròpies de l'autor i no del gènere o de l'època. Això, en la pràctica, és impossible d'aconseguir del tot. Als subdirectoris *TesiCp*, per capítols, i *TesiBl*, per blocs, hi tenim els fitxers que contenen les unitats que hem analitzat:

- Llargada de paraula, mesurada en nombre de lletres per paraula, que analitzarem en el capítol 8.
- Llargada de frase, mesurada en nombre de paraules per frase, que analitzarem en el capítol 9.
- Llargada de capítol, mesurada en nombre de paraules per capítol, que analitzarem en el capítol 10.
- Freqüència d'ús de lletres, que analitzarem en el capítol 11.
- Freqüència d'ús de les paraules més freqüents, que analitzarem en el capítol 12.
- Mesures de riquesa i diversitat de l'estil, que analitzarem en el capítol 13.

El fitxer amb les dades per la subdivisió del *Tirant* en blocs de 1000 paraules consecutives, *TesiBlocs.mpj*, està format per quatre fulls, cadascun format per 397 files, una per cada bloc. Les dades pels capítols del *Tirant* estan organitzades en 3 arxius tots ells formats per sis fulls. Els tres arxius contenen, respectivament:

1) les dades per tots els capítols, en el fitxer *TesCap.mpj*, format per 489 files,
2) les dades per tots els capítols excepte aquells que estan editats íntegrament en cursiva, en el fitxer *TCpNg0.mpj*, format per 470 files. És el que es fa servir per a l'anàlisi de la llargada de capítol.
3) les dades per als capítols amb llargada superior a les *200* paraules, en el fitxer *TCpNg200.mpj*. Està format per 425 files i és el full principal, amb el que s'han fet pràcticament totes les anàlisis,

Dels sis fulls, tres contenen, bàsicament, la mateixa informació per blocs i per capítols i tres són diferents. En tots els fulls, les primeres dues columnes contenen el número, be del capítol be del bloc, i la seva llargada. A continuació es donen detalls dels sis fulls.

4.3.1.1 Llargada de bloc

Aquest full es dóna només per als blocs. Conté:

- Llargada de bloc, en nombre de paraules. Tots els blocs tenen 1000 paraules, tret de l'últim, que té una llargada de 349 paraules.
- Per cada bloc, quins capítols abraça.

4.3.1.2 Llargada de Paraula

El full amb les distribucions de llargada de paraula es dóna tant per capítols com per blocs.

Hem comptat per a totes les paraules del Tirant la seva llargada, mesurada en nombre de lletres. Les unitats utilitzades han sigut:

- Llargada mitjana de paraula.
- Nombre de paraules *k* lletres, per *k=1,...,17*, i on les paraules de llargada 10 i superior han sigut ajuntades en una sola categoria.

4.3.1.3 Llargada de frase i de capítol

Aquest full només es dóna per als capítols.

Per a tots els capítols del *Tirant* hem comptat el nombre de frases, i per cada frase hem comptat el nombre de paraules. Aquestes dades s'analitzen en el capítol 9.

La llargada de frase només ha sigut estudiada per als capítols, atès que aquests estan formats per un nombre enter de frases. Pels blocs en canvi, l'inici i el final difícilment coincideix amb els d'una frase. Si eliminéssim les fraccions de frase a l'inici i final de bloc, llavors els blocs deixarien de tenir llargada constant. Per tant, aquestes dades només es troben en els fitxers de dades per capítols.

S'han agrupat les frases en deu categories segons la seva llargada, tal i com mostra la taula 4.1:

Categoria	Llargada Frase
Cat. 1	1-9 paraules
Cat. 2	10-15 paraules
Cat. 3	16-20 paraules
Cat. 4	21-25 paraules
Cat. 5	26-30 paraules
Cat. 6	31-40 paraules
Cat. 7	41-50 paraules
Cat. 8	51-60 paraules
Cat. 9	61-70 paraules
Cat. 10	>70 paraules

Taula 4.1: Categories en les que ha estat classificada la llargada de frase pel *Tirant.*

Per cada capítol tenim la següent informació relativa a la llargada de frase:

- Llargada mitjana de frase per capítol.
- Nombre de frases per capítol
- Nombre de frases en cadascuna de les 10 categories en les que s'han agrupat.

Les dades relatives a la llargada de capítol contenen, per cada capítol:

- Llargada de capítol, en nombre de paraules.

4.3.1.4 Freqüència d'ús de lletres

Són dos fulls que només es donen per als capítols.

Hem comptat el nombre d'ocurrències en cada capítol de cadascuna de les lletres. Els fulls contenen:

- les 24 lletres que formen l'alfabet, excloses la *k* i la *w*.
- les 36 lletres amb grafia diferent, excloses la *k* i la *w*. Conté les 14 grafies diferents de les vocals.

4.3.1.5 Freqüència d'ús de paraules més freqüents

El full amb el nombre d'ocurrències de les paraules més freqüents es dóna tant per capítols com per blocs.

Hem comptat el nombre de vegades que apareixen en cada capítol o bloc cadascuna de les paraules que surten en el Tirant. S'han fet servir les 25 més paraules més freqüents, després d'excloure els substantius, per estudiar la homogeneïtat d'estil i, un cop estimat el punt en què l'estil canvia, s'ha determinat quines paraules, entre les 101 més abundants, són útils per discriminar entre els dos estils.

Per cada capítol i bloc es disposa de:

- Nombre d'ocurrències de les 101 paraules més abundants.

4.3.1.6 Mesures de Riquesa i Diversitat de vocabulari

El full amb les mesures de riquesa i diversitat es dóna tant per capítols com per blocs.

Els deu índexs de riquesa i diversitat que s'analitzaran en el capítol *12* de la tesi són:

- nombre de paraules amb grafia diferent, V,
- índex de Simpson, D,
- Entropia, H^s,
- nombre de paraules que apareixen una vegada, V_1,
- nombre de paraules que apareixen dues vegades, V_2,
- índex de Brillouin, B,
- índex de McIntosh, M,
- índex de Honoré, R,
- relació entre el nombre de paraules que surten dues vegades i el nombre de paraules diferents, V_2/V,
- índex de Brunet, W.

Tots deu índexs s'han calculat tant per blocs com per capítols.

4.3.2 Distribució de Vocabulari

Els fitxers amb les distribucions de vocabulari es troben en els subdirectoris *DVocabCp* per capítols i *DVocabBl* per blocs. Tot i que aquestes dades no s'analitzen a la tesi, tenim previst fer-ho en un futur inmediat.

Amb el nombre de vegades que apareixen en cada capítol o bloc cadascuna de les paraules del *Tirant* es construeixen les distribucions del vocabulari, que compten quantes paraules apareixen una sola vegada, quantes dues vegades, i així fins al nombre màxim d'aparicions d'una mateixa paraula en un capítol o bloc.

Els fitxers amb les dades estan formats per un únic full que conté, per tots els capítols o blocs, la seqüència de les V_d, nombre de paraules que apareixen d vegades, per $d=1,..., d_{max}$, on d_{max} és igual a *354* pels capítols i *94* pels blocs. Per tant els fulls estan composats per d_{max} files i una primera columna on hi ha la seqüència de les d, i 489 per capítols, i 397 per blocs, columnes més amb les V_d.

La taula 4.2 conté, com a exemple, la distribució del vocabulari per al primer bloc de 1000 paraules. Per a fer-la més llegible s'han eliminat les files amb $V_d=0$.

D	V_d
1	280
2	78
3	17
4	13
5	11
6	6
7	5
8	3
9	1
10	2
12	1
13	1
14	2
20	1
27	1
35	1
41	1
53	2

Taula 4.2: Distribució del vocabulari per al primer bloc de 1000 paraules. V_d, el nombre de paraules que apareixen d vegades. De la taula s'han eliminat les files amb $V_d=0$.

4.3.3 Freqüència d'ús de totes les paraules

Als subdirectoris *FrqTotCp*, per capítols, i *FrqTotBl*, per blocs, hi tenim els fitxers amb les distribucions de vocabulari. Tot i que aquestes dades tampoc no s'analitzen a la tesi, tenim previst fer-ho en un futur inmediat.

S'ha comptabilitzat, per totes les 13.828 paraules diferents del *Tirant*, quants cops surten en cada capítol o bloc. El resultat és una taula amb 13.828 files, on la primera columna conté les paraules ordenades, la segona el nombre total d'aparicions de la paraula en el *Tirant* i les següents contenen el nombre de vegades que apareix la paraula en cada capítol o bloc.

Es disposa de dues ordenacions per a les paraules, alfabètica i per freqüència. Pels capítols, els fitxers que contenen les dues ordenacions són *FrqTCpAlf.mtw* i *FrqTCpFrq.mtw*, per blocs són *FrqTBlAlf.mtw* i *FrqTBlFrq.mtw*.

La taula 4.3 conté un fragment de la taula ordenada per freqüència, amb les 27 paraules més abundants per als primers 10 capítols.

Paraula	**Freq.**	**Cap.1**	**Cap.2**	**Cap.3**	**Cap.4**	**Cap.5**	**Cap.6**	**Cap.7**	**Cap.8**	**Cap.9**	**Cap.10**
e	22114	12	26	66	33	63	35	20	13	12	44
de	14890	15	28	46	29	46	15	20	9	9	27
la	14202	9	19	48	34	42	27	10	13	9	29
que	13556	8	9	53	13	34	23	16	6	7	21
lo	9413	10	10	26	9	33	27	3	1	6	16
en	7765	6	12	20	21	17	16	6	9	4	14
a	7528	1	11	22	13	16	13	4	6	4	19
per	6871	4	8	20	11	21	11	5	6	7	18
no	5849	1	3	19	5	8	7	5	4	3	11
l	5169	7	2	9	7	12	10	5	5	4	15
los	4666	5	1	13	3	20	6	0	1	4	6
com	4379	2	3	11	4	16	3	2	4	0	10
ab	4342	1	7	8	12	12	2	0	0	1	19
les	3813	3	6	15	5	7	1	3	3	2	13
d	3702	0	3	6	6	16	5	2	1	2	5
tirant	2913	0	0	0	0	0	0	0	0	0	0
li	2680	1	3	4	5	3	8	0	0	0	1
qui	2615	1	1	11	2	11	3	2	2	1	7
del	2430	3	1	3	3	5	4	1	0	1	6
se	2409	1	4	3	5	7	3	0	0	0	2
és	2316	6	3	9	5	2	1	2	0	1	2
molt	2285	1	8	2	8	10	6	1	3	0	4
gran	2225	0	2	7	4	19	4	1	3	1	5
jo	2099	0	1	10	2	1	1	0	3	0	12
rei	1949	0	2	0	0	21	11	1	1	4	11
si	1855	3	3	2	3	2	4	3	1	2	5
tots	1729	0	3	2	3	3	3	0	0	0	5

Taula 4.3: fragment de la taula amb les distribucions de freqüències de totes les paraules, ordenada per freqüència, amb les 27 paraules més abundants per als primers 10 capítols.