PROJECTE FINAL DE CARRERA

LLICENCIATURA EN CIÈNCIES I TÈCNIQUES ESTADÍSTIQUES

Títol: **DISTRIBUCIÓ DE SICHEL I RIQUESA DE VOCABULARI: APLICACIÓ AL *TIRANT LO BLANC***

Alumnes: **Emma Arcos Fuster**
Ignasi Serra Pons

Director: **Josep Ginebra Molins**

Data: **Juliol 2004**

Facultat de Matemàtiques i Estadística
UNIVERSITAT POLITÈCNICA DE CATALUNYA

DADES DEL PROJECTE:

Nom de l'estudiant: Emma Arcos Fuster
Ignasi Serra Pons

DNI:

Títol del Projecte:

Director del Projecte:

Tutor del Projecte:

QUALIFICACIÓ

Excel·lent (9'5)

MEMBRES DEL TRIBUNAL (nom i signatura)

President: Pere Grima

Vocal: José A. Lubary

Secretari Josep GINEBRA

Data: 23 de juliol de 2004

DISTRIBUCIÓ DE SICHEL I RIQUESA DE VOCABULARI: APLICACIÓ AL *TIRANT LO BLANC*

Emma Arcos Fuster
Ignasi Serra Pons

Director de projecte

Josep Ginebra i Molins

Juliol 2004

Després de tot aquest temps treballant amb aquest projecte, no volem acabar sense donar les gràcies a la gent que ens ha ajudat desinteressadament. Gràcies a ells, la feina ha resultat menys feixuga i es fa possible l'entrega d'aquest treball.

Gràcies Josep,
per oferir-nos aquest treball tan interessant i original, i sobretot per les xerrades que hem compartit al teu despatx. Ara ja estem a punt per fer el cim del Pedraforca,

a l'Àlex Riba,
per proporcionar-nos el material necessari,

a l'Adela,
per assessorar-nos en el Matlab,

als nostres familiars,
per motivar-nos i suportar-nos en moments prou difícils, com per deixar-ho estar, i a l'Àlex que ha volgut posar el seu granet de sorra,

als amics,
que han intentat treure'ns un somriure quan més ho necessitàvem,

i sobretot, a les costes del del Garraf i a la Serra del Catllaràs

ÍNDEX

CAPÍTOL 1:
Introducció

1.1 Objectius

Fa temps que els experts en literatura medieval intenten resoldre les diferents incògnites que han sorgit arran de l'autoria del *Tirant lo Blanc*. El llibre es va escriure entre el 1460 i el 1464, però no va ser editat fins al 1490. Actualment hi ha moltes hipòtesis diferents que van des de l'autoria única d'en Joanot Martorell (1414-1465), que és la proposta més arrelada actualment i defensada per experts com Martí de Riquer, fins a la hipòtesi d'una doble autoria amb Martí Joan de Galba completant la part final del llibre redactat per Joanot Martorell. Els que defensen la primera de les hipòtesis es recolzen en la dedicatòria de la primera edició del llibre, en canvi, els que defensen la segona es basen en el colofó, on explica que la última de les cinc parts que formen el llibre va ser escrita per Martí Joan de Galba després de la mort de Martorell.

El grau de participació de Galba en el *Tirant* ha estat objecte d'intensos debats. Experts com Coromines (1956), els quals es recolzen en el colofó, diuen que existeix un canvi d'estil entre els capítols 300 i 350. Molts d'altres, com Marinesco (1979) defensen l'autoria única. Nosaltres intentarem reduir tota aquesta incertesa amb eines estadístiques que permeten modelar el vocabulari i serviran per mesurar la diversitat i la riquesa del vocabulari d'un autor. Les eines que emprem, veurem que ens serveixen per detectar un possible canvi d'autor o d'estil al voltant del capítol 380 del *Tirant lo Blanc*, confiramnt els resultats trobats per Riba (2002) i Riba i Ginebra (2003) mitjançant altres mètodes.

L'anàlisi estadística de l'estil literari busca característiques quantificables d'un text, que l'autor rarament controli, i les aprofita per caracteritzar l'evolució de l'estil. En el nostre cas, caracteritzem l'estil del *Tirant lo Blanc* a base de modelar la distribució del seu vocabulari i així poder quantificar la seva riquesa i diversitat.

L'objectiu principal del nostre projecte és detectar si realment existeix un canvi en la diversitat del vocabulari en el *Tirant lo Blanc* a partir de l'estimació dels paràmetres d'una distribució que ajusti l'ús del vocabulari. Això ho farem a partir d'un inventari de totes les paraules que apareixen en el llibre i del nombre d'ocurrències de cadascuna d'elles.

L'assumpció bàsica de la qual partim, és que un cert autor coneix un conjunt de paraules que formen el seu vocabulari, les quals tenen probabilitats diferents de ser escrites, és a dir, hi ha certes paraules que escriu més freqüentment que d'altres. Agafant una mostra d'un text de l'autor, denotem per N el nombre de paraules que conté i per V el nombre de paraules diferents. Comptant la freqüència de cada paraula obtenim V_1, que són les paraules que només apareixen una vegada, V_2, que són les que apareixen dues vegades, V_3, que són les que apareixen tres vegades i així successivament, fins a obtenir la seqüència $V_1, V_2, V_3, ..., V_r$.

Yule (1944) va trobar que la distribució per la proporció de paraules que apareixen r vegades en un text de tamany N podria ser modelada a través d'una mixtura de Poissons. Sichel (1986), basant-se amb les hipòtesis de Yule, proposa modelar la seqüència $V_1, V_2, V_3, ..., V_r$ a partir d'una mixtura de Poissons truncada en el 0, on la distribució de barreja és la Gaussiana inversa generalitzada. Aquesta distribució és la que utilitzem en el projecte per intentar assolir els nostres objectius

1.2 Descripció dels continguts

En el capítol 2 expliquem la problemàtica de l'autoria del *Tirant lo Blanc* i introduim possibles maneres per ajudar a desvelar les incògnites que amaga. Aquests camins es basen, per exemple, en l'estudi de la llargada de frase, la llargada de capítol, ambdues mesurades amb número de paraules, la freqüència d'ús de les lletres, etc. Nosaltres intentarem mesurar la riquesa del vocabulari en el *Tirant lo Blanc* mitjançant la freqüència d'ús de les paraules.

En el capítol 3 presentem la base de dades amb totes les paraules del llibre resumides per freqüències, és a dir, tenim el nombre de vegades que surt cada paraula en cadascun dels capítols del *Tirant lo Blanc*.

Estudiant la naturalesa de les dades hem de decidir quina família de distribucions pot servir per ajustar la freqüència d'ús de les paraules. Per això escollim la distribució de Sichel, que es basa en una barreja de Poissons que expliquem amb tot detall en el capítol 4 del projecte, en el qual introduim la teoria amb la que ens basarem per estimar els seus paràmetres.

En el capítol 5 presentem quatre mètodes per estimar els paràmetres d'aquesta distribució per les dades corresponents als diferents capítols del llibre. Tot seguit, en el capítol 6, mitjançant el test Khi-quadrat, avaluaem si aquestes distribucions ajusten correctament les dades. A partir d'aquí seleccionarem els estimadors dels paràmetres dels mètodes que ens han ajustat millor.

En el capítol 7 calculem diferents índexs de la riquesa en el *Tirant lo Blanc*. Com veurem, ens interessen aquells índexs pels que el seu valor esperat no depengui del tamany de la mostra, és a dir, del número de paraules del text. En aquest punt del projecte esbrinarem si existeix cap relació entre els paràmetres estimats en el capítol 5 i els indicadors de riquesa calculats i, d'aquesta manera entendre més clarament el significat dels paràmetres estimats.

En el capítol 8 incrementem el tamany mostral a base d'ajuntar els 487 capítols en 35 blocs de text més llarg, de forma que en cada bloc tenim un major nombre de paraules. Amb aquestes dades ens trobem que amb els dos paràmetres de la distribució de Sichel que hem estimat fins ara pels capítols del *Tirant*, no podem recollir tota la informació i ens plantegem la distribució de Sichel amb tres paràmetres. Així doncs, procedim altra vegada a estimar els paràmetres i a analitzar les dades com a la resta de la tesi amb l'objectiu de detectar un canvi en la diversitat del vocabulari a partir d'un cert bloc.

Finalment, tenint en compte els resultats obtinguts, en l'últim capítol del projecte expliquem possibles propostes futures per treballar amb aquest tipus de dades.

CAPÍTOL 2:

La problemàtica del *Tirant lo Blanc*

Al voltant de l'autoria del *Tirant lo Blanc* hi ha una sèrie d'incògnites que encara no han estat resoltes. De les moltes propostes que s'han estudiat, la que s'havia considerat com a més versemblant fins fa poc és la que afirma que el gruix de l'obra fou escrit per Joanot Martorell (1414-1465), entre 1460 i la seva mort, mentre que Martí Joan de Galba (¿-1490) s'encarregà de l'edició de l'obra el 1490, després de reelaborar alguns passatges i potser acabar l'obra. Així doncs, el grau de participació de Galba ha estat objecte d'intens debat.

La doble autoria Martorell/Galba es sustenta en el que diu el colofó, segons el qual Galba hauria acabat la novel.la de Martorell afegint-hi la quarta part. Entre els que accepten, o acceptaven, la participació, en major o menor mesura, de Galba cal citar Martínez i Martínez (1916), que explica com, en l'inventari dels béns de Galba, es parla:

> *"... de un libro encuadernado en pergamino llamado Tirant, y a continuación de otro acabado, el que tienen por original los impresores; bien claramente demuestra la frase del segundo tot acabat, que el primero no lo estaba, y no hay que dudar que éste era escrito por Martorell, al que le faltaba la quarta parte, que la muerte no le dio tiempo a escribir, y a la segunda, toda terminada consistía en la copia de la s tres partes escritas por mossen Joanot, corregidas i arregladas por Galba, con el fin de que resultase la unidad tan pregonada por los comentaristas, y la última parte del libro es de su cosecha propia."*

Entwistle (1927) atribueix quatre capítols a Galba, del 410 al 413, que corresponen a l'aventura del cavaller Espèrcius, sense tancar la possibilitat que fos autor d'altres.

Moll (1933), en estudiar els refranys del *Tirant*, observa com aquests apareixen molt sovint en l'obra, però que manquen a partir del capítol 326, i considera que aquesta pot ser una raó, i una frontera, per a precisar la intervenció de Galba. Guia (1996) creu que aquesta afirmació és errònia, en aparèixer refranys en capítols posteriors al 326.

Menéndez y Pelayo (1943) reconeix una paternitat important a Galba, que estaria al voltant d'una quarta part de l'obra.

Martí de Riquer (1947) opta per una intervenció progressiva de Galba en la novel.la, tot i que més endavant rectifica (Riquer, 1990) i es decideix per un sol autor. Coromines (1956), seguint la interpretació literal de colofó, accepta la veu autoritzada de Riquer i, a partir de l'anàlisi estilística de l'obra, defensa que la intervenció de Galba es fa més patent a partir del capítol 320. Pel que fa a la frontera entre Martorell i Galba, la situa entre els capítols 300 i 350, en la quarta de les cinc parts en què s'organitzen les edicions modernes del *Tirant*. Pel que fa a l'estil, Coromines afirma que:

> *"és amb Galba quan apareix verament dins del Tirant lo Blanc l'anomenat 'Estil de la Valenciana Prosa' o almenys quan s'accentua fins a un amanerament de mal gust"*

Coromines observa que, a partir del capítol 417, tendeix a ésser mecànicament constant l'ús d'epítets anteposats i pot atribuir-se a Galbala sovintejada col·lecció del verb final i el gust pels cultismes, i assegura que el llenguatge del continuador és *"redundant, emfàtic, farragós i feixugament declamatori"*.

D'Owler (1961), Goertz (1967), Bosch (1987), Ferrando (1989), Rubiera (1990, 1992) i Wittlin (1990), van acceptar després de 1947 l'autoritzada opinió que llavors mantenia Riquer sobre la paulatina i progressiva intervenció de Galba a partir de l'aventura africana de Tirant.

En front a la teoria de la doble autoria del *Tirant*, altres investigadors, com Givanel (1916), afirmen, en contra de la declaració del colofó del llibre, que la redacció es deu a Martorell, i que Galba es limità a fer la correcció i preparació del manuscrit abans de lliurar-lo a la impremta.

Marinesco (1979), investigador romanès especialitzat en els aspectes històrics del *Tirant*, defensa l'autoria única basant-se en l'anàlisi de la dedicatòria i del colofó amb la successió d'aventures narrades en la novel.la.

Riquer (1990), després de sospesar totes les possibilitats, canvia la hipòtesi formulada el 1947 i es decanta per un *Tirant lo Blanc* d'un sol autor, Joanot Martorell, i d'un Galba que es limita a donar una lleugera revisió, i molt hipotèticament, a escadusseres intromissions. Riquer, doncs, aposta ara per una novel.la d'un sol autor, encara que existeixen en el manuscrit lleugeres contradiccions, que segons l'estudiós, en escriure un llibre tan llarg és molt possible que l'autor s'hagués contagiat de l'anomenada *"Valenciana Prosa"* d'escriptors més joves. Amb Riquer canvien d'idea, i adopten la de l'autoria única, alguns dels seguidors, com Lola Badia.

En una línia rupturista, Guia (1996,98) planteja una nova teoria i defensa que un Martorell a les acaballes podria haver proporcionat al jove Joan Rois de Corella el manuscrit del *Guillem de Varoic*, lletres de batalla i altres materials sobre el món de cavalleries que Corella aprofita per escriure fins al capítol 154 del Triant lo Blanc, quan mor Martorell i Corella en continua la redacció tot sol.

2.1 Com afrontar aquesta problemàtica amb eines estadístiques

La hipòtesi que està a la base de tots els estudis d'estilometria i, més en general, de l'estilística computacional, és que l'autor disposa d'un vocabulari format per un nombre finit de paraules, i que a l'hora de triar-ne una, ho fa seguint patrons no conscients, patrons que poden deprendre del context o del gènere. La feina de l'analista de l'estil és trobar trets característics d'un autor dels que ell, probablement, no n'és conscient. Aquests trets poden ser mesurats quantitativament, que serà utilitzat per comparar-lo amb d'altres escriptors.

2.1.1 Caracterització de l'estil literari

La quantificació de l'estil es pot fer a tres nivells, en funció del grau de complexitat a caracteritzar. Un primer nivell compara les freqüències d'ús d'unitats lingüístiques fàcils d'identificar i de comptar, que siguin freqüents i difícilment controlables concsientment per l'autor, per exemple la llargada de paraula o frase, o bé la proporció d'ús d'algunes paraules. Escollir les unitats a estudiar és un problema obert, que convé adaptar a cada autor, gènere i temps. En un segon nivell, que obliga a fer un inventari de totes les paraules i a comptabilitzar les vegades que apareix cadascuna en el text, es vol caracteritzar la riquesa i diversitat de vocabulari. L'objecte d'estudi no és res més, que identificar la freqüència d'ús de paraules determinades.
I un tercer nivell modela l'ordre d'aparició del vocabulari i d'altres unitats lingïístiques. Estudien el llenguatge com un procés estocàstic.

Una altra possibilitat consisteix en comparar les proporcions d'ús de noms, verbs, adjectius, preposicions, conjuncions, articles i altres parts morfo-sintàctiques del llenguatge. La gent més cultivada empra més substantius, una actitud més activa es pot traduïr en un percentatge de verbs més alt. El principal inconvenient d'aquesta variable respecte les anteriors és que no és fàcil reconèixer automàticament la funció gramatical de les paraules.

2.1.2 Diversitat i Riquesa de Vocabulari

A l'hora de comparar textos, on hi ha més informació és a nivell de lèxic, perquè és on hi ha més dades. Es parteix de la hipòtesis que l'autor disposa d'un vocabulari, composat per una llista de paraules que les utilitzarà a l'hora d'escriure. L'autor tendeix a utilitzar-ne unes més que d'altres i aquest fet es pot ajustar i quantificar. L'objectiu recauria en quantificar la llista de paraules de què disposa l'autor a partir de les freqüències d'aparició, amb això, caracteritzem la diversitat del text.

El concepte de diversitat s'ha estudiat en altres camps de la ciència. Té l'origen en l'ecologia i l'estudi de la biodiversitat, i ha estat aplicat també, entre d'altres contextos científics, a la genètica, a l'anàlisi de la concentració industrial, a l'economia, a les desigualtats econòmiques i a la diversitat lingüística.

Un índex de diversitat és una mesura de "dispersió qualitativa" d'una població d'indivuidus que pertanyen a vàries categories qualitativament diferents. De la mateixa manera que estadístics com la variància, la desviació estàndard o el rang, mesuren variabilitat d'una variable discreta o contínua, els índexos de diversitat mesuren variabilitat en variables de naturalesa categòrica. En el capítol 7, s'expliquen els indicadors de riquesa que s'utilitzaran per analitzar el canvi d'estil en el decurs del *Tirant*.

Sichel (1986) i Yule (1944) plantegen el debat sobre si els estudis de riquesa de vocabulari d'un autor o text s'han de fer per paraules determinades, d'una classe

morofo-sintàctica especial com ara substantius, adverbis,verbs, preposicions i conjuncions, o bé pel global del vocabulari.
Riba i Ginebra (2000,b) analitzen l'evolució dels índexos de diversitat per estudiar l'homogeneïtat d'estil en el *Tirant*.

2.2 *Estudis Precedents sobre l'autoria del Tirant*

La problemàtica de l'autoria del *Tirant lo Blanc* ja ha estat estudiada mitjançant l'anàlisi estadística. Cabos, Ginebra i Riba han estudiat la possibilitat d'un canvi en la riquesa d'estil en la quarta part del llibre . En aquest apartat s'intenta fer un breu resum dels resultats que van obtenir.

En una primera aproximació a l'anàlisi estadística de l'estil de *Tirant lo Blanc*, Ginebra i Cabos (1998) documenten l'existència d'una frontera entre els capítols 300 i 390. Ho fan comparant la distribució de la llargada de paraula, de la llargada de frase i la freqüència d'ús d'una sèrie de paraules. En un primer intent, s'estudien aquestes variables d'interés, agafant textos per blocs d'unes 4000 paraules, prenent com a mostra el *Tirant* editat per Martí de Riquer. Troben que les diferències més grans apareixen entre els set primers blocs i els tres últims.

Per confirmar l'existència d'una frontera a la quarta, Ginebra i Cabos (1998) fan un test de permutacions agrupant els 10 blocs de text en grups de set blocs i de tres blocs de totes les 120 maneres possibles. Troben que la combinació que agafa els set primers blocs per un cantó i els tres últims per l'altre, dóna la distància entre parelles de distribucions més gran d'entre totes les 120 combinacions possibles. Fent l'anàlisi de correspondències, s'observa que les paraules llargues són més abundants a la última part , que a la resta del llibre i que la freqüència d'ús de paraules *molt, e, tant, l', sobre* i *de* augmenta a mesura que arribem al final del llibre, contràriament, les paraules *puix, sia, ja, als,k tal, si* i *cap* disminueix.

Riba centra tota la seva tesi a l'estudi de l'homogeneïtat d'estil al llarg del *Tirant*. Donada la llargada de l'obra, vol trobar una frontera interior que podria indicar tant l'existència de dos autors com de dues etapes d'escriptura diferenciades, com també, determinar en quin punt hi ha la frontera i què els caracteritza. En la segona part de la tesi aplica tècniques estilomètriques a l'estudi de l'homogeneïtat d'estil en el *Tirant*: analitza la llargada de paraula, de frase i de capítol. Posteriorment, fa una introducció a la freqüència d'us de lletres i de paraules, i finalment estudia l'evolució de mesures de riquesa i de diversitat del vocabulari, que en el nostre projecte també s'estudiaran. Els resultats més rellevants de Riba i Ginebra (2000, 2003) foren els que descriurem a continuació.

2.2.1 Resultats més rellevants

- **Estudi de la llargada de paraula**

Riba estudia l'existència d'un punt de canvi sobres la llargada mitjana de paraula dels capítols, el nombre de paraules i la proporció que representa respecte el total del capítol, per paraules de i lletres, $i=1,2,...,17$.

Els resultats foren els següents: s'observava l'existència d'un punt de canvi a aquestes dues variables entre els capítols 345 i 371, aquest punt variava en funció del mètode o la unitat d'estadística textual analitzada. Així doncs, intuïa a indicar que en aquests dos capítols es barregen dos estils. En general, en els pirmers capítols la llargada mitjana de paraules és més petita que al final del llibre.

- **Estudi de la llargada de la frase**

Riba analitza la mesura de la llargada de frase en nombre de paraules per frase. En aquest cas, no es veu cap frontera clara. Cal destacar que aquesta unitat d'estadística textual no és bona per a discriminar entre estils literaris diferents.

Cal tenir en compte que es considerava una frase, tot el conjunt de paraules que acabava en un punt, un signe d'interrogació o bé un signe d'exclamació. Així doncs, aquests indicadors depenen de la puntuació, i per tant, queda sota el control conscient de l'autor. A més a més, en els textos medievals, com el *Tirant*, la puntuació no s'introdueix fins més tard, i és obra de l'editor.

- **Estudi de la llargada de capítol**

Riba mesurava la llargada de capítol, Ni, en nombre de paraules. Els resultats no foren gaire prometedors. En les análisis gráfiques s'observaven diferències en la distribució de la llargada de capítol entre els capítols del principi i els del final del llibre, però es fèia difícil assegurar que aquestes diferències fossin significatives. A més a més, no demostrava l'existència d'un punt en el que la distribució de la llargada de capítol canviés. Així doncs, no va poder demostrar la doble autoria del *Tirant*, tal i com assegura Coromines (1956), des del punt de vista estilomètric de la llargada del capítol.

- **Freqüència d'ús de les lletres**

Riba va estudiar la freqüència en l'ús de les lletres, i analitza gràficament la proporció de lletres diferents buscant punts de canvi. Els resultats que va obtenir no eren gaire determinants a l'hora de detectar un canvi en l'estil. Coincideix, però, amb els resultats d'altres unitats d'estadística textual. Troba dos punts màxims en els capítols 119 i 382.

CAPÍTOL 3:

Anàlisi exploratori de les dades

En aquest capítol es descriu la manera com es van recollir les dades i com s'estructuren les matrius sobre les quals treballarem. Explorarem les dades per tal de descobrir quin és el camí que haurem d'emprendre per intentar assolir els nostres objectius.

Essencialment s'utilitzen dues bases de dades diferents, la primera, que és la més gran, conté la informació de la freqüència de totes les paraules del *Tirant lo Blanc* per cada capítol. Aquestes dades no són encara amb les que realment treballarem i seran modificades en el capítol 9 on, com ja veurem més endavant, ens interesserà treballar per blocs de capítols, és a dir, agregar els capítols de 14 en 14 de forma que disposem de trossos de text més llargs.

A partir d'aquesta base de dades s'obté la segona matriu de dades que és amb la que realment treballarem, ja que ens dona informació sobre el número de vegades que es repeteixen les paraules del llibre en cada capítol. Aquesta informació serà interessant, tant en el capítol 5 com en el capítol 9, de cara a estimar els paràmetres de la distribució de vocabulari per cada capítol.

3.1 Descripció de la base de dades

Per al nostre estudi farem servir la versió del *Tirant lo Blanc* editada per Martí Riquer d'Edicions 62 (1983), que van ser tractades pel Dr. Àlex Riba. Per facilitar la feina del comptatge de les paraules per mitjà de programes informàtics, les dades es van recollir d'una versió digital obtinguda de la biblioteca virtual de l'Institut Joan Lluís Vives que es trobava a la pàgina http://www.lluisvives.com.

Es van comptar les freqüències de totes les paraules emprades amb l'objectiu de quantificar l'estil literari i la riquesa del text. Un cop recollida aquesta informació, el primer que es va fer va ser eliminar totes les paraules que hi apareixien en cursiva, perquè són citacions o paraules escrites en llatí que no es corresponen al propi vocabulari de l'autor.
Per la identificació de les paraules, es consideraven paraules diferents totes aquelles que tenien grafia diferent, per exemple, *home* i *homes* serien dues paraules diferents. I també les paraules separades per guió o apòstrofs són considerades paraules diferents.

El *Tirant lo Blanc* consta de 415.293 paraules, de les quals se n'empren 398.242 que són les que no estan en cursiva. D'aquestes, n'hi ha un total de 13.828, que són diferents. Hi ha 5.599 paraules que només apareixen una vegada en tot el text, n'hi ha 1959 que s'usen dues vegades, 1.114 que surten tres vegades, i així successivament. En el nostre cas hi ha una paraula que apareix 354 vegades, és la conjunció *e*, que equival a la paraula *i* del català actual. Es troba en el capítol 189, el qual té un total de 6521 paraules, de les quals 1365 són diferents. El llibre el formen 487 capítols, però nosaltres considerem que en tenim 489 perquè el 71 i el 107 estan subdividits en 71a, 71b, 107a i 107b. Per tant, estem considerant que els capítols 71a i 71b són capítols diferents.

La nostra base de dades original està formada per una primera columna amb tota la llista de les 13.828 paraules diferents del *Tirant lo Blanc*. En una segona columna hi tenim la freqüència total de cadascuna de totes aquestes paraules al llarg de tot el llibre i, a continuació, hi tenim representada aquesta mateixa freqüència per cadascun dels capítols. És a dir, si sumem la freqüència de tots els capítols, obtenim la freqüència total per

cadascuna de les paraules del llibre representada en la segona columna. Aquesta informació no és la que estudiarem a fons en la memòria, però és la base de la qual hem de partir per organitzar-nos les dades amb les que realment treballarem. En la taula 3.1 presentem una part d'aquesta informació per tal d'entendre millor la base de dades original de la qual partim.

	Paraula	**Freqüència total**	**Freq. Cap.1**	**Freq. Cap.2**	**...**	**Freq. Cap.67**	**...**	**Freq. Cap.107a**	**...**	**Freq. Cap.487**
1	***e***	22114	12	26	...	138	...	50	...	29
2	***de***	14890	15	28	...	86	...	41	...	13
3	***la***	14202	9	19	...	69	...	22	...	8
4	***que***	13556	8	9	...	81	...	23	...	10
5	***lo***	9413	10	10	...	69	...	21	...	8
6	***en***	7765	6	12	...	43	...	19	...	4
7	***a***	7528	1	11	...	55	...	17	...	4
8	***per***	6871	4	8	...	38	...	11	...	4
9	***no***	5849	1	3	...	26	...	5	...	2
10	***l***	5169	7	2	...	32	...	14	...	10
11	***los***	4666	5	1	...	40	...	13	...	4
12	***com***	4379	2	3	...	25	...	10	...	3
13	***ab***	4342	1	7	...	32	...	10	...	2
14	***les***	3813	3	6	...	32	...	4	...	1
15	***d***	3702	0	3	...	20	...	7	...	3
16	***tirant***	2913	0	0	...	30	...	3	...	0
17	***li***	2680	1	3	...	22	...	4	...	3
18	***qui***	2615	1	1	...	16	...	5	...	3
19	***del***	2430	3	1	...	13	...	6	...	1
...	...	...	...	...	...	...	...	...	...	...
13828	***zels***	1	0	0	...	1	...	0	...	0

***Taula 3.1**: Exemple de la base de dades on presentem la freqüència total i per capítols de les paraules.*

En la taula 3.1 observem, per exemple, que la paraula *e* apareix un total de 22.114 vegades en tot el llibre, 12 vegades en el capítol 1, 26 en el 2, 138 en el 67, 50 en el 107a i 29 en el darrer capítol. El fet que hi hagi tantes diferències en l'aparició de les paraules en els diferents capítols del llibre és degut a què els capítols tenen llargades molt variables, és a dir, el nombre total de paraules en cada capítol pot ser molt diferent. En aquesta taula no mostrem el tamany dels capítols, però més endavant, en la taula 3.3, la qual presenta les dades que realment analitzem en aquest projecte, si que representem el tamany dels capítols, *N*.

Aquestes dades les utilitzarem en el capítol 9 on formarem blocs de capítols amb l'objectiu d'augmentar el nostre tamany mostral. Simplement sumarem les freqüències per cada grup de 14 capítols i un últim bloc de 13 capítols. Així doncs, obtindrem una altra base de dades que en lloc de 489 capítols tindrà 35 blocs de capítols. La taula 3.2 ens mostra l'esquema de la base de dades per blocs de capítols de la qual partim en el capítol 9. Hi observem, per exemple, que la paraula *e* es repeteix 431 vegades en el primer bloc, el qual conté la freqüència de les paraules dels 14 primers capítols. En el segon bloc, comprès entre els capítols 15 i 28, apareix un total de 657 vegades i en l'últim bloc, comprès per els 13 darrers capítols del llibre, apareix 733 vegades.

Igual que ens passa per capítols, les dades representades per blocs ens mostren diferències força grans entre les freqüències de paraules pels diferents blocs. A l'ajuntar les freqüències per capítols que tenen llargades diferents, els blocs que en resulten també tenen tamanys molt variables. El bloc de menys paraules és el 15 , amb 3441 paraules i el que en té més és el 12, amb 28.976 paraules.

	Paraula	Freqüència total	Freq. Bloc 1 (Cap. 1-14)	Freq. Bloc 2 (Cap. 15-28)	...	Freq. Bloc 35 (Cap. 477-489)
1	*e*	22114	431	657	...	733
2	*de*	14890	307	392	...	492
3	*la*	14202	298	414	...	362
4	*que*	13556	264	399	...	290
5	*lo*	9413	208	340	...	198
6	*en*	7765	153	223	...	203
7	*a*	7528	152	199	...	204
8	*per*	6871	144	182	...	177
9	*no*	5849	82	155	...	88
10	*l*	5169	113	109	...	167
11	*los*	4666	95	149	...	126
12	*com*	4379	94	136	...	56
13	*ab*	4342	85	133	...	97
14	*les*	3813	87	131	...	98
15	*d*	3702	66	113	...	81
16	*tirant*	2913	0	0	...	55
17	*li*	2680	50	71	...	51
18	*qui*	2615	51	81	...	69
19	*del*	2430	36	73	...	74
...	...	...	...	...	...	...
13828	*zels*	1	0	0	...	0

***Taula 3.2**: Exemple de la base de dades on presentem la freqüència total per blocs de capítols de les paraules. Cada bloc està format per 14 capítols exepte l'últim que en conté 13.*

A partir d'ara ens centrem en les dades referents als capítols del *Tirant lo Blanc*, ja que no és fins al capítol 9 on comencem a parlar dels blocs de capítols.

A partir del conjunt de dades representat en la taula 3.1, se'n deriven les dades que farem servir per estimar els paràmetres de la distribució amb la qual treballem i expliquem amb tot detall en el capítol 4. Com ja veurem, es tracta d'obtenir informació sobre el nombre d'aparicions de cada paraula en cadascun dels capítols del *Tirant lo Blanc*. Així doncs, tenint en compte la divisió del llibre per capítols, observem que la paraula més freqüent apareix un total de 354 vegades en el capítol 189.

Això ens porta a definir la base de dades que ens serà útil per estimar els paràmetres de la nostra distribució de vocabulari. Aquestes dades consten d'una primera columna enumerada de l'1 fins al 489 que ens indica el número de capítol i 354 columnes més que ens diuen el nombre de paraules que apareixen repetides r vegades, des de r=1 fins a r=354, que corresponen al comptatge del nombre de paraules en cada una de les aparicions dels capítols del llibre, és a dir, la seqüència V_1, V_2, V_3, ..., V_{354}. Per entendre millor aquestes dades en presentem una part en la taula 3.3. També hi mostrem la llargada dels capítols N, per fer èmfasi en els variables tamanys d'aquests.

Per exemple, observem que en el primer capítol tenim 107 paraules que apareixen una vegada, 16 que es repeteixen dues vegades, 6 que es repeteixen tres vegades i així successivament. La paraula que es repeteix més, com ja hem dit, es troba en el capítol 189 i surt un total de 354 vegades.

Cap	N_i	r																
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	...
1	255	107	16	6	2	2	2	2	1	1	1	0	1	0	0	1	0	0	...
2	476	172	26	19	7	2	2	2	2	1	1	1	1	0	0	0	0	0	...
3	1174	299	70	32	16	10	5	4	2	5	1	2	0	3	0	1	0	0	...
4	670	205	52	20	7	10	3	2	2	1	0	1	1	2	0	0	0	0	...
5	1089	302	54	27	18	7	4	4	1	1	1	1	2	1	0	0	3	1	...
6	615	238	37	18	6	2	2	1	1	0	1	2	0	1	0	1	1	0	...
7	283	123	20	7	3	3	1	0	0	0	1	0	0	0	0	0	1	0	...
8	237	97	11	9	4	1	3	0	1	2	0	0	0	2	0	0	0	0	...
9	223	111	16	2	6	0	1	2	0	2	0	0	1	0	0	0	0	0	...
10	867	275	67	15	12	8	4	2	1	0	1	2	1	1	1	1	1	0	...
11	579	156	47	12	5	4	4	1	2	2	1	2	1	0	0	0	0	0	...
12	404	134	27	13	7	7	3	1	3	1	1	0	0	2	0	0	0	0	...
13	1085	281	56	33	15	9	2	2	2	2	0	0	1	3	0	0	2	2	...
14	485	167	27	5	8	5	2	2	0	3	1	0	0	1	0	2	1	0	...
15	289	120	13	7	5	1	5	0	1	2	1	0	1	0	0	0	0	0	...
16	680	191	45	20	8	6	3	0	2	6	0	3	0	2	0	0	0	0	...
17	566	176	41	13	8	3	3	1	2	1	2	0	2	0	1	1	0	0	...
18	1367	321	91	26	17	13	8	6	3	0	3	2	0	0	0	1	0	0	...
19	539	222	34	12	4	2	4	5	0	0	0	0	0	1	1	2	0	1	...
20	590	191	35	18	11	3	1	4	2	2	1	1	0	0	0	0	3	0	...
21	719	107	52	17	8	6	4	3	2	1	1	0	1	0	3	0	0	0	...
22	892	172	63	26	13	8	2	5	2	2	0	1	2	0	4	1	0	0	...
23	682	299	38	16	15	5	4	3	0	0	1	1	3	1	0	2	0	1	...
24	987	205	50	24	11	9	7	5	1	3	3	3	1	0	2	2	0	0	...
25	1401	302	78	33	22	13	5	6	3	5	1	1	1	2	2	1	1	2	...

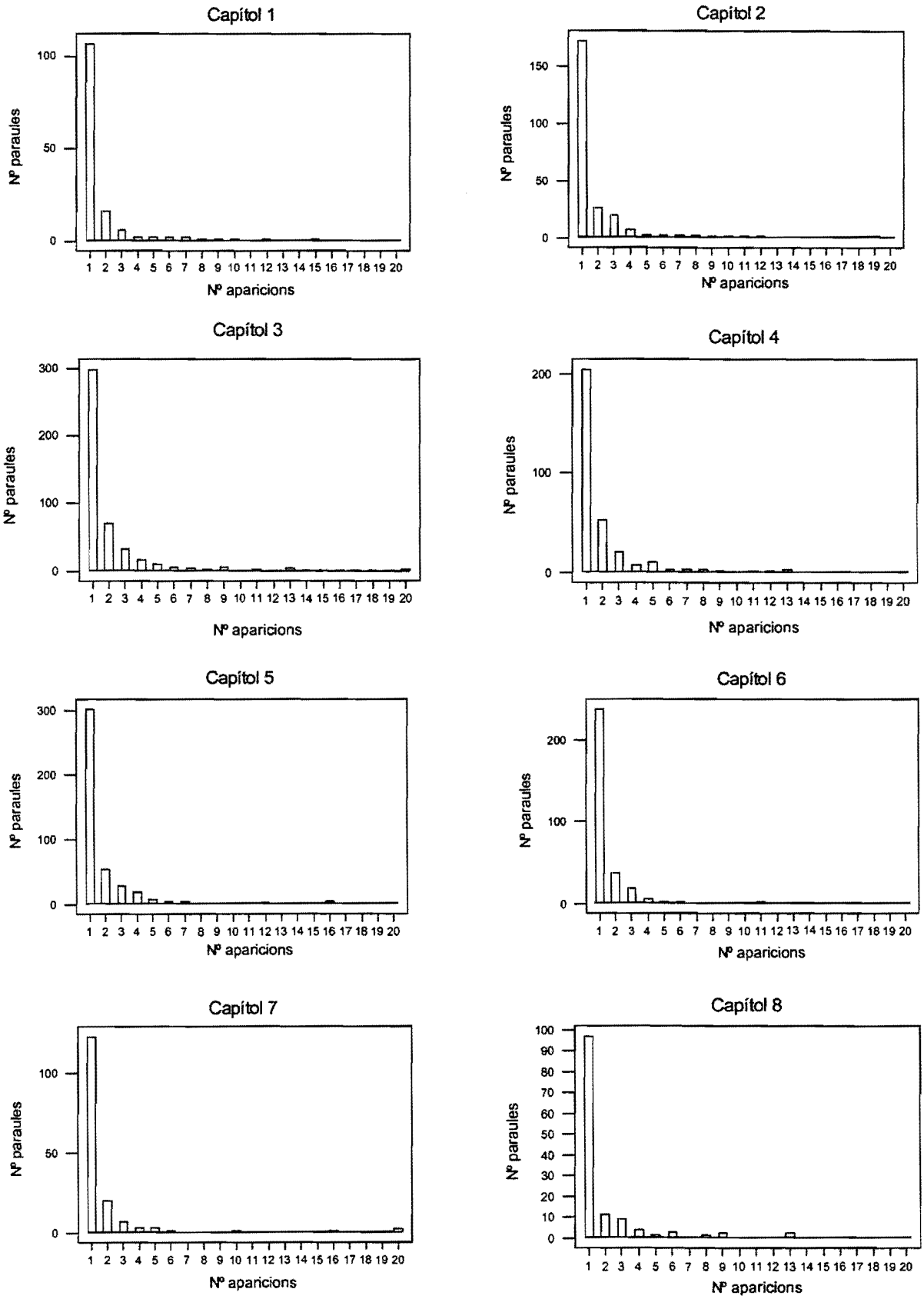
***Taula 3.3**: Exemple de la base de dades on presentem la seqüència V_r per r=1 fins a r=17 dels 25 primers capítols del Tirant lo Blanc.*

3.2 Exploració de les dades

Així doncs, com hem esmentat anteriorment, les dades que analitzem són la seqüència V_1, V_2, V_3, ..., V_r en la que V_r és el nombre de paraules diferents que apareixen r vegades en un text. Per exemple, V_1 ens indica el nombre de paraules que apareixen només una vegada, V_2 ens indica el nombre de paraules que es repeteixen dues vegades i així successivament.

Dels 489 capítols originals, s'han eliminat una sèrie de capítols que no utilitzarem perquè tot el capítol està escrit en cursiva, quedant-nos finalment amb 470. Així doncs, obtenim una matriu de dades amb una primera columna que conté el nombre d'aparicions de les paraules, r, ordenat de l'1 a 354, i 470 columnes que contenen el nombre de paraules que apareixen r vegades, V_r, dels 470 capítols, que contenen escrit.

Els valors de la matriu de dades disminueixen en augmentar el nombre de repeticions (hi ha moltes paraules que no es repeteixen i poques que es repeteixen molt sovint), això fa que a partir de les files 15 i 20, aproximadament, la nostra base de dades contingui molts zeros i uns. A continuació presentem uns gràfics on podem veure aquest fet clarament reflectit, són els gràfics 3.1. Observem que, com és d'esperar, a mesura que augmenten el nombre de repeticions, disminueix el nombre de paraules. Ho mostrem per els vuit primers capítols del *Tirant lo Blanc*, en tots ells veiem que el número de paraules a partir de la decena aparició rarament és més gran que dos.



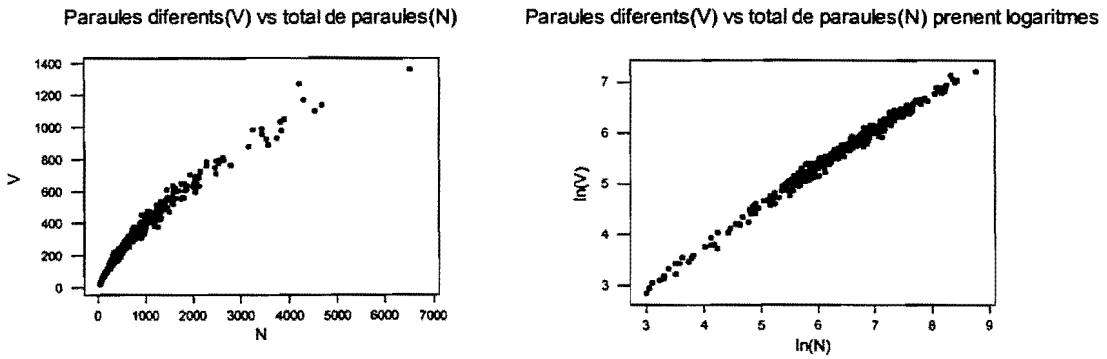
Gràfics 3.1: *es presenten la freqüència de paraules en funció del nombre d'aparicions en un text. Representem gràficament, la freqüència dels primers vuit capítols.*

A partir d'aquestes dades hem calculat el nombre de paraules diferents V, el nombre total de paraules que s'escriuen, N, i el nombre de paraules diferents que es repeteixen r vegades, V_r, per cadascun dels capítols. Observem que V i N són tals que:

$$V = \sum_r V_r$$

$$N = \sum_r r * V_r$$

Hi ha una relació lineal molt clara entre el logaritme de les diferents paraules (ln(V)) que apareixen en el text i el logaritme de totes les paraules (ln(N)). Si observem els resultats de la regressió lineal ens n'adonem, que una explica el 99% de la variabilitat de l'altre. Ho veiem en els gràfics 3.2.



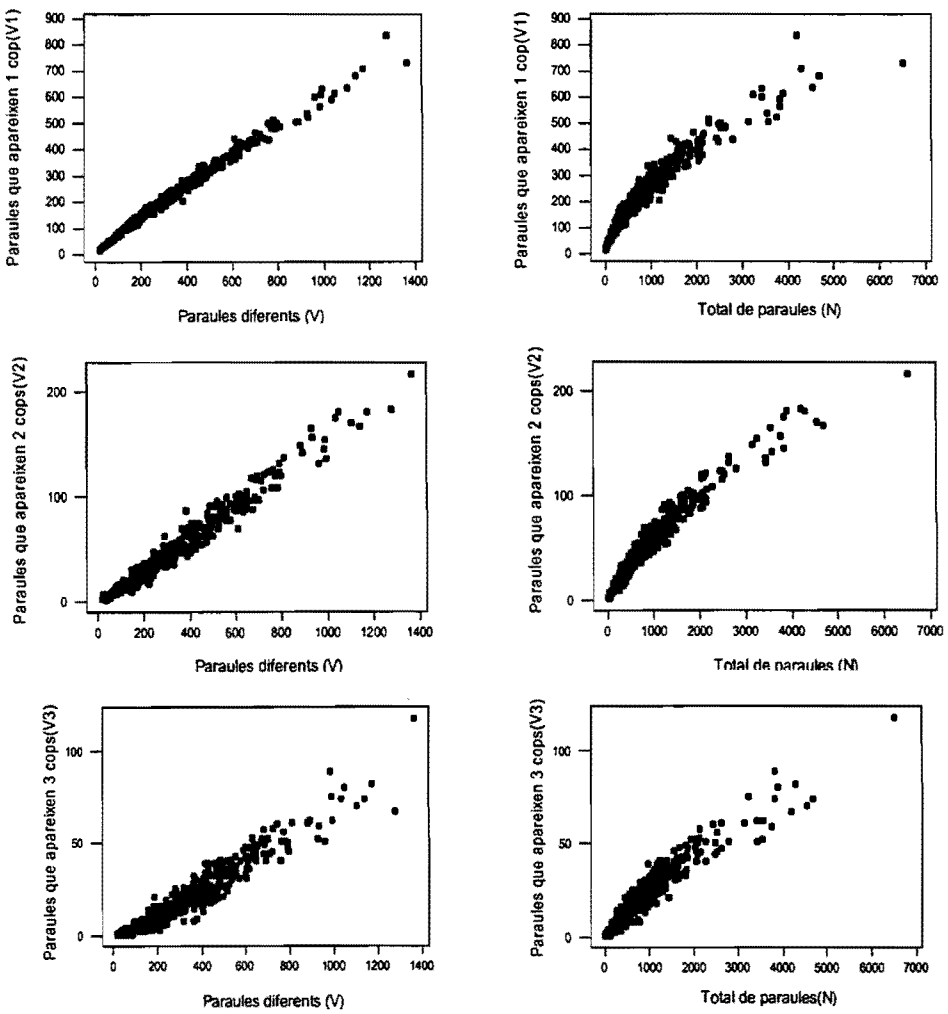
ln(V) = 0,854 + 0,745 ln(N)

	Coeficient	D.S	T	P
Constant	0,85	0,025	34,43	0,000
ln(N)	0,74	0,004	192,60	0,000

S = 0.08178 R-Sq = 98.8% R-Sq(adj) = 98.8%

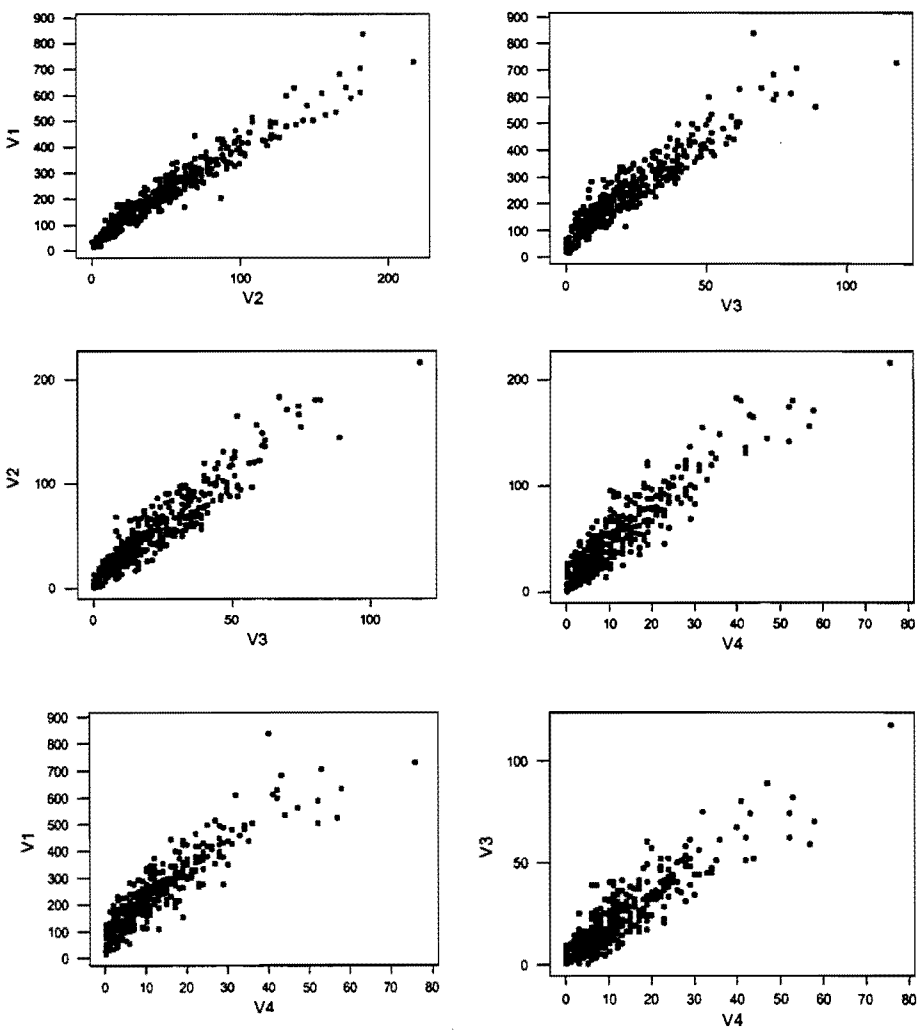
***Gràfics 3.2**: A dalt a l'estquerra, representem el nombre de paraules diferents (V) en funció del total de paraules que surten en el text (N), a la dreta, el logaritme neperià de V en funció de N. I finalment, el model lineal del logaritme de V en funció de V.*

Per altra banda, el nombre de paraules diferents V està directament correlacionada amb el nombre de paraules que només apareixen una vegada en un capítol (la correlació de Pearson és 0,949). Fet totalment coherent, si tenim en compte que el nombre de paraules diferents que apareixen en un capítol V s'ha calculat a partir de la suma de les V_r's, i que V_1 és la variable que té més pes sobre V, explicant un 98.7 % de la variabilitat. Tanmateix la relació entre V i V_1 no és lineal.



Gràfics 3.3: *es representen el nombre de paraules que apareixen un, dos i tres cops (V1, V2 i V3 respectivament), en funció de les paraules diferents (V), i el total de paraules (N).*

En els gràfics 3.3 observem que a mesura que augmenta r, la relació entre V_r i V cada cop és menys evident. Això és degut, tal i com hem vist amb tota claredat en els gràfics 3.1, a què el nombre de paraules que apareixen una sola vegada en el text és molt més elevat que les que hi apareixen dues, tres, quatre, etc vegades. De la mateixa manera, en els gràfics 3.4, on mostrem les relacions entre V1, V2, V3 i V4, també s'aprecia clarament aquest fet.



Gràfics 3.4: *es representen les relacions entre el nombre de paraules que apareixen un, dos, tres i quatre cops (V_1, V_2, V_3 i V_4).*

Més endavant, en el capítol 7, veurem que aquests índexos ens poden ser útils per quantificar la riquesa del vocabulari d'un autor. És molt lògic pensar que com més gran és V_1, és a dir, el número de paraules que apareixen només una vegada en un text, més ric i divers és aquest text. Però ens interesserà, com veurem, trobar índexos de riquesa que no depenguin de N, ja que si depenen de N aquesta riquesa es pot confondre amb el nombre total de paraules del text.

CAPÍTOL 4:

Distribucions de vocabulari

Fa temps que s'intenta representar la freqüència de les paraules, V_r, que apareixen en un text r vegades mitjançant diferents tipus de distribucions. En aquest capítol mostrem diferents opcions d'ajust de distribucions de vocabulari. Ens centrarem de ple en la distribució de Sichel, que és la que ajustem en les nostres dades del *Tirant Lo Blanc*.

4.1 Algunes distribucions proposades

Zipf (1932) va ser el primer en trobar que existeix una relació entre el nombre d'ocurrències r i les seves freqüències V_r. Proposa la relació:

$$V_r * r = k,$$

sent k una constant. Doncs la distribució de probabilitats per r basada en aquesta relació és :

$$p_r = \frac{k}{N} * \frac{1}{V_r^2},$$

que indica la probabilitat de que una paraula aparegui r vegades en un text de llargada N. Aquesta llei es coneix com la *llei de Zipf* i ha estat criticada, entre d'altres,
per Herdan (1966), que argumenta que l'ajust és molt dolent en la cua de la distribució, que es correspon amb les paraules més freqüents.

Yule (1944) va dir que la distribució correcta per la distribució de les freqüències de paraules hauria de ser una barreja de distribucions de Poisson, però mai ho va provar. Un ajust bo es pot obtenir utilitzant el que en la literatura estadística es coneix com la distribució de Waring. En reconeixement al treball fet per Herdan (1964) en aplicar aquesta distribució a la freqüència d'ús de paraules, la distribució també s'ha conegut en la literatura lingüística com el model de Waring-Herdan.

La distribució de probabilitats de Waring-Herdan ve donada per:

$$p_r = \frac{(x-a)a(a+1)...(a+r-2)}{x(x+1)(x+2)...(x+r-1)},$$

on els paràmetres x i a caracteritzen la llargada del text, l'abast del vocabulari emprat i la riquesa del vocabulari. Si V és el número de paraules diferents, el nombre esperat de paraules que apareixen r vegades és $V*p_d$. Herdan va proposar com a estimadors:

$$a = \left[\frac{V}{V-V_1} - \frac{V}{N} - 1\right]^{-1},$$

$$x = \frac{aV}{V-V_1},$$

on N és la llargada del text i V_1 és el número de paraules que només apareixen una vegada.

Muller (1969) mostra com aquest model és raonablement bo per texts de llargada $1000<N<100.000$, tot i que Herdan va cometre un error en els seus càlculs que va ser demostrat per Dolphin (veure Muller, 1975), el qual va corregir l'estimador d'a per:

$$a = \left[1 - \frac{V_1}{V}\right] * \left[\frac{N}{V} - 1\right] * \left[\frac{NV_1}{V^2} - 1\right]^{-1}.$$

L'error de Herdan va donar a Muller i Dolphin la idea d'avaluar l'abast total del lèxic d'un autor, calculant el nombre de paraules que l'autor coneix i que no surten en el text, és a dir V_0. Les lleis de Waring-Herdan i Dolphin-Muller han estat considerades com els millors models existents per ajustar corbes de vocabulari.

Efron i Thisted (1976) opta per un model empirical-bayes paramètric i un de no paramètric per determinar quantes paraules coneixia Shakespeare, és a dir, quin era V_0 de l'escriptor. Al 1987 examinen la consistència en l'ús de paraules d'un poema recentment descobert i atribuït a Shakespeare amb el corpus bibliogràfic del mateix Shakespeare, usant un model empirical-bayes no paramètric. També analitzen poemes de poetes elisabetians (Jonson, Marlowe i Donne) i quatre poemes atribuïts sense dubtes al mateix Shakespeare.

4.2 Discució sobre la distribució de les dades

Les dades de les quals disposem quantifiquen el nombre de paraules que surten r vegades. El nostre objectiu és trobar la llei de probabilitat que s'ajusti bé al nombre d'aparicions d'una paraula que coneix l'autor. Com ja hem esmentat anteriorment, la quantitat de paraules decreix a mesura que augmenten les repeticions. Hi ha moltes paraules que no es repeteixen i poques que es repeteixen molt sovint. Així doncs, tenim una distribució decreixent, deixant una cua molt llarga en augmentar el nombre de repeticions. En aquest moment se'ns planteja un problema: cap de les distribucions que hem vist fins ara s'ajusten bé.

Es tracta de descriure la llei de probabilitat pel nombre d'aparicions d'una paraula que coneix l'autor. S'intueix que la llei que busquem es relacionarà amb una distribució binomial, que tindrà com a paràmetres N, que serà el nombre de paraules d'un text i π la probabilitat que una paraula s'utilitzi.

Per entendre-ho millor, imaginem que tenim una caixa amb V_a paraules i que cada vegada que hem d'escriure una paraula, triem a l'atzar una de les paraules que hi ha a la caixa i

posteriorment la retornem, així doncs, repetim l'experiment N vegades, tantes vegades com paraules hem d'escriure. La funció de probabilitat es podria escriure de la següent forma:

$$\phi(r|N) = \frac{N!}{r!(N-r)!}\pi^r(1-\pi)^{N-r}, \qquad (4.1)$$

En aquest cas, però, el paràmetre π no és constant, perquè cadascuna de les paraules té una probabilitat de ser utilitzada diferent, que l'anomenem π_i. Així doncs, si el vocabulari de l'autor és V_a, tenim V_a pis diferents: $\pi_1,\pi_2,\pi_3,...,\pi_{Va}$, on

$$\sum_{i=1}^{Va}\pi_i = 1$$

$$0<\pi_i<1$$

Ens trobem en el cas d'una barreja de distribucions, que anomenem distribució mixtura. Podríem ajustar aquestes dades per una distribució binomial d'un únic paràmetre que no és constant i que seguirà una distribució de probabilitat. Denotem doncs, per $p(r,\pi)$ la distribució inicial, i per $\varphi(\pi)$ la distribució del paràmetre. La distribució mixtura es pot escriure:

$$prob(Mixtura) = C\int_0^1 \pi^r(1-\pi)^{N-r}\psi(\pi)d\pi \quad , (4.2)$$

Pel què intuïm, hem de recòrrer a noves distribucions plantejades per a resoldre dades amb aquestes característiques. Sichel va proposar, l'any 1975, una nova distribució que calcula aquesta probabilitat, que una paraula surti r vegades. Aquesta distribució ha estat desenvolupada i provada al llarg de molts anys, obtenint resultats satisfactoris. Aquest fet la converteix en una distribució candidata per ajustar-se a aquest tipus de dades. Es tracta d'una barreja de distribucions, en la qual, la distribució inicial és una poisson de paràmetre $N\pi$, i la distribució del paràmetre π és una Gaussiana inversa generalitzada, i per tant, es pot escriure com:

$$\phi(r|N) \approx \frac{N^r}{r!} \int_0^1 \pi^r e^{-N\pi} \psi(\pi) d\pi \quad , \qquad (4.3)$$

El què fa Sichel, és aproximar la distribució binomial (4.2), per una distribució Poisson de paràmetre $N\pi$, amb la justificació que prové d'una població amb un paràmetre π molt petit, i un tamany de mostra N prou gran, amb $N\pi$ finit i definir una distribució pel paràmetre π. Fins aquí tot és conegut, en l'apartat 4.3 d'aquest capítol s'explica amb tot detall la distribució de la mixtura, i més concretament, la distribució Gaussiana inversa generalitzada del paràmetre π.

4.3 Distribució de Sichel

Sichel (1975) reprén la idea de Yule d'una barreja de Poissons i proposa una nova família de distribucions com a model per a les freqüències d'ús de paraules. Sichel proposa que la distribució del paràmetre π , que s'ha introduït en el punt 4.2, sigui una Gausiana inversa. Així doncs, la distribució Sichel es calcula a partir d'una barreja de distribucions, en el qual la distribució inicial és una poisson de paràmetre $N\pi$, i la distribució del paràmetre π és una Gaussiana inversa. El paràmetre π és la probabilitat que s'utilitzi una paraula, així doncs, si el vocabulari de l'autor consta de Va paraules, tenim v_a pis diferents: $\pi_1,\pi_2,\pi_3,...,\pi_{Va}$, i compleix que:

$$\sum_{i=1}^{Va} \pi_i = 1 \quad ,$$

amb $0<\pi_i<1$. Sabent això, la distribució de mixtura es pot escriure:

$$\Phi(r|N) \approx \frac{N^r}{r!} \int_0^1 \pi^r e^{-N\pi} \Psi(\pi) d\pi \quad , \qquad (4.4)$$

on $\Psi(\pi)$ és la inversa-gaussiana generalitzada que depén de tres paràmetres b,c i γ:

$$\Psi(\pi) = \frac{(2/bc)^{\gamma}}{2K_{\gamma}(b)} \pi^{\gamma-1} \exp(-\frac{\pi}{c} - \frac{b^2 c}{4\pi}) \quad , \quad (4.5)$$

on el domini de b,c i γ és $: -\infty < \gamma < \infty, \quad b \geq 0, c > 0$ i $0 < x < \infty$. $K_{\gamma}(b)$ és la funció recursiva modificada de Bessel de segon ordre γ i argument b. Cal dir que per b igual a 0, la distribució és una Gamma i quan $c \rightarrow \infty$ és una distribució Pearson del tipus *V*.

Els moments de primer i segon ordre del paràmetre π són:

$$\mu_1(\pi) = E(\pi) = \left(\frac{bc}{2}\right) \frac{K_{\gamma+1}(b)}{K_{\gamma}(b)}, \qquad (4.6)$$

$$\mu_2(\pi) = E(\pi^2) = \left(\frac{bc}{2}\right)^2 \frac{K_{\gamma+2}(b)}{K_{\gamma}(b)}, \qquad (4.7)$$

El quocient entre les funcions bessel $K_{\gamma+1}(b)/K_{\gamma}(b)$, és més petit que 1, i per valors de b molt petits, el quocient tendeix a b. A més a més, recordant que:

$$\sum_{i=1}^{Va} \pi_i = 1 \ ,$$

deduïm que el valor esperat ha de ser molt petit, perquè és lògic que un autor coneix moltes paraules, i les probabilitats s'han de repartir en el domini [0,1]. Cal dir també, que Sichel afirma, en l'article escrit l'any 1985, que el valor teòric més alt d'aquest paràmetre és 0.07. L'esperança també es pot escriure:

$$E(\pi) \approx \frac{1}{V_a} \sum_{i=1}^{Va} \pi_i = V_a^{-1}$$

Tornant a la distribució mixtura que descriu Sichel i resolent la integral 4.4, s'obté la funció densitat d'aquesta distribució, que es defineix:

$$\Phi(r|N) = \left[(1+cN)^{\gamma/2} K_\gamma(b)\right]^{-1} \frac{1}{r!}\left(\frac{bcN}{2\{1+cN\}^{1/2}}\right) k_{r+\gamma}\left(b\{1+cN\}^{1/2}\right), \quad (4.8)$$

on r=0,1,2,3...,∞. $\Phi(r|N)$ indica la probabilitat que una paraula aparegui r vegades en un text de llargada N. Aquest model es pot simplificar, si utilitzem la següent reparemetrització dels paràmetres b i c:

$$\alpha = b\{1+cN\}^{1/2},$$

$$\theta = \frac{cN}{1+cN},$$

amb $\alpha > 0$ i $0 < \theta < 1$. Com a conseqüència de la reparametrització, s'obté la funció densitat de la distribució Sichel següent:

$$\Phi(r|N) = \frac{(1-\theta)^{\gamma/2}}{k_\gamma\left(\alpha\{1-\theta\}^{1/2}\right)} \frac{\left(\frac{1}{2}\alpha\theta\right)^r}{r!} k_{r+\gamma}(\alpha), \quad (4.9)$$

per r=0,1,2,...,∞.

Observi's que els nous paràmetres α i θ depenen de la llargada del text N. A valors alts de N, li corresponen valors alts de α i θ, per altra banda, θ no serà més gran que 1.

Recordem que $\phi(r|N)$, representa la probabilitat que qualsevol paraula es repeteixi r vegades, i per tant $\phi(0|N)$ és la probabilitat que una paraula no s'utilitzi en el text i que forma part del vocabulari de l'autor, així doncs, en la nostra mostra no es pot observar aquesta probabilitat.

Això fa que el domini observat de les diferents aparicions de paraules sigui $r=1,2,..,\infty$.

Hem de calcular la probabilitat, condicionant pel fet que el domini de la població està truncat en el 0.

$$\phi(r|N)_{r=1..\infty} = \frac{\phi(r|N)}{\sum_{i=1}^{\infty}\phi(i|N)} = \frac{\phi(r|N)}{1-\phi(0|N)},$$

la probabilitat del 0 es reparteix entre r=1,2,3,...,∞ i d'aquesta manera obtenim la distribució Sichel truncada en r igual a 1:

$$\Phi(r|N) = \left[(1-\theta)^{-\gamma/2} k_j\left(\alpha\{1-\theta\}^{1/2}\right) - k_j(\alpha)\right]^{-1} \frac{(0.5\alpha\theta)^r}{r!} k_{+\gamma}(\alpha), \qquad (4.10)$$

per r=1,2,...,∞

Sichel ha demostrat que en molts casos, encara que no sempre és així, la distribució de la freqüència en l'aparició de paraules està ben ajustada, si fixem el paràmetre γ igual a –1/2. Saben que la funció bessel és simètrica i que la imatge de -1/2 és la següent:

$$K_{-1/2}(z) = K_{-1/2}(z) = \left[\frac{\pi}{2z}\right]^{1/2} \exp(-z),$$

se'ns simplifica el model de la distribució obtenint el següent:

$$\Phi(r|N) = \frac{\left(2\alpha/\pi\right)^{1/2}\exp(\alpha)}{\exp\left(\alpha\left[1-(1-\theta)^{1/2}\right]\right)-1} \frac{(0.5\alpha\theta)^r}{r!} k_{r-1/2}(\alpha), \qquad (4.11)$$

per r=1,2,...,∞ (Cas truncat)

$$\Phi(r|N)=\left(2\alpha/\pi\right)^{1/2}\exp(\alpha(1-\theta)^{\frac{1}{2}}\frac{(0.5\alpha\theta)^r}{r!}k_{r-1/2}(\alpha),$$

per r=0,2,...,∞ (Cas no truncat)

Ara passem a estudiar els moments de la distribució mixtura. Recordem que es tracta d'una barreja de distribucions, en la qual la distribució inicial és una Poisson de paràmetre $N\pi$, i la distribució del paràmetre π és una Gaussiana inversa generalitzada. A partir dels moments definits per la Gaussiana inversa 4.6 i 4.7, i tenint en compte que la Gaussiana inversa es barreja amb una Poisson positiva, és a dir, truncada en el 0, s'obtenen les següents equacions que ens defineixen el moment de primer i segon ordre:

$$\mu_1(r|N=E(r|N)=\frac{1}{2}bcN\left[1-\exp(-b(\{1+cN\}^{1/2}-1))\right]^{-1}, \quad (4.12)$$

$$\mu_2(r|N)=E(r^2|N)=\frac{1}{2}bcN\left[1+\frac{1}{2}(1+b)cN\right]\left[1-\exp\left(-b\left(\{1+cN\}^{1/2}-1\right)\right)\right]^{-1}, \quad (4.13)$$

Aquestes equacions les utilitzarem a l'hora d'estimar els paràmetres de la distribució pel mètode dels moments.

L'objecte del nostre estudi és determinar un canvi d'estil, i per fer-ho ens serà útil, ajustar la distribució de vocabulari als capítols del *Tirant,* per determinar si els seus paràmetres canvien en algun punt del llibre o no. La distribució que proposa Sichel és candidata per ajutar-s'hi. Ara bé, els paràmetres de la mixtura, α i θ, no són coneguts i per tant s'hauran d'estimar. Per fer-ho, emprem diferents mètodes d'estimació, que s'expliquen amb detall en el punt 5.1 del projecte, com també es mostren els resultats obtinguts. Cal fer ressó al perquè només estimarem dos paràmetres, si la distribució en té tres: fixem el paràmetre γ a -1/2, perquè Sichel ja va intuïr i comprovar que la distribució s'ajustava significativament millor a aquest tipus de dades.

CAPÍTOL 5:

Ajust de la distribució de Sichel als capítols del Tirant lo Blanc

En aquest capítol passem a estimar els paràmetres α i θ de la distribució de Sichel per cadascun dels capítols del *Tirant lo Blanc.* Expliquem i implementem quatre mètodes diferents que ens permetran fer-ho. A partir dels estimadors d'α i θ estimarem els paràmetres b i c i en graficarem els resultats. Recordem que el nostre objectiu és determinar possibles punts de canvi en la distribució del vocabulari en la seva riquesa i per tant, una presentació gràfica de la seqüència dels estimadors α, θ, b i c, al llarg de tots els capítols, ens serà molt útil per detectar, tant canvis de variabilitat, com de tendència en els seus valors.

A continuació, passarem a veure gràficament les relacions entre els paràmetres estimats per mètodes diferents, per exemple, relacionarem l'α estimada per el primer mètode amb l'α estimada per el segon.

Abans de passar a treballar directament amb les nostres dades del *Tirant lo Blanc*, hem resolt un exemple fet per Sichel per verificar el bon funcionament de les funcions que hem hagut d'implementar.

5.1 Explicació dels mètodes d'ajust

Per estimar els paràmetres α i θ, de la distribució de Sichel, plantegem el problema en forma d'un sistema de dues equacions i dues incògnites. La resolució del sistema plantejat no és del tot senzilla de resoldre degut a la naturalesa no lineal que té la distribució amb la qual treballem. Cal dir, que les diferents maneres d'estimar els paràmetres que introduïm en aquesta secció no són úniques, és a dir, ens podíem haver plantejat altres mètodes igualant algun valor teòric de la distribució de Sichel al seu estimador d'interés, corresponent.

A continuació presentem i expliquem els mètodes que nosaltres hem utilitzat per estimar α i θ. El primer, que s'explica en la secció 5.1.1, fou proposat per Sichel al 1986 i es basa en la probabilitat d'aparició de paraules que només surten una única vegada en un text de llargada N. En la secció 5.1.2 s'explica el mètode dels moments, en la secció 5.1.3 s'explica el mètode basat en l'estimació dels paràmetres mitjançant la proporció de paraules que surten una vegada i dues vegades en el text i finalment, en la secció 5.1.4 implementem el mètode de la màxima versemblança.

5.1.1 Estimació de α i θ basada en la mitjana i la proporció de paraules que apareixen una vegada

Sichel (1986) planteja un sistema d'equacions no lineals en el que, per una banda, igualem la probabilitat que una paraula aparegui una sola vegada amb el seu valor mostral o observat en el text, i per altra banda, igualem el valor esperat de r vegades amb el seu valor mostral. El sistema de dues no lineals equacions que en resulta és:

$$\Phi(1|N) = \frac{1}{2}\alpha\theta\left[\exp\left(\alpha(1-(1-\theta)^{\frac{1}{2}})\right)\right]^{-1} = \frac{V_1}{V}$$

i:

$$E(r|N) = \frac{\alpha\theta}{2(1-\alpha)^{\frac{1}{2}}}\left[1-\exp\left(-\alpha(1-(1-\theta)^{\frac{1}{2}})\right)\right]^{-1} = \frac{N}{V}$$

on *r* és el número de repeticions d'una certa paraula, V_1 el número de paraules que només surten una vegada en un capítol, *V* el número de paraules diferents i *N* el número de totes les paraules escrites al capitol.

El fet que l'esperança a nivell mostral en la segona equació és *N/V* es justifica amb la següent equació:

$$\overline{X} = \sum_i r * p_r = \sum_i r * \frac{V_r}{V} = \frac{N}{V}$$

on p_r calcula la probabilitat mostral, que una paraula aparegui *r* vegades en un text de llargada *N*.

Per resoldre aquest sistema d'equacions no lineals s'ha executat la rutina *lsqnonlin* de les llibreries del Matlab, que adjuntem en l'Annexe del projecte. Aquesta funció l'hem hagut de cridar 470 vegades a través d'un bucle per tal d'estimar els paràmetres per tots els capítols del *Tirant*. Abans però, hem hagut d'implementar el sistema en una subrutina.

5.1.2 Estimació d' α i θ basada en el mètode dels moments

Aquest mètode està basat en els moments de primer i segon ordre, $\mu_1(r|N)=E(r|N)$ i $\mu_2(r|N)=E(r^2|N)$ respectivament. Per tant, aprofitant la segona equació del mètode anterior obtenim el següent sistema d'equacions no lineals:

$$E(r|N)=\frac{\alpha\theta}{2(1-\alpha)^{\frac{1}{2}}}\left[1-\exp\left(-\alpha(1-(1-\theta)^{\frac{1}{2}})\right)\right]^{-1}=\frac{N}{V}$$

$$E(r^2|N)=\frac{\alpha_N\theta}{2(1-\theta)^{\frac{1}{2}}}\left[1+\frac{1}{2}(1+\alpha(1-\theta)^{\frac{1}{2}})\frac{\theta}{1-\theta}\right]\left[1-\exp\left(-\alpha(1-\theta)^{\frac{1}{2}}\left(\left\{1+\frac{\theta}{1-\theta}\right\}^{\frac{1}{2}}-1\right)\right)\right]^{-1}=\sum_i r^2*\frac{V_r}{V}$$

L'expressió pel moment d'ordre dos mostral es justifica mitjançant:

$$\mu_2(r|N)=\sum_i r^2*\frac{V_r}{V}=\sum_i r^2*p_r$$

Per resoldre aquest sistema d'equacions no lineals, hem escrit la rutina en llenguatge Matlab, que adjuntem en l'Annexe de la memòria. Igual que en el mètode anterior, hem executat la funció *lsqnonlin* que resol sistemes d'equacions no lineals. En aquest cas hem implementat una subrutina diferent corresponent al sistema descrit en aquest mètode d'estimació.

5.1.3 Estimació de α i θ basada en la proporció de paraules que apareixen una i dues vegades en el text

En aquest mètode d'estimació, aprofitem la primera equació del mètode plantejat en l'apartat 5.1.1, en el qual igualem la probabilitat que una paraula aparegui una sola vegada amb la proporció de paraules que apareixen només una vegada al text. La segona equació iguala la proporció de paraules que apareixen dues vegades al text amb la probabilitat teòrica de la distribució $\phi(2|N)$. El sistema plantejat i que s'haurà de resoldre és el següent:

$$\Phi(1|N) = \frac{1}{2}\alpha\theta * \left[\exp\left[\alpha * (1-(1-\theta)^{\frac{1}{2}})\right]-1\right]^{-1} = \frac{V_1}{V}$$

$$\Phi(2|N) = \frac{1}{8}\frac{(\alpha^2\theta^2)*(1-\theta)^{-1}}{\left(1+\frac{\theta}{1-\theta}\right)} * \left[1+\left(\alpha^{-1}(1-\theta)^{\frac{-1}{2}}\right)*\left(1+\frac{\theta}{1-\theta}\right)^{\frac{-1}{2}}\right] * \left[\exp\left[\alpha(1-\theta)^{\frac{1}{2}}\left[\left(1+\frac{\theta}{1+\theta}\right)^{\frac{1}{2}}-1\right]-1\right]\right]^{-1} = \frac{V_2}{V}$$

Anàlogament amb els altres mètodes, per resoldre aquest sistema d'equacions no lineals, hem escrit la rutina en llenguatge Matlab, que adjuntem en l'Annexe de la memòria. Tal i com hem fet en els mètodes que el precedeixen, implementem una rutina que executa la funció *lsqnonlin* en cadascun dels capítols. La funció *lsqnonlin* crida una subrutina que descriu el sistema. En aquest cas, el sistema d'equacions és bastant més llarg degut a la complexitat de les equacions.

5.1.4 Mètode de la màxima versemblança

Aquest mètode passa per calcular la funció de versemblança, $L(\alpha,\theta)$, corresponent a la densitat de Sichel, $\Phi(r|N)$, que depén dels paràmetres a estimar.
En aquest punt considerem la distribució Sichel com una distribució multinomial, que generalitza a la Binomial pel cas de més de dues categories. En el nostre cas, tenim tantes categories com el nombre de repticions de paraules més gran que hi pot haver en un text. Aquesta distribució ens diu, donades r categories, quants resultats s'han observat de cadascuna d'elles, repetint l'esdeveniment de forma independent tantes vegades com paraules diferents utilitza l'autor (V vegades).

Per tant, la funció de versemblança la calculem com a producte dels valors que pren la densitat de Sichel per $r = 1, 2, 3, ..., \infty$. És a dir,

$$L(\alpha,\theta)=\prod_{i=1}^{V_r}\Phi(i)^{V_i}=\Phi(1|N)^{V1}*\Phi(2|N)^{V2}*\Phi(3|N)^{V3}*...*\Phi(r|N)^{Vr},$$

on V_i és el número de paraules que apareixen 1, 2, 3, ...,fins a r vegades en un text de llargada N. Per estimar els paràmetres mitjançant el mètode de la màxima versemblança, procedim a derivar la funció $L(\alpha,\theta)$ respecte α i respecte θ i igualant a 0, les dues expressions. Per tenir una funció més senzilla de derivar, pendrem el logaritme de $L(\alpha,\theta)$ obtenint la següent expressió:

$$l(\alpha,\theta)=\log(L(\alpha,\theta))=\sum_{i=1}^{r}V_i*\log(\Phi(i))$$

Observació:Cal dir que per r gran, en molts casos, el terme no intervé en el càlcul de L,ja que $Vr = 0$. Així doncs, obtenim un sistema d'equacions en funció dels nostres paràmetres derivant $l(\alpha,\theta)$ respecte α i θ, que és el següent:

$$\frac{dl(\alpha,\theta)}{\partial\theta} = \frac{\alpha}{2*(1-\theta)^{1/2}} * \frac{\exp\left[\alpha*\left[1-(1-\theta)^{\frac{1}{2}}\right]\right]}{\exp\left[\alpha*\left[1-(1-\theta)^{\frac{1}{2}}\right]\right]-1} * \frac{\theta*V}{N} = 0$$

$$\frac{dl(\alpha,\theta)}{\partial\alpha} = V*\left[1+\frac{1}{\alpha}-\frac{\exp\left[\alpha*\left(1-(1-\theta)^{\frac{1}{2}}\right)\right]*\left(1-(1-\theta)^{\frac{1}{2}}\right)}{\exp\left[\alpha*\left(1-(1-\theta)^{\frac{1}{2}}\right)\right]-1}\right] - \sum_{r=1}^{r\max} V_r * \wedge(r/\alpha) = 0$$

$$\text{on,}\quad \wedge(r/\alpha) = \frac{K_{r-\frac{3}{2}}(\alpha)}{K_{r-\frac{1}{2}}(\alpha)}$$

i $K_r(\alpha)$ és la funció modificada de Bessel de segon ordre. A l'hora de derivar aquesta funció respecte d'α hem utilitzat la següent fórmula extreta de l'article *Atkinson (1982)*.

$$\frac{dK_\gamma(\alpha)}{\partial\alpha} = -\frac{\gamma}{\alpha} * K_\gamma(\alpha) - K_{\gamma-1}(\alpha)$$

5.2 Ajust a un exemple de Sichel(1975)

Abans de presentar els resultats dels paràmetres estimats per cada capítol del *Tirant lo Blanc*, passarem a resoldre un exemple fet per Sichel per tal de verificar el bon funcionament dels programes que hem hagut d'implementar. Un cop comprovada la correcció dels nostres programes, presentarem els resultats pels quatre mètodes definits en la secció 5.1.

Fins ara, en aquest capítol hem estimat α i θ. És hora de recordar la reparametrització que hem explicat en el capítol anterior. És a dir, no només estimarem els resultats de les estimacions de α i θ pels quatre mètodes, sinó que també mitjançant α i θ estimem les reparametritzacions b i c. Recordem doncs les següents equacions:

$$b = \alpha(1-\theta)^{\frac{1}{2}} \qquad \text{i} \qquad c = \frac{\theta}{(1-\theta)N} \ ,$$

on N és el número de paraules d'un capítol.

5.2.1 Aplicació dels mètodes d'estimació a l'exemple extret de Sichel (1975)

Les dades de l'exemple que utilitzem són una mostra de paraules en anglès de l'obra *Essay on Bacon* de l'escriptor Macaulay. El text té 2048 paraules diferents, entre les quals, n'hi ha una que apareix 255 vegades, i n'hi ha 990 que només surten una vegada. Es vol ajustar la distribució mixtura de Sichel a la freqüència d'aparicions de les paraules. Les dades es representen resumides en 17 categories, tal i com es pot veure en la taula 5.1. De cada categoria s'adjunta el valor observat i el valor esperat per la distribució de Sichel, havent estimat els paràmetres per quatre mètodes diferents, que corresponen als partats 5.1 d'aquest capítol.

En aquest exemple verifiquem els resultats del mètode 5.1.1, basat en el valor esperat i en la proporció de paraules que apareixen una vegada. Així doncs, ens assegurem que els resultats que obtindrem seran correctes i consitents. A més a més, hi adjuntem els ajustos de la distribució amb els valors dels paràmetres estimats pels mètodes restants. En aquestes dades, la distribució de Sichel només ajusta bé, quan els paràmetres s'estimen mitjançant el mètode 5.1.1. Observi's que els valors esperats amb aquests

paràmeters disten poc dels valors observats. Això fa, que el valor de l'estadístic χ^2 sigui molt proper a 1. El valor de l'estadístic χ^2 de discrepància entre valors observats i valors esperats, sobre la distribució estimant els paràmetres mitjançant els mètodes restants 5.1.2,5.1.3 i 5.1.4 és molt més elevat.

Freqüència de les paraules	*Valors observats*	*Valors esperats basats en la mitjana i V1*	*Valors esperats pel mètode dels moments*	*Valors esperats pel mètode basat en V1 i V2*	*Valors esperats per màxima versemblança*
1	990	990.0	1148.3	990.0000	214.4869
2	367	353.0	297.3	338.5590	359.8992
3	173	179.4	145.4	171.7110	417.9772
4	112	110.0	88.9	105.9447	377.9417
5	72	74.9	60.8	72.7948	283.7673
6	47	54.5	44.6	53.4837	184.2508
7	41	41.5	34.2	41.1291	106.3890
8	31	32.7	27.2	32.6904	55.7515
9	34	26.4	22.2	26.6415	26.9271
10	17	21.7	18.4	22.1418	12.1321
11	24	18.1	15.5	18.6949	5.1487
12	19	15.4	13.3	15.9907	2.0744
13	10	13.1	11.5	13.8271	0.7986
14	10	11.3	10.0	12.0669	0.2954
15	13	9.9	8.8	10.6146	0.1055
16-20	31	34.5	31.7	38.1479	0.0545
21-30	31	32.6	32.2	38.6715	0.0002
31	26	29.0	37.7	44.8902	0.0000
Totals	2048	2048.0	2048.0	2048.0	2048.0
χ^2		10.7257	72.8203	21.6854	1.0089e+012
Paràmetres estimats		α =0.4747 θ=0.9671	α =0.0594 θ=0.9775	α =0.3999 θ=0.9772	α =26.1399 θ=0.2473

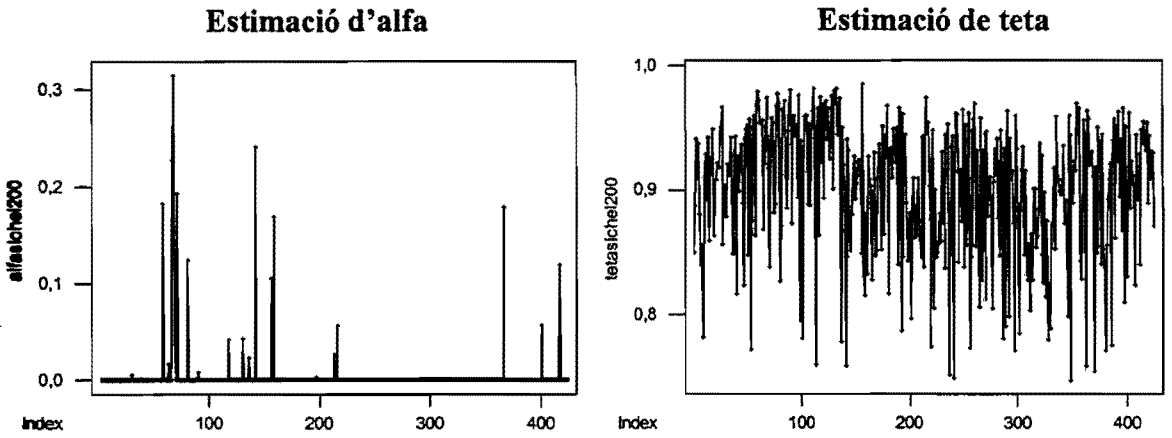
Taula 5.1: *Exemple del càlcul dels paràmetres pels diferents mètodes.*

5.3 Ajust basat en el mètode 5.1.1

En l'annex 5 presentem una taula dels valors estimats per α i θ per cada capítol del *Tirant lo Blanc*, calculats a partir d'aquest primer mètode. En el gràfic 5.1 observem l'evolució d'aquests valors al llarg del llibre.

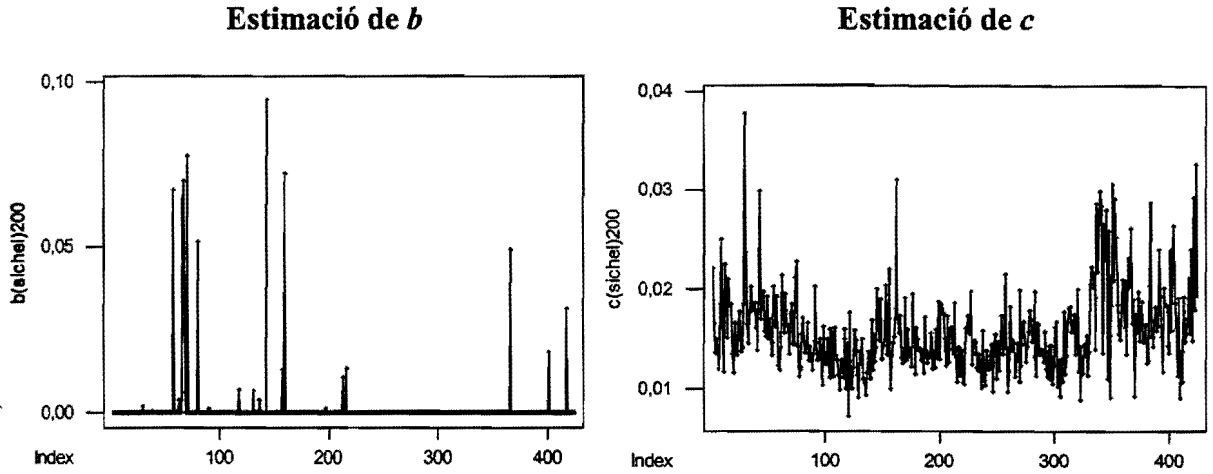
A partir d'ara graficarem els estimadors pels capítols que tenen més de 200 paraules, ja les estimacións d'α dels capítols amb menys de 200 paraules són més grans i no ens deixa veure l'evolució natural de la tendència i la variabilitat de l'estimador (en l'annexe esmentat es pot comprovar aquest fet). Cal dir però, que el fet de veure un canvi de tendència sobre aquests valors no ens determinaran un canvi en els paràmetres de la distribució, perquè és sabut que α i θ depenen de la llargada del text (N).

En l'apartat del gràfic 5.1, que presenta l'evolució dels valors estimats per θ, hi podem veure un canvi de tendència a partir del capítol 140. Les α's són gairebé sempre 0, i no hi podem observar cap canvi, malgrat tot presentem els capítols de més de 200 paraules.



Gràfic 5.1: *Evolució dels paràmetres α i θ estimats en funció de l'ordre dels capitols que tenen més de 200 paraules*

A continuació, graficarem els paràmetres b *i* c que són els paràmetres de base de la distribució barreja i per tant, no depenen de N. Els paràmetres b i c els hem estimat mitjançant α i θ, basant-nos en les fórmules comentades en el capítol 4. Observem que l'estimador b consta d'un producte on hi intervé α, per tant, per tots els capítols que tenen una α igual a 0, el valor de b també ho serà **(veure 5.2)**.



Gràfic 5.2: *Evolució dels paràmetres b i c estimats en funció de l'ordre dels capitols que tenen més de 200 paraules.*

De la mateixa manera que ens passava en l'estimació d'α, en el gràfic 5.2 podem veure que l'estimador b tampoc ens dóna cap informació sobre possibles punts de canvi, ja que el seu valor és gairebé sempre nul. L'estimació del paràmetre c és la més informativa, la seva variabilitat és més baixa que la de θ i s'observen clarament els diferent canvis. Hi ha dos valors de c força més elevats que la resta, el primer és la observació 29 corresponent a un capítol que només té 244 paraules i el segon és la 162 que es correspon a un capítol de 265 paraules.

Deixant de banda aquests dos capítols, observem que la c té tendència a disminuir entre les observacions 60 i 130 i que a partir d'aquí, és manté força estable fins a l'observació 320 aproximadament. A partir d'aquest moment és on podem veure el canvi més significatiu del valor estimat del paràmetre, ja que de cop s'incrementa el seu valor i

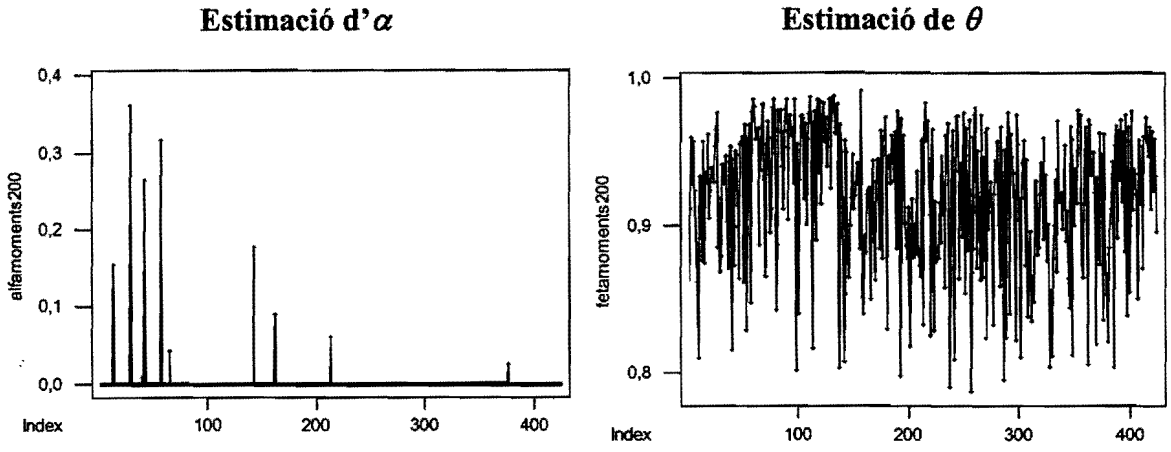
també la seva variabilitat. En les primeres observacions també hi podem veure una variabilitat de c lleugerament superior.

5.4 Ajust basat en el mètode 5.1.2

En l'annexe 5 presentem una taula dels estimadors d'α i θ per cada capítol del *Tirant lo Blanc*, calculats a partir del mètode dels moments.

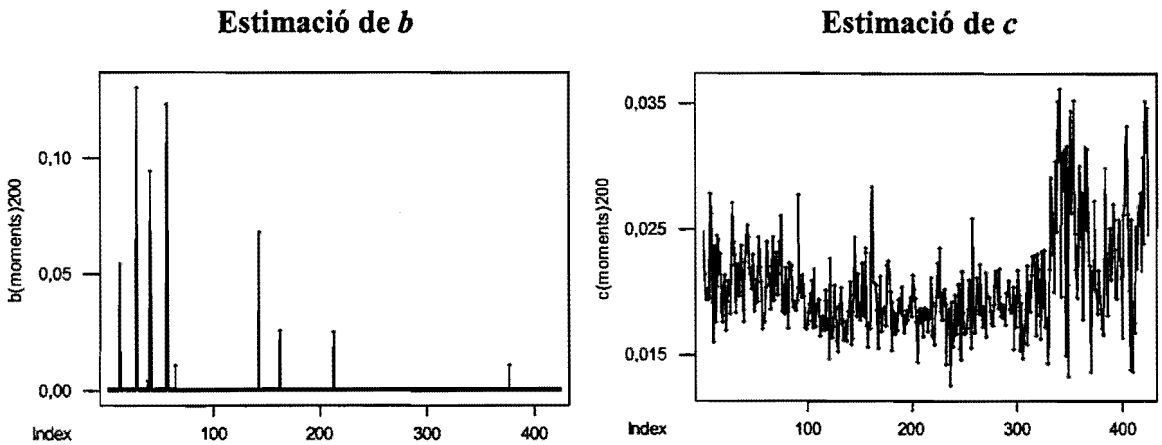
En el gràfic 5.3 observem l'evolució dels estimadors a través dels capitols del llibre. Igual que en la secció 5.2.2 anterior grafiquem els resultats per els capítols que tenen més de 200 paraules.

Tot i haver descartat els capitols de més de 200 paraules, en el gràfic 5.3 podem apreciar que en l'estimació d'α, al ser pràcticament sempre zero, no hi podem observar cap evolució. En canvi en l'estimació de les θ's hi tornem a veure un canvi de tendència a partir de l'observació 140 aproximadament, igual que ens passava en el primer mètode.



Gràfic 5.3: *Evolució dels paràmetres α i θ estimats en funció de l'ordre dels capitols que tenen més de 200 paraules*

Cal dir però, que el fet de veure un canvi de tendència sobre aquests valors no ens determinaran un canvi en la riquesa del vocabulari, perquè és sabut que α i θ depenen de la llargada del text (N). Tot seguit, passem a graficar els valors estimats de b i c a partir dels paràmetres α i θ, que són els paràmetres de base de la distribució de barreja, i per tant, no depenen del tamany del text. Els resultats els podem veure en el gràfic 5.4.



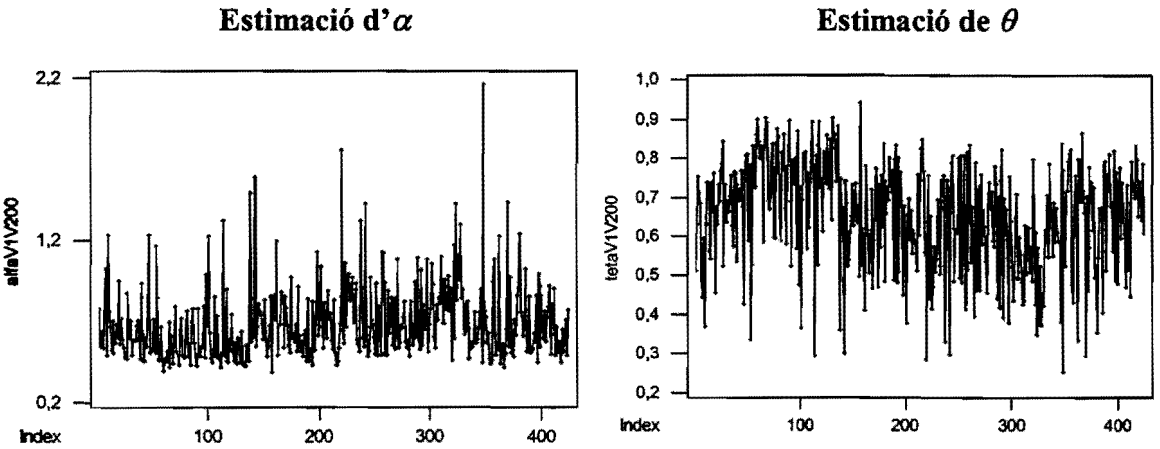
Gràfic 5.4: *Evolució dels paràmetres b i c estimats en funció de l'ordre dels capitols que tenen més de 200 paraules*

En el gràfic 5.4 observem que, coincidint amb l'anterior mètode, el valor estimat per b, tampoc ens dóna informació, perquè és 0.

En l'estimació del paràmetre c veiem, encara més clarament que en el primer mètode, el canvi que es produeix a partir de l'observació 320 aproximadament. La variabilitat és molt més elevada i també sembla que s'incrementa el seu valor. En el tram inicial observem també un petit canvi de tendència entre les observacions 90 i 100. Cal dir que per el mètode dels moments l'estimador c presenta, en general, més variabilitat que per l'anterior .

5.5 Ajust basat en el mètode 5.1.3

En l'annexe 5 presentem una taula dels estimadors d'α i θ per cada capitol del *Tirant lo Blanc*, calculats a partir de la proporció de paraules que surten una i dues vegades en el text.

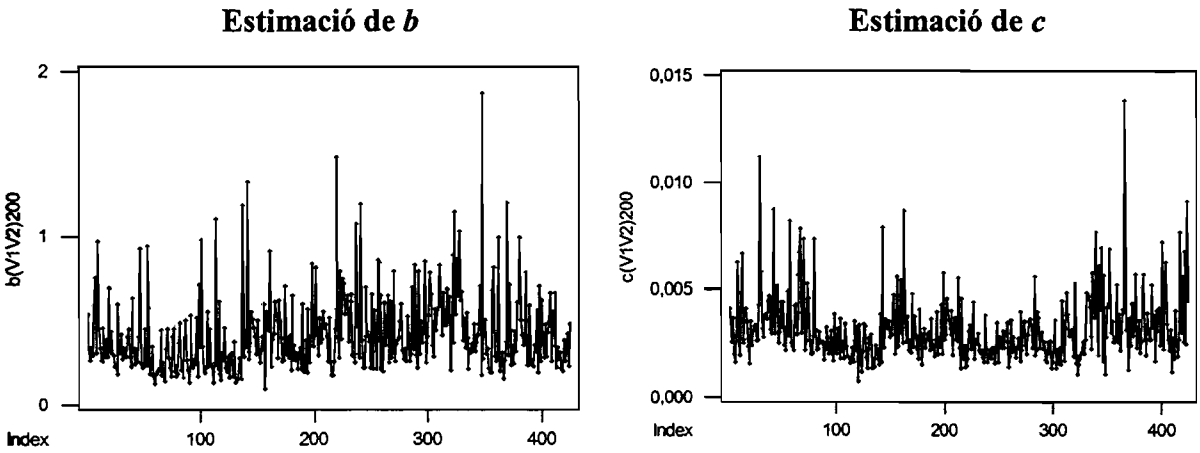


Gràfic 5.5: *Evolució dels paràmetres α i θ estimats en funció de l'ordre dels capitols que tenen més de 200 paraules*

En el gràfic 5.5 observem l'evolució dels estimadors a través dels capitols del llibre. Igual que en les seccions anteriors grafiquem els resultats per els capítols que tenen més de 200 paraules.

En el gràfic 5.5 podem veure el canvi de tendència que ja véiem anteriorment en θ, es troba a partir de l'observació 140 aproximadament. Amb aquest mètode l'estimació d'α ja no té valors nuls, però tot i això continuem sense veure-hi canvis significatius, tret d'un lleuger augment de la variabilitat.

Cal considerar, que veure un canvi de tendència sobre aquests valors no ens determinaran un canvi en els paràmetres de la distribució, perquè és sabut que α i θ depenen de la llargada del text (N). Tot seguit, passem a graficar els valors estimats de b i c mitjançant els paràmetres α i θ. Ho podem veure en el gràfic 5.6.

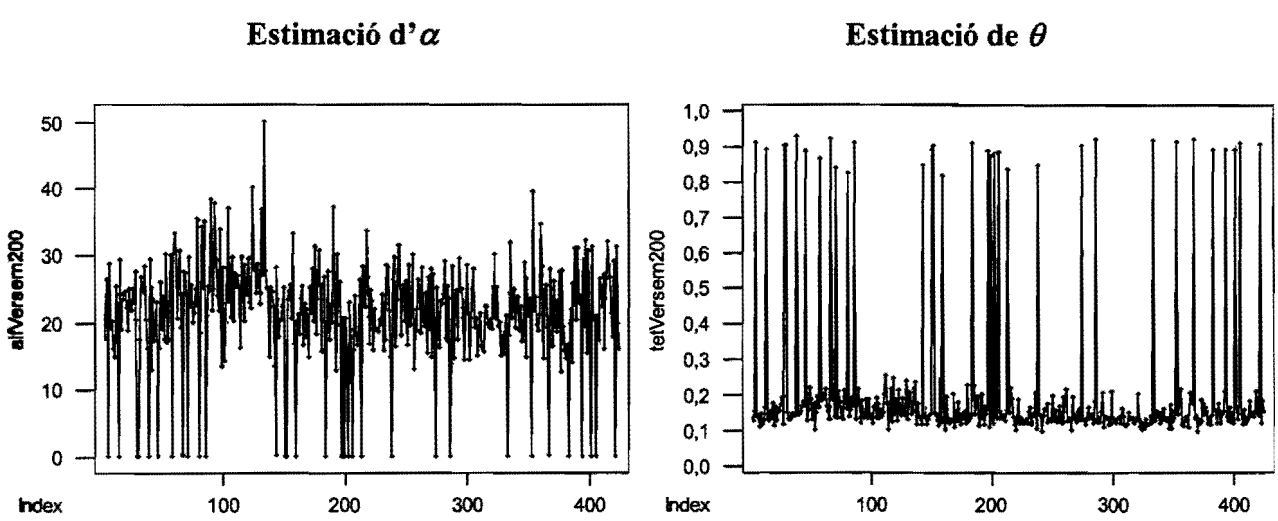


Gràfic 5.6: *Evolució dels paràmetres b i c estimats en funció de l'ordre dels capitols que tenen més de 200 paraules*

Per aquest mètode, l'estimador c presenta molta més variabilitat en les primeres observacions, tal i com véiem en el gràfic 5.6. Però en aquest cas, no s'observa tant clarament el canvi a partir de l'observació 320 que podiem veure tant en el mètode emprat per Sichel com en el dels moments. El paràmetre b té el mateix comportament que l'α però canviat d'escala.

5.6 Ajust basat en el mètode 5.1.4

Pesentem gràficament els estimadors d'α i θ per els capítols del *Tirant lo Blanc* que tenen més de 200 paraules. A l'annex podem veure els valor dels estimadors per tots els capítols del llibre calculats a partir d'aquest mètode.



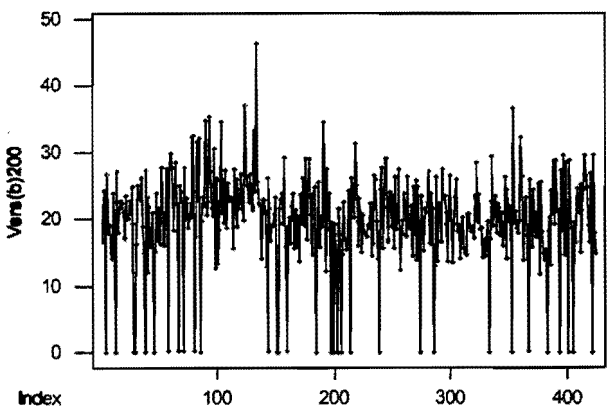
***Gràfic 5.7:** Evolució dels paràmetres α i θ estimats en funció de l'ordre dels capitols que tenen més de 200 paraules*

Per aquest mètode, igual que en l'anterior, els valors d'α no són nuls. A més a més, en el gràfic 5.7 observem un fet curiós, el canvi que observàvem en l'estimació de θ a partir de l'observació 140 aproximadament, en aquest cas l'observem en el paràmetre α. En els resultats gràfics de les θ's no hi podem veure tendència degut a uns 40 capítols aproximadament, els quals tenen un paràmetre molt proper a 1. En qualsevol cas, cal considerar, que veure un canvi de tendència sobre aquests valors no ens determinaran un canvi en els paràmetres de la distribució, perquè és sabut que α i θ depenen de la

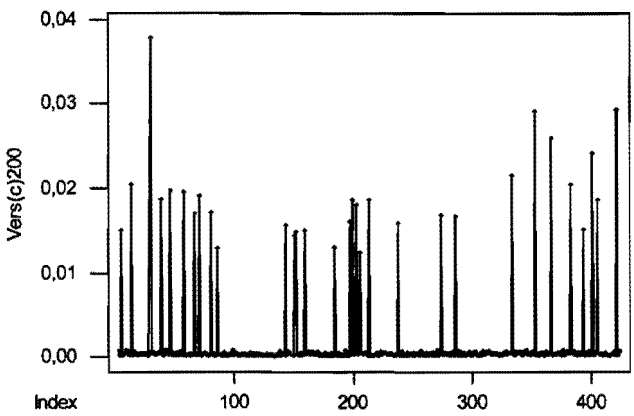
llargada del text (N). Tot seguit, passem a graficar els valors estimats per b *i* c mitjançant els paràmetres α i θ.

En aquest cas, observant el gràfic 5.8, continuem veient el mateix comportament pel valor estimat de b que pel valor estimat de α . En la representació de c, que recordem que fins ara és l'unic paràmetre que ens ha donat pistes sobre una possible diferència entre els últims capítols del llibre i la resta. En aquest cas no hi podem apreciar cap mena de patró ni canvi, a causa de la gran quantitat de valors tant propers a 0.

Estimació de *b*



Estimació de *c*



Gràfic 5.8: *Evolució dels paràmetres b i c estimats en funció de l'ordre dels capitols que tenen més de 200 paraules*

5.7 Exploració dels paràmetres

A continuació passem a explorar i a fer una descripció dels resultats dels paràmetres que hem obtingut, calculant-ne els estadístics més rellevants, com ara la mitjana, la variabilitat, la mediana, el mínim i el màxim. L'objectiu d'això és començar a veure quin és el comportament dels nostres paràmetres, ja que els que tinguin menys variabilitat seran els més útils per poder detectar els possibles canvis d'estil en el *Tirant lo Blanc*. Els resultats els mostrem en la taula 5.2.

En la taula 5.2 observem que l'estimador de la mitjana, tant en la primera metodologia com en el mètode dels moments (5.1.2), és força semblant. En el mètode per la proporció de paraules que surten una vegada i dues vegades en el text (5.1.3), el valor de la mitjana és diferent. També és important veure que l'estimador que té menys variabilitat és el corresponent al paràmetre c en els quatre mètodes utilitzats. Aquest fet ja l'observavem en els gràfics fets en la secció 5.3.

En aquest capítol hem vist que la majoria de les α 's i de les b's eren 0 o molt properes a 0, tant per el primer mètode (5.1.1) com per el mètode dels moments (5.1.2). Per tant, no és d'estranyar que la mediana pels dos primers mètodes sigui 0. Cal dir també que la variabilitat més gran la presenten els paràmetres calculats a partir del tercer mètode (5.1.3) i el mètode de la versemblança (5.1.4), els que presentava el pitjor ajust dels tres.

		DESCRIPTIVA UNIVARIANT DELS ESTIMADORS				
Mètode	**Estimador**	**Mitjana**	**Variab.**	**Mediana**	**Mínim**	**Màxim**
5.1.1	α	0,00509	0,02991	0,00000	0,00000	0,31420
	θ	0,90087	0,05389	0,91180	0,74710	0,98490
	b	0,00156	0,00967	0,00000	0,00000	0,09455
	c	0,01571	0,00424	0,01479	0,00724	0,03781
5.1.2	α	0,00353	0,02928	0,00000	0,00000	0,36010
	θ	0,92013	0,04786	0,93080	0,78690	0,99120
	b	0,00128	0,01080	0,00000	0,00000	0,13053
	c	0,02047	0,00398	0,01954	0,01249	0,03630
5.1.3	α	0,6841	0,2263	0,6225	0,3898	2,1593
	θ	0,65253	0,13819	0,67260	0,25350	0,93920
	b	0,4181	0,2283	0,3552	0,0961	1,8656
	c	0,00313	0,00151	0,00272	0,00070	0,01381
5.1.4	α	20,073	8,858	21,242	0,0000	50,317
	θ	0,21036	0,20525	0,14430	0,05140	0,9279
	b	18,481	8,096	19,707	0,0000	46,455
	c	0,00220	0,00599	0,00030	0,00004	0,03789

Taula 5.2: *Descriptiva univariant dels paràmetres estimats per cadascun dels mètodes.*

Ens intereseria obtenir paràmetres poc correlacionats entre ells i amb poca dependència del tamany mostral, N, així com del número de paraules diferents V. Així doncs, a continuació passem a calcular la correlació de Pearson entre b i c i entre α i θ i a graficar-los respecte N.

Mètode	paràmetres b i c		paràmetres α i θ	
	correlació	p-valor	correlació	p-valor
5.1.1	0,070	0,152	-0,011	0,824
5.1.2	0,019	0,699	-0,135	0,005
5.1.3	-0,178	0,000	-0,894	0,000
5.1.4	-0,722	0,000	-0,646	0,000

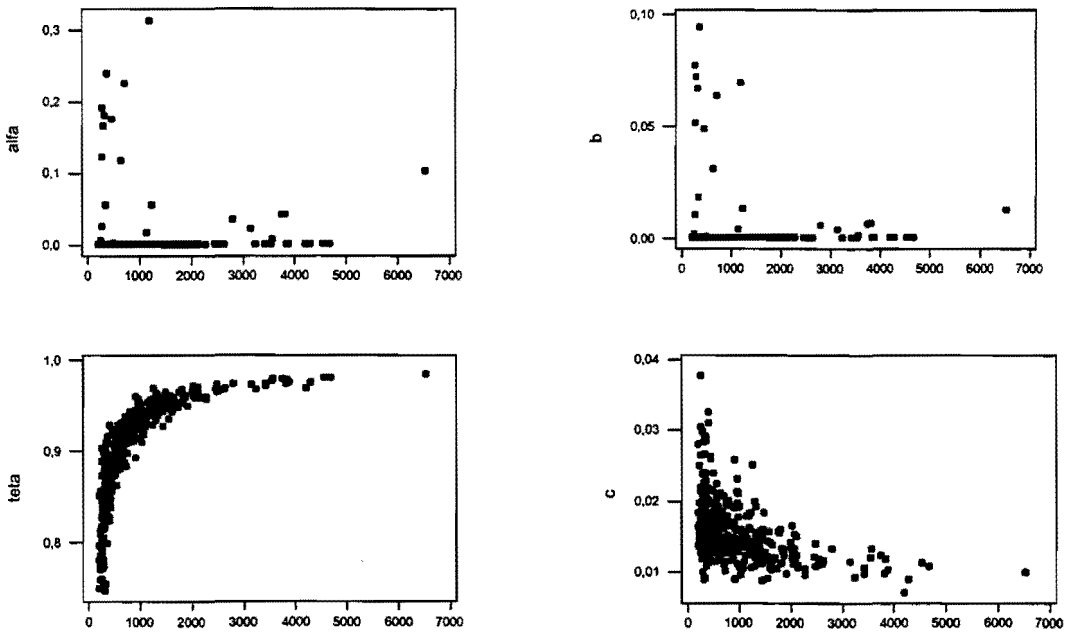
Taula 5.3: *Correlació entre els paràmetres per cadascun dels mètodes*

En la taula 5.3, observem que els paràmetres més poc correlacionats els trobem entre els estimadors b i c trobats a partir dels dos primers mètodes. Ara passarem a representar aquests estimadors en funció del tamany del capitol N.

En els gràfics 5.9, s'observa clarament, que els paràmetres α's i b's són molt propers a 0 (en tot el decurs del llibre). D'altra banda, els paràmetres rellevants són els θ's i c's. El paràmetre θ es correlaciona positivament amb N. Aquesta relació no és lineal, creix més depressa per N's petites, és a dir, per capítols curts i s'estabilitza per capítols llargs.

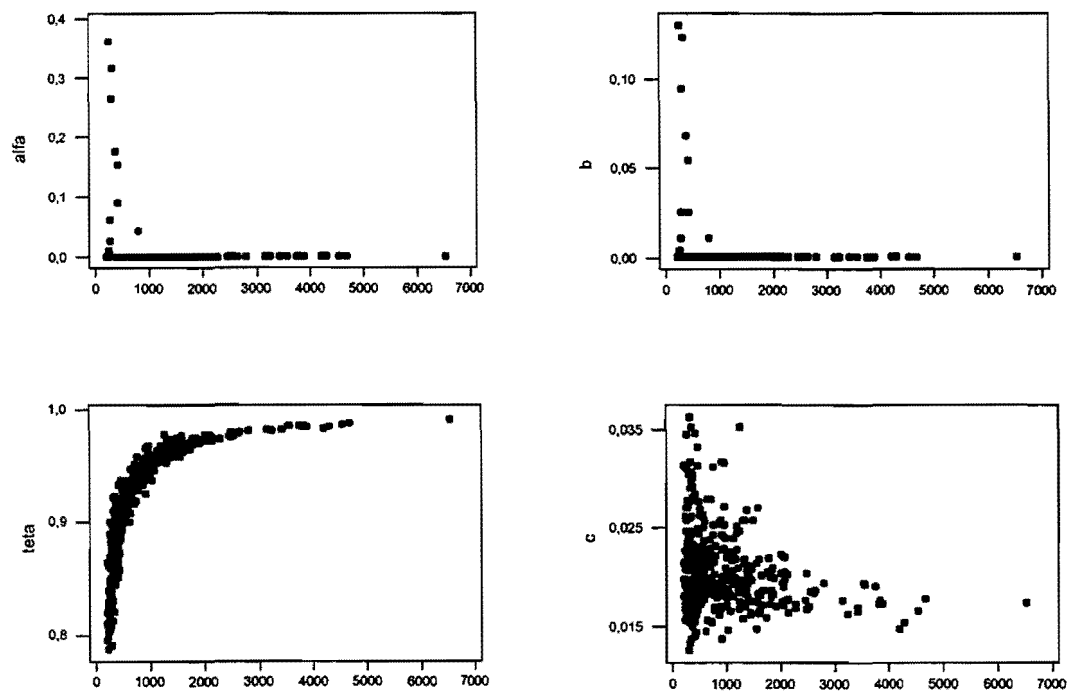
En canvi, el valor de c es manté constant per tots els capítols, tot i que la seva variabilitat és força més elevada per capítols amb poques paraules. Observem també que el primer mètode (5.1.1) presenta una variabilitat més petita en el paràmetre c que no pas el mètode dels moments (5.1.2).

Mètode 5.1.1



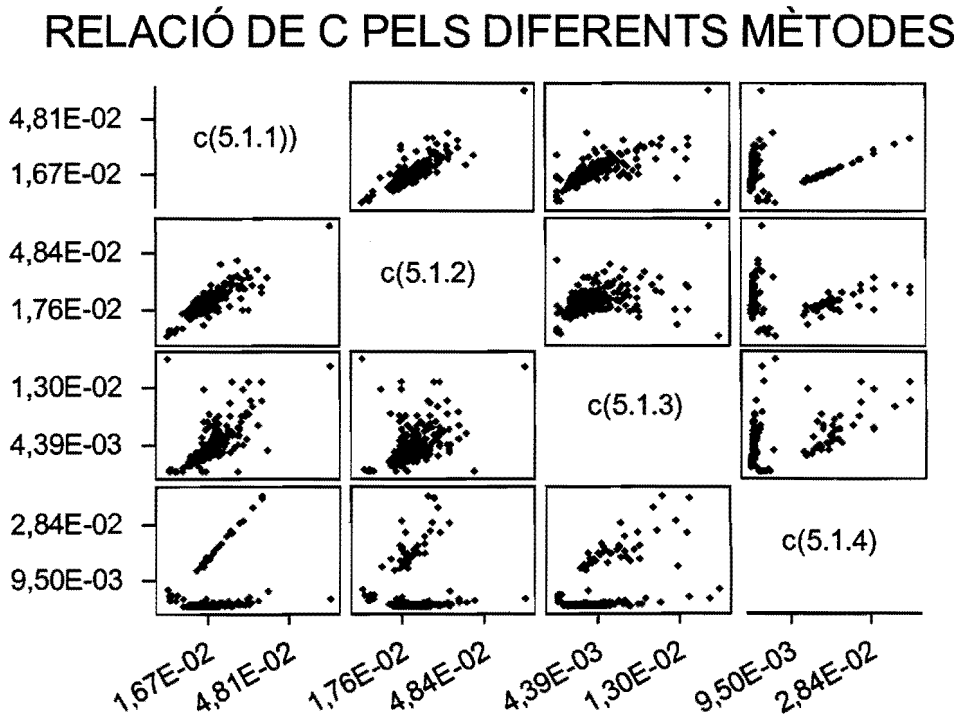
Gràfics 5.9: *Valor dels paràmetres estimats igualant al valor esperat i proporció d'una aparició, en funció dels capítols de més de 200 paraules. (Mètode 5.1.1)*

Mètode 5.1.2



Gràfics 5.10:*Valor dels paràmetres estimats pel mètode dels moments en funció dels capítols de més de 200 paraules.(Mètode 5.1.2)*

Tot seguit passem a veure la relació entre els diferents mètodes sobre el paràmetre *c*, el qual es comporta de forma menys variable que els altres. En el gràfic 5.11 veiem aquestes relacions. Observem que els mètodes més correlacionats pel que el paràmetre *c* són els dos primers. El mètode 5.1.3 i el 5.1.1 també presenten força correlació, encara que no és tan evident.



Gràfics 5.11: *Relació del paàmetre c pels quatre mètodes emprats*

5.8 Valoració dels resultats obtinguts

Fins ara, hem presentat els resultats a nivell gràfic, la qual cosa ens permet comparar els mètodes emprats a nivell descriptiu. Així doncs, en un primer cop d'ull, podem dir que en els dos primers mètodes utilitzats obtenim uns resultats molt semblants en tots els paràmetres estimats. Els valors d'α i de b són molt propers a 0 en ambdós casos, i els de θ i de c presenten la mateixa tendència, tot i que el primer mètode presenta uns valors de c no tan variables. Aquest paràmetre ens permet dir gràficament que a partir de l'observació 320, existeix un canvi.

Per altra banda, pels dos darrers mètodes implementats, els valors d'α i de b presenten un lleuger increment respecte als dos primers, aquests no són tant propers a zero. En el sistema basat en la proporció de les paraules que apareixen una i dues vegades en el text, el paràmetre c presenta més variabilitat i nivell gràfic no podem dir tant clarament que hi ha un canvi de tendència. En contra del què semblava, l'últim mètode, el de màxima versemblança, ens mostra els resultats més diferents i variables i en cap dels paràmetres podem dir que hi ha un canvi en la part final del *Tirant lo Blanc*.

De fet, en la implementació de l'últim mètode hem estat suposant premisses que no són certes, com són: que el vocabulari que utilitza l'autor està fixat, com també el nombre de repeticions r. Això és el que ens podria fer pensar que l'estimació dels paràmetres pel mètode de la màxima versemblança no és del tot satisfactori.

En el següent punt tractarem la bondat d'ajust per saber realment quin dels quatre mètodes és el que ens troba els millors estimadors dels paràmetres, és a dir, els que s'ajusten millor a les nostres dades.

ANNEXE Capítol 5:

Taula 5.1: Estimació d'α i θ pel m ètode de la probabilitat d'una aparició i el valor esperat, mètode dels moment, mètode de la probabilitat d'una i dues aparicions i pel mètode de la màxima versemblança, respectivament.

Obs.	5.1.1		5.1.2		5.1.3		5.1.4	
	α	θ	α	θ	α	θ	α	θ
1	0,0000	0,8490	0,0000	0,8630	0,7728	0,5133	17,8149	0,1324
2	0,0000	0,8868	0,0000	0,9108	0,5489	0,6321	21,8660	0,1314
3	0,0000	0,9409	0,0000	0,9589	0,5419	0,7525	26,4367	0,1573
4	0,0000	0,9093	0,0000	0,9284	0,6372	0,7107	0,0001	0,9093
5	0,0000	0,9371	0,0000	0,9566	0,5354	0,6848	28,7768	0,1414
6	0,0000	0,8799	0,0000	0,9231	0,8405	0,4949	19,2443	0,1427
7	0,0000	0,8394	0,0000	0,8872	1,0196	0,4450	20,2800	0,1132
8	0,0000	0,8554	0,0000	0,8613	0,4945	0,5963	16,1153	0,1487
9	0,0000	0,7816	0,0000	0,8102	1,2322	0,3708	14,9581	0,1195
10	0,0000	0,9103	0,0000	0,9328	0,7641	0,6310	25,4578	0,1307
11	0,0000	0,9287	0,0000	0,9318	0,6664	0,7385	22,3476	0,1660
12	0,0000	0,8917	0,1544	0,8764	0,5194	0,7283	0,0000	0,8917
13	0,0000	0,9422	0,0000	0,9568	0,5096	0,7326	29,3228	0,1450
14	0,0000	0,9058	0,0000	0,9222	0,7008	0,5632	24,1785	0,1336
15	0,0000	0,8587	0,0000	0,8748	0,5647	0,5417	19,1447	0,1298
16	0,0000	0,9258	0,0000	0,9358	0,5643	0,7384	23,4352	0,1559
17	0,0000	0,9126	0,0000	0,9289	0,6776	0,6649	24,4021	0,1377
18	0,0000	0,9484	0,0000	0,9619	0,6009	0,7635	24,6920	0,1787
19	0,0000	0,8623	0,0000	0,9048	0,9444	0,4561	22,2291	0,1153
20	0,0000	0,9072	0,0000	0,9233	0,5595	0,6748	18,8628	0,1679
21	0,0000	0,9080	0,0000	0,9388	0,7194	0,6327	25,1005	0,1306
22	0,0000	0,9221	0,0000	0,9378	0,5925	0,7321	21,8930	0,1615
23	0,0000	0,9181	0,0000	0,9290	0,5338	0,6903	21,8597	0,1576
24	0,0000	0,9460	0,0000	0,9538	0,4755	0,7681	24,9864	0,1730
25	0,0000	0,9507	0,0000	0,9653	0,5085	0,7851	23,1685	0,1932
26	0,0000	0,9661	0,0000	0,9761	0,4724	0,8431	27,5792	0,1971
27	0,0000	0,8556	0,0000	0,8850	0,8739	0,5235	20,5768	0,1194
28	0,0000	0,9036	0,0000	0,9181	0,6444	0,6858	0,0000	0,9037
29	0,0059	0,9022	0,3601	0,8686	0,5991	0,7309	0,0046	0,9024
30	0,0000	0,8784	0,0000	0,8793	0,7008	0,6380	17,6016	0,1525
31	0,0000	0,9206	0,0000	0,9417	0,4918	0,6915	26,8096	0,1334
32	0,0000	0,9118	0,0000	0,9285	0,5832	0,6630	24,9401	0,1345
33	0,0000	0,9172	0,0000	0,9316	0,5852	0,7036	24,6641	0,1405
34	0,0000	0,9422	0,0000	0,9471	0,5827	0,7561	28,3755	0,1493
35	0,0000	0,8933	0,0000	0,9019	0,6679	0,6285	20,4688	0,1442
36	0,0000	0,8481	0,0000	0,8709	0,6979	0,5755	16,0390	0,1442
37	0,0011	0,9278	0,0000	0,9426	0,5932	0,7648	0,0000	0,9279
38	0,0000	0,9432	0,0000	0,9537	0,4715	0,7625	29,4403	0,1456
39	0,0000	0,8166	0,0094	0,8157	0,9344	0,5355	13,0523	0,1513

40	0,0001	0,9270	0,0000	0,9447	0,6378	0,7322	25,1742	0,1475
41	0,0000	0,9248	0,0000	0,9368	0,6160	0,6896	23,5440	0,1543
42	0,0000	0,8982	0,2641	0,8731	0,4567	0,7200	17,2835	0,1720
43	0,0000	0,9274	0,0000	0,9502	0,5999	0,6923	19,7291	0,1839
44	0,0513	0,7310	0,8059	0,5943	0,6944	0,5378	6,5736	0,2046
45	0,2201	0,7752	0,1169	0,7916	0,9017	0,5886	0,1983	0,7787
46	0,0933	0,8159	0,3588	0,7767	0,7480	0,6238	0,0926	0,8160
47	2,8385	0,0582	0,2950	0,1496	103,9318	0,0030	2,6231	0,0603
48	0,1364	0,7489	0,4106	0,7006	0,8237	0,5536	0,1476	0,7470
49	0,2344	0,1095	0,1876	0,1071	37,4655	0,0049	1,6829	0,0514
50	0,0000	0,4849	0,0000	0,5473	9,1685	0,0441	8,6755	0,0805
51	0,0000	0,9359	0,0000	0,9477	0,5424	0,7686	23,0715	0,1704
52	0,0000	0,8888	0,0000	0,8990	0,7186	0,6751	0,0000	0,8888
53	0,0000	0,8232	0,0000	0,8641	1,2306	0,4260	16,0619	0,1304
54	0,0000	0,9415	0,0000	0,9555	0,5038	0,7632	26,0137	0,1604
55	0,0000	0,9485	0,0000	0,9565	0,5171	0,8079	19,0811	0,2239
56	0,0000	0,9515	0,0000	0,9603	0,5318	0,8094	23,8485	0,1897
57	0,0000	0,8468	0,0000	0,8615	0,7060	0,5884	17,4392	0,1330
58	0,0000	0,9568	0,0000	0,9681	0,5328	0,7753	30,2244	0,1625
59	0,0000	0,9405	0,0000	0,9578	0,5102	0,7836	23,1238	0,1766
60	0,0000	0,7719	0,0000	0,8289	1,1631	0,3380	17,0315	0,1036
61	0,0000	0,8740	0,0000	0,8929	0,8449	0,5270	17,4413	0,1511
62	0,0000	0,9505	0,0000	0,9563	0,4599	0,8310	22,8138	0,1953
63	0,0000	0,9594	0,0000	0,9681	0,5035	0,8046	30,0560	0,1682
64	0,1817	0,8629	0,3159	0,8477	0,6632	0,7276	0,1675	0,8644
65	0,0000	0,9679	0,0000	0,9774	0,4633	0,8615	30,4241	0,1850
66	0,0012	0,9784	0,0000	0,9854	0,3961	0,8990	33,3158	0,2033
67	0,0187	0,8646	0,1138	0,8534	0,7275	0,6559	0,0168	0,8649
68	0,0000	0,7621	0,0000	0,7754	1,3295	0,4038	12,4267	0,1300
69	36,2462	0,0317	15,1083	0,0676	2,1160	0,3639	6,0026	0,1561
70	0,0000	0,9723	0,0000	0,9806	0,4729	0,8435	30,8979	0,1949
71	0,0000	0,9534	0,0000	0,9583	0,4531	0,8022	20,7004	0,2188
72	0,0170	0,9534	0,0000	0,9618	0,5165	0,8306	27,4887	0,1718
73	0,0000	0,9556	0,0000	0,9658	0,4580	0,8007	30,8568	0,1573
74	0,0000	0,8671	0,0000	0,8864	0,6951	0,5878	19,3623	0,1329
75	0,0000	0,9399	0,0429	0,9374	0,4155	0,8192	24,1740	0,1689
76	0,2273	0,9203	0,0000	0,9371	0,6081	0,8262	0,1879	0,9230
77	0,3142	0,9506	0,0000	0,9663	0,5531	0,9022	27,6818	0,1888
78	0,0360	0,9740	0,0000	0,9818	0,4381	0,8913	27,3411	0,2266
79	0,0000	0,9167	0,0000	0,9308	0,6009	0,6893	24,3210	0,1418
80	0,1921	0,8373	0,0000	0,8652	0,7910	0,6698	0,1622	0,8411
81	0,0000	0,9394	0,0000	0,9548	0,5151	0,7000	29,8765	0,1393
82	0,0000	0,9573	0,0000	0,9707	0,5076	0,8034	22,4803	0,2122
83	0,0929	0,7952	0,0001	0,8103	0,9632	0,5534	0,0819	0,7969
84	7,0415	0,0661	3,0561	0,1150	80,4460	0,0085	4,5412	0,0908
85	3,4558	0,0913	1,9740	0,1222	106,3667	0,0052	3,8532	0,0811
86	0,0000	0,4694	0,0000	0,5421	79,3859	0,0071	7,4988	0,0841
87	0,0000	0,6746	0,0000	0,6996	2,2941	0,2202	11,3813	0,1078

88	0,1355	0,6527	0,8814	0,5136	0,9866	0,4353	0,1911	0,6411
89	0,0000	0,9537	0,0000	0,9590	0,4311	0,8369	25,5702	0,1824
90	0,0000	0,4448	0,1654	0,4153	6,5426	0,0645	5,5462	0,0981
91	0,0000	0,3629	0,1057	0,3449	52,4087	0,0083	4,5831	0,0857
92	0,0000	0,5727	0,0000	0,5992	1,0218	0,2690	8,6406	0,1025
93	0,0000	0,8814	0,0000	0,8948	0,6431	0,5982	20,0769	0,1381
94	0,0000	0,8875	0,0000	0,9114	0,7158	0,5941	21,4209	0,1344
95	0,0000	0,9670	0,0000	0,9797	0,4933	0,8400	25,0900	0,2167
96	0,0000	0,9486	0,0000	0,9661	0,4764	0,7639	22,6102	0,1936
97	0,0000	0,9772	0,0000	0,9856	0,4789	0,8759	35,6603	0,1870
98	0,0000	0,9695	0,0000	0,9784	0,4765	0,8538	35,3983	0,1653
99	0,1234	0,8257	0,0008	0,8425	0,6463	0,6707	0,1236	0,8257
100	0,0000	0,8702	0,0000	0,8865	0,7609	0,5888	18,5908	0,1397
101	0,0000	0,9645	0,0000	0,9780	0,4871	0,8143	34,3248	0,1588
102	0,0000	0,9576	0,0000	0,9718	0,5588	0,7781	25,2779	0,1922
103	0,0000	0,9712	0,0000	0,9783	0,4288	0,8600	35,1930	0,1707
104	0,0000	0,9124	0,0000	0,9400	0,6942	0,6717	0,0003	0,9124
105	0,0000	0,8853	0,0000	0,9109	0,7770	0,5821	21,1861	0,1341
106	0,0000	0,9466	0,0000	0,9638	0,5688	0,7794	25,5056	0,1707
107	0,0000	0,9678	0,0000	0,9792	0,4814	0,8169	24,8039	0,2211
108	0,0075	0,9795	0,0000	0,9856	0,4228	0,8974	38,4447	0,1836
109	0,0000	0,9362	0,0000	0,9530	0,4990	0,6895	29,5733	0,1369
110	0,0000	0,8727	0,0000	0,9042	0,7736	0,5249	21,8776	0,1226
111	0,0000	0,9540	0,0000	0,9658	0,5211	0,7935	24,9814	0,1869
112	0,0000	0,9598	0,0000	0,9749	0,4866	0,7902	37,9781	0,1367
113	10,3376	0,0433	2,6458	0,1177	104,2053	0,0061	4,5195	0,0841
114	0,0000	0,9433	0,0000	0,9612	0,5382	0,7343	29,3849	0,1460
115	0,0000	0,9519	0,0000	0,9657	0,5504	0,8014	24,0115	0,1894
116	0,0000	0,8938	0,0000	0,9282	0,7912	0,5680	21,8796	0,1364
117	0,0000	0,9756	0,0000	0,9852	0,4746	0,8679	33,9621	0,1896
118	0,0000	0,7946	0,0000	0,8022	0,9831	0,4764	13,5060	0,1358
119	0,0000	0,9388	0,0000	0,9553	0,6678	0,7274	28,3084	0,1453
120	0,0000	0,7808	0,0000	0,8399	1,2285	0,3666	14,3036	0,1241
121	0,0000	0,9138	0,0000	0,9309	0,6253	0,6927	20,0345	0,1655
122	0,0000	0,9587	0,0000	0,9737	0,4675	0,8074	28,2870	0,1761
123	0,0000	0,9597	0,0000	0,9744	0,4858	0,7927	37,1357	0,1393
124	0,0000	0,9506	0,0000	0,9617	0,4469	0,8129	23,3199	0,1919
125	0,0000	0,9590	0,0000	0,9722	0,5198	0,8144	28,5779	0,1751
126	0,0000	0,8881	0,0000	0,9176	0,8483	0,5685	20,9393	0,1375
127	0,0000	0,9527	0,0000	0,9688	0,4888	0,7610	29,7342	0,1580
128	0,0000	0,8865	0,0000	0,9004	0,7339	0,6232	20,2013	0,1407
129	0,0000	0,9415	0,0000	0,9582	0,4663	0,7394	27,4490	0,1530
130	0,0000	0,9630	0,0000	0,9752	0,4846	0,7923	25,3642	0,2040
131	0,0000	0,9809	0,0000	0,9868	0,4167	0,8925	27,0609	0,2560
132	0,0000	0,9498	0,0000	0,9702	0,5061	0,7477	25,2430	0,1776
133	2,7860	0,0664	0,4965	0,1469	126,8953	0,0028	2,6378	0,0674
134	0,0000	0,7589	0,0000	0,8165	1,3244	0,2960	16,4269	0,1028
135	0,0000	0,9506	0,0000	0,9631	0,5017	0,7753	29,9021	0,1538

136	0,0000	0,9656	0,0000	0,9773	0,4818	0,8064	24,0456	0,2207
137	0,0000	0,8625	0,0000	0,8901	0,8964	0,5270	20,2564	0,1250
138	0,0000	0,9579	0,0000	0,9663	0,4566	0,8328	28,5471	0,1731
139	0,0418	0,9743	0,0000	0,9850	0,4465	0,8916	24,8079	0,2481
140	0,0000	0,9221	0,0000	0,9360	0,5924	0,7261	24,1873	0,1477
141	0,0000	0,9682	0,0000	0,9841	0,5247	0,7461	29,6250	0,1903
142	0,0000	0,8926	0,0000	0,9146	0,7357	0,6124	23,0508	0,1291
143	0,0000	0,9583	0,0000	0,9723	0,5109	0,8147	22,3631	0,2153
144	0,0000	0,9714	0,0000	0,9825	0,4523	0,8003	40,2578	0,1515
145	0,0000	0,9396	0,0000	0,9571	0,5644	0,7343	28,4890	0,1455
146	0,0000	0,9651	0,0000	0,9729	0,4392	0,8565	27,8449	0,1928
147	0,0000	0,9246	0,0000	0,9393	0,5944	0,7317	24,3938	0,1491
148	0,0000	0,9422	0,0000	0,9588	0,5882	0,7247	28,9037	0,1468
149	0,0000	0,9750	0,0000	0,9850	0,4500	0,8496	25,5073	0,2407
150	0,0000	0,9636	0,0000	0,9739	0,4427	0,8251	24,5354	0,2114
151	0,0000	0,9001	0,0000	0,9249	0,6374	0,6431	22,8841	0,1355
152	0,0429	0,9789	0,0000	0,9861	0,4359	0,9030	36,9370	0,1908
153	0,0000	0,9697	0,0000	0,9774	0,4271	0,8616	27,2621	0,2092
154	0,0000	0,9808	0,0000	0,9881	0,4230	0,8611	50,3166	0,1476
155	0,0000	0,9444	0,0000	0,9627	0,5508	0,7413	27,7870	0,1551
156	0,0000	0,9680	0,0000	0,9813	0,4873	0,8165	25,2153	0,2186
157	0,0229	0,9732	0,0000	0,9823	0,4560	0,8819	25,1396	0,2389
158	0,0000	0,7775	0,0000	0,8035	1,4908	0,3614	14,9472	0,1172
159	0,0000	0,9502	0,0000	0,9682	0,5736	0,7487	25,2903	0,1780
160	0,0000	0,9193	0,0000	0,9498	0,6979	0,6065	24,7289	0,1424
161	0,0000	0,9002	0,0000	0,9177	0,6051	0,6726	22,4712	0,1377
162	0,0000	0,7586	0,0000	0,8073	1,5948	0,3007	13,6039	0,1206
163	0,2404	0,8453	0,1769	0,8529	0,6261	0,7368	0,2278	0,8469
164	0,0000	0,9396	0,0000	0,9577	0,5426	0,7421	28,2629	0,1465
165	0,0000	0,9311	0,0000	0,9497	0,6613	0,6855	23,5197	0,1617
166	0,0000	0,8601	0,0000	0,8822	0,8094	0,5256	17,9726	0,1378
167	0,0000	0,8497	0,0000	0,8650	0,7588	0,5383	20,2809	0,1179
168	0,0000	0,8796	0,0000	0,8994	0,7367	0,6106	20,4925	0,1340
169	0,0000	0,9093	0,0000	0,9188	0,6894	0,6509	22,5350	0,1451
170	0,0000	0,9269	0,0000	0,9482	0,5909	0,7313	25,2319	0,1471
171	0,0000	0,8922	0,0000	0,9115	0,8263	0,6286	0,0000	0,8922
172	0,0000	0,9040	0,0000	0,9205	0,5907	0,6898	0,0001	0,9040
173	0,0000	0,9220	0,0000	0,9282	0,4876	0,7340	23,4376	0,1519
174	0,0000	0,9233	0,0000	0,9425	0,5466	0,6413	24,7122	0,1464
175	0,0000	0,9109	0,0000	0,9325	0,6711	0,6339	25,5585	0,1308
176	0,0000	0,8932	0,0000	0,8977	0,6111	0,6784	20,7190	0,1424
177	0,0000	0,8480	0,0000	0,8844	0,8509	0,4979	20,6140	0,1155
178	0,0972	0,7780	0,0000	0,8031	0,9696	0,5369	0,0863	0,7798
179	0,1041	0,9849	0,0000	0,9912	0,3898	0,9392	33,4696	0,2449
180	0,0000	0,8299	0,0000	0,8512	0,8129	0,5261	16,8545	0,1280
181	0,1672	0,8141	0,0000	0,8404	0,8617	0,6199	0,1410	0,8178
182	0,0000	0,8806	0,0000	0,8926	0,7068	0,5999	20,2894	0,1361
183	0,0000	0,8128	0,0086	0,8126	0,9096	0,5235	13,6949	0,1432

184	0,0000	0,8318	0,0000	0,8820	1,1938	0,4101	21,2815	0,1050
185	0,0000	0,9270	0,0904	0,9207	0,4894	0,7793	18,3283	0,1952
186	0,0000	0,7297	0,0000	0,7724	1,9596	0,2469	14,0642	0,1068
187	0,0000	0,4965	0,0000	0,6537	0,0007	0,2098	13,5916	0,0622
188	0,0000	0,5481	0,0000	0,6585	61,5376	0,0087	13,4446	0,0661
189	0,4303	0,5986	0,8766	0,5200	1,2111	0,4294	0,5126	0,5830
190	0,0000	0,7665	0,0000	0,7974	1,4319	0,3433	14,3696	0,1174
191	0,0000	0,9085	0,0000	0,9236	0,6763	0,6664	23,6926	0,1379
192	0,0000	0,9024	0,0000	0,9249	0,6783	0,6044	25,5310	0,1245
193	0,1109	0,7170	0,4477	0,6536	1,1116	0,4593	0,1149	0,7163
194	0,0000	0,6672	0,0421	0,6655	1,0556	0,3423	9,2708	0,1255
195	0,0000	0,6529	0,0000	0,6706	1,7239	0,2622	10,0493	0,1120
196	0,0000	0,7266	0,0000	0,7710	2,0730	0,2467	13,9309	0,1064
197	0,0000	0,8269	0,0000	0,8501	0,8810	0,5051	16,7654	0,1271
198	0,0000	0,8978	0,0000	0,9254	0,8481	0,5914	22,9203	0,1335
199	0,0000	0,8894	0,0000	0,9083	0,8111	0,6152	22,0708	0,1318
200	0,0000	0,9299	0,0000	0,9434	0,5401	0,7199	17,9273	0,2034
201	0,0000	0,8460	0,0000	0,8890	0,8539	0,4681	21,1754	0,1120
202	0,0000	0,8504	0,0000	0,8634	0,7811	0,5891	14,8688	0,1553
203	0,0074	0,7748	0,1158	0,7558	1,1065	0,4779	11,7930	0,1420
204	0,0000	0,9092	0,0000	0,9295	0,6703	0,6440	25,2776	0,1307
205	0,0000	0,9364	0,0000	0,9444	0,5529	0,7748	21,4230	0,1827
206	0,0000	0,9228	0,0000	0,9442	0,5091	0,7225	28,0956	0,1296
207	0,0000	0,8508	0,0000	0,8839	0,9674	0,4715	20,3410	0,1184
208	0,0000	0,9505	0,0000	0,9645	0,5940	0,7335	31,4337	0,1469
209	0,0000	0,8638	0,0000	0,8779	0,6154	0,5708	18,2914	0,1380
210	0,0000	0,9429	0,0000	0,9595	0,5775	0,7565	26,9508	0,1573
211	0,0000	0,9265	0,0000	0,9594	0,6810	0,6261	30,8670	0,1223
212	0,0000	0,9669	0,0000	0,9731	0,5145	0,8371	23,4307	0,2294
213	0,0000	0,8163	0,0000	0,8297	0,9097	0,4861	15,7021	0,1294
214	0,0000	0,9303	0,0000	0,9448	0,5425	0,7407	25,5977	0,1488
215	0,0000	0,9205	0,0000	0,9392	0,6100	0,6591	25,5378	0,1393
216	0,0000	0,9335	0,0000	0,9467	0,5121	0,7277	26,8750	0,1460
217	0,0000	0,9096	0,0000	0,9286	0,5707	0,6946	0,0000	0,9097
218	0,0000	0,9491	0,0000	0,9609	0,4823	0,8017	18,7214	0,2288
219	0,0000	0,9392	0,0000	0,9576	0,6225	0,7427	27,6264	0,1490
220	0,0000	0,9219	0,0000	0,9301	0,6411	0,7136	17,5318	0,1960
221	0,0000	0,9498	0,0000	0,9636	0,4523	0,7692	22,6877	0,1953
222	0,0000	0,8400	0,0000	0,8843	0,8390	0,4879	20,0430	0,1146
223	0,0000	0,9655	0,0000	0,9770	0,4519	0,7992	37,3485	0,1490
224	0,0000	0,9609	0,0000	0,9713	0,4958	0,8325	24,6370	0,2039
225	0,0000	0,7859	0,0000	0,7979	0,5328	0,4819	13,0400	0,1374
226	0,0000	0,9606	0,0000	0,9726	0,4338	0,8021	30,1435	0,1701
227	0,0000	0,8457	0,0000	0,8755	0,8221	0,5090	19,6482	0,1194
228	0,0000	0,9291	0,0000	0,9463	0,6360	0,7213	25,1727	0,1498
229	0,0000	0,9437	0,0000	0,9610	0,5249	0,7640	26,0055	0,1636
230	0,0029	0,8886	0,0000	0,9016	0,7208	0,6800	0,0000	0,8889
231	0,0000	0,8423	0,0000	0,8857	1,1301	0,4498	20,6701	0,1124

232	0,0000	0,8726	0,0000	0,8778	0,6125	0,6789	0,0000	0,8726
233	0,0000	0,8811	0,0000	0,9117	0,6630	0,6030	20,6960	0,1340
234	0,0000	0,7963	0,0000	0,8186	1,0387	0,3776	15,2455	0,1245
235	0,0000	0,7009	0,0000	0,7153	1,5144	0,3146	10,8206	0,1212
236	0,0000	0,8813	0,0000	0,8929	0,7169	0,6543	0,0001	0,8814
237	0,0000	0,9092	0,0000	0,9173	0,5371	0,6966	23,1527	0,1415
238	0,0000	0,8609	0,0000	0,8785	0,8129	0,5936	17,9471	0,1378
239	0,0000	0,8857	0,0000	0,9002	0,6428	0,6661	0,0001	0,8858
240	0,0000	0,8628	0,0000	0,8845	0,5946	0,5582	15,8862	0,1558
241	0,0000	0,8736	0,0000	0,8811	0,7550	0,6170	19,1201	0,1383
242	0,0000	0,9088	0,0000	0,9363	0,8864	0,6062	24,0922	0,1361
243	0,0000	0,8743	0,0000	0,8837	0,7140	0,5888	18,3152	0,1446
244	0,0000	0,8584	0,0000	0,8731	0,7465	0,5799	16,7491	0,1452
245	0,0000	0,8457	0,0000	0,8653	0,5846	0,5128	17,8556	0,1307
246	0,0000	0,9421	0,0000	0,9575	0,5297	0,7578	26,4519	0,1589
247	0,0263	0,8373	0,0608	0,8326	0,8257	0,6042	0,0236	0,8376
248	0,0000	0,9436	0,0000	0,9591	0,5071	0,7430	28,4098	0,1510
249	0,0000	0,6418	0,0000	0,7018	1,9498	0,2178	12,3061	0,0919
250	0,0000	0,9736	0,0000	0,9830	0,4306	0,8238	26,9554	0,2245
251	0,0562	0,9456	0,0000	0,9574	0,4508	0,8483	22,4695	0,1941
252	0,0000	0,9539	0,0000	0,9704	0,5371	0,7666	33,8137	0,1422
253	0,0000	0,9024	0,0000	0,9333	0,6860	0,5609	26,7562	0,1194
254	0,0000	0,8751	0,0000	0,9069	0,8504	0,5508	20,9803	0,1285
255	0,0000	0,7738	0,0000	0,8252	1,7524	0,2870	16,9948	0,1042
256	0,0000	0,9182	0,0000	0,9458	0,6393	0,6030	24,9536	0,1402
257	0,0000	0,9481	0,0000	0,9649	0,5650	0,7565	23,2450	0,1880
258	0,0000	0,8052	0,0000	0,8287	1,0600	0,4399	15,9792	0,1222
259	0,0000	0,8992	0,0000	0,9161	0,6633	0,6525	22,1363	0,1388
260	0,0000	0,8454	0,0000	0,8756	0,9771	0,4143	19,1675	0,1226
261	0,0000	0,8493	0,0000	0,8769	0,9951	0,4698	19,1572	0,1242
262	0,0000	0,8572	0,0000	0,8778	0,8307	0,5745	18,6484	0,1310
263	0,0000	0,8798	0,0000	0,9156	0,8972	0,5098	20,4134	0,1351
264	0,0000	0,8583	0,0000	0,8882	0,9696	0,5433	19,9647	0,1239
265	0,0000	0,9306	0,0000	0,9474	0,6907	0,6936	24,3553	0,1561
266	0,0000	0,8776	0,0000	0,9080	0,8386	0,5905	15,8433	0,1667
267	0,0000	0,8723	0,0000	0,9063	0,8933	0,5227	22,8684	0,1174
268	0,0000	0,8370	0,0000	0,8580	0,9343	0,5024	17,4841	0,1276
269	0,0000	0,9437	0,0000	0,9611	0,5827	0,7456	28,6855	0,1497
270	0,0000	0,9380	0,0000	0,9526	0,5621	0,7081	28,1453	0,1452
271	0,0000	0,9530	0,0000	0,9692	0,5175	0,7552	21,9346	0,2074
272	0,0000	0,7520	0,0000	0,7905	1,3198	0,3312	14,9109	0,1087
273	0,0000	0,8474	0,0000	0,8706	0,9070	0,5696	0,0000	0,8476
274	0,0000	0,9341	0,0000	0,9534	0,5964	0,7433	26,4488	0,1488
275	0,0000	0,7545	0,0000	0,7827	1,2268	0,3306	13,3658	0,1211
276	0,0000	0,9432	0,0000	0,9639	0,5433	0,7254	29,8155	0,1441
277	0,0000	0,7487	0,0000	0,8093	1,4263	0,2978	16,4666	0,0988
278	0,0000	0,9146	0,0000	0,9373	0,6650	0,6608	23,4653	0,1446
279	0,0000	0,9616	0,0000	0,9736	0,4807	0,7897	31,6420	0,1647

280	0,0000	0,9606	0,0000	0,9741	0,5148	0,8075	31,6651	0,1626
281	0,0000	0,9142	0,0000	0,9363	0,6664	0,6195	18,3355	0,1798
282	0,0000	0,8416	0,0000	0,8637	0,9714	0,4843	18,0559	0,1266
283	0,0000	0,9087	0,0000	0,9373	0,7055	0,6110	24,5704	0,1338
284	0,0000	0,9158	0,0000	0,9422	0,6813	0,6202	25,7510	0,1341
285	0,0000	0,8741	0,0000	0,8921	0,6043	0,6103	20,2750	0,1319
286	0,6464	0,6571	1,4559	0,5364	0,8074	0,6168	0,8171	0,6291
287	0,0000	0,9644	0,0000	0,9764	0,5026	0,8005	26,3055	0,2010
288	0,0000	0,8372	0,0000	0,8544	0,9320	0,4930	16,8114	0,1323
289	0,0000	0,9515	0,0000	0,9661	0,4905	0,8073	28,7242	0,1609
290	0,0000	0,8553	0,0000	0,8838	0,7899	0,4826	19,6959	0,1247
291	0,0000	0,8842	0,0000	0,9022	0,6337	0,5740	21,1732	0,1337
292	0,0000	0,9609	0,0000	0,9714	0,4883	0,8076	30,2191	0,1702
293	0,0000	0,7724	0,0000	0,7869	1,1294	0,4136	13,1673	0,1288
294	0,0000	0,8457	0,0000	0,8682	1,1205	0,4305	19,0120	0,1233
295	0,0000	0,9479	0,0000	0,9602	0,5053	0,8167	22,0451	0,1965
296	0,0000	0,8552	0,0000	0,8840	0,9440	0,5150	18,9696	0,1282
297	0,0000	0,9567	0,0000	0,9745	0,4910	0,7620	26,3639	0,1834
298	0,0000	0,9688	0,0000	0,9800	0,4887	0,8349	25,2425	0,2211
299	0,0000	0,8539	0,0000	0,8716	0,8882	0,5331	18,4896	0,1304
300	0,0000	0,9310	0,0000	0,9503	0,6284	0,6747	28,3061	0,1368
301	0,0000	0,8761	0,0000	0,8974	0,6740	0,5731	20,3787	0,1327
302	0,0000	0,8056	0,0000	0,8630	0,8429	0,3953	19,4316	0,1038
303	0,0000	0,9579	0,0000	0,9748	0,5528	0,7892	24,4778	0,1982
304	0,0000	0,8272	0,0000	0,8648	0,8364	0,4611	15,5043	0,1372
305	0,0000	0,9316	0,0000	0,9460	0,5283	0,7197	26,8782	0,1439
306	0,0000	0,8508	0,0000	0,8766	0,8414	0,4707	19,0373	0,1260
307	0,0000	0,9447	0,0000	0,9632	0,6011	0,7296	27,9826	0,1546
308	0,0000	0,8112	0,0000	0,8236	1,0822	0,4614	14,9356	0,1324
309	0,0000	0,9470	0,0000	0,9660	0,5380	0,7605	27,1447	0,1621
310	0,0000	0,8779	0,0000	0,8975	0,5657	0,5803	20,4298	0,1338
311	0,0000	0,9048	0,0000	0,9177	0,6688	0,6660	0,0000	0,9048
312	0,0000	0,9104	0,0000	0,9287	0,6593	0,6264	25,1687	0,1324
313	0,0000	0,8788	0,0000	0,9090	0,7260	0,5353	20,8312	0,1321
314	0,0000	0,8045	0,0000	0,8332	0,8184	0,4567	16,3161	0,1201
315	0,0000	0,9062	0,0000	0,9210	0,5765	0,6825	23,2781	0,1382
316	0,0000	0,9327	0,0000	0,9436	0,5351	0,7381	22,9216	0,1674
317	0,0000	0,9120	0,0000	0,9299	0,5790	0,6724	24,3003	0,1378
318	0,0000	0,9414	0,0000	0,9572	0,4816	0,7139	29,2329	0,1444
319	0,0000	0,8767	0,0000	0,8994	0,8229	0,5884	17,8120	0,1497
320	0,0000	0,8461	0,0000	0,8589	0,6797	0,6093	17,3767	0,1328
321	0,0000	0,9013	0,0000	0,9315	0,6215	0,5758	25,4994	0,1241
322	0,0000	0,9482	0,0000	0,9669	0,5494	0,7789	23,7700	0,1845
323	0,0000	0,9207	0,0000	0,9301	0,5986	0,7352	0,0000	0,9207
324	0,0000	0,7808	0,0000	0,7954	0,9398	0,4421	13,8670	0,1270
325	0,0000	0,9331	0,0000	0,9509	0,6121	0,7032	28,4220	0,1384
326	0,0000	0,7903	0,0000	0,8243	1,0867	0,4162	14,7329	0,1246
327	0,0000	0,9307	0,0000	0,9509	0,5872	0,7448	24,2136	0,1570

328	0,0000	0,9635	0,0000	0,9763	0,5266	0,8226	24,7982	0,2093
329	0,0000	0,7978	0,0000	0,8402	1,0169	0,3927	17,4684	0,1106
330	0,0000	0,7691	0,0434	0,7639	0,8362	0,4789	11,2789	0,1453
331	0,0000	0,9411	0,0000	0,9614	0,5482	0,6965	29,6845	0,1420
332	0,0000	0,9144	0,0000	0,9460	0,7985	0,6068	24,8289	0,1373
333	0,0000	0,9149	0,0000	0,9358	0,6500	0,6696	25,1580	0,1360
334	0,0000	0,8725	0,0000	0,9071	0,7416	0,5436	22,3361	0,1201
335	0,0000	0,8863	0,0000	0,9162	0,6816	0,5970	20,7542	0,1374
336	0,0000	0,7705	0,0000	0,8222	1,0862	0,3796	14,5540	0,1180
337	0,0000	0,9596	0,0000	0,9751	0,5133	0,7528	23,1024	0,2126
338	0,0000	0,9325	0,0000	0,9555	0,6779	0,6414	28,5405	0,1373
339	0,0000	0,8743	0,0000	0,9028	0,7738	0,5217	21,2710	0,1268
340	0,0000	0,7843	0,0000	0,8108	1,0521	0,4349	14,4558	0,1236
341	0,0000	0,8826	0,0000	0,9188	0,9252	0,4937	20,7228	0,1353
342	0,0000	0,9160	0,0000	0,9422	0,6064	0,6496	25,0192	0,1378
343	0,0000	0,9346	0,0000	0,9580	0,5386	0,7098	28,0539	0,1417
344	0,0000	0,8476	0,0000	0,8728	0,8007	0,4923	19,2293	0,1230
345	0,0000	0,9146	0,0000	0,9446	0,7096	0,5997	20,7473	0,1617
346	0,0000	0,8319	0,0000	0,8383	0,8438	0,5545	15,1832	0,1411
347	0,0000	0,8347	0,0000	0,8730	0,8338	0,5091	19,3974	0,1152
348	0,0000	0,8270	0,0000	0,8678	0,8171	0,4814	18,6851	0,1158
349	0,0000	0,6514	0,0000	0,6737	1,4444	0,2726	10,4262	0,1089
350	0,0000	0,8558	0,0000	0,8956	1,0966	0,4274	21,4902	0,1154
351	0,0000	0,8025	0,0000	0,8356	0,8564	0,4364	16,6074	0,1176
352	0,0000	0,8636	0,0000	0,8655	0,6845	0,6282	16,1292	0,1534
353	0,0000	0,8266	0,0000	0,8481	0,9523	0,5026	15,7850	0,1338
354	0,0000	0,9007	0,0000	0,9296	0,7752	0,6215	22,4355	0,1384
355	0,0000	0,9007	0,0000	0,9296	0,7752	0,6215	22,4355	0,1384
356	0,0000	0,8527	0,0000	0,8803	0,9811	0,5079	19,8496	0,1217
357	0,0000	0,8622	0,0000	0,8849	0,7593	0,5707	19,3287	0,1301
358	0,0000	0,8844	0,0000	0,9207	0,9036	0,4867	20,4044	0,1386
359	0,0000	0,9375	0,0000	0,9425	0,4618	0,7977	19,1789	0,2029
360	0,0000	0,9127	0,0000	0,9418	0,8230	0,5862	25,2310	0,1339
361	0,0000	0,8470	0,0000	0,8907	1,1734	0,4250	21,2140	0,1122
362	0,0000	0,9274	0,0000	0,9591	0,5980	0,6098	30,2848	0,1254
363	0,0000	0,8249	0,0000	0,8924	1,4246	0,3492	20,6540	0,1052
364	0,0000	0,8973	0,0000	0,9340	0,8291	0,5719	25,1861	0,1224
365	0,0000	0,8139	0,0000	0,8758	1,1046	0,3752	19,5504	0,1062
366	0,0000	0,8743	0,0000	0,9002	0,9540	0,5255	20,1216	0,1329
367	0,0000	0,7797	0,0000	0,8043	1,3012	0,3717	15,1438	0,1172
368	0,0000	0,8239	0,0000	0,8560	0,8402	0,4536	17,8620	0,1195
369	0,0000	0,7887	0,0000	0,8119	0,8957	0,4235	15,3585	0,1198
370	0,0000	0,8709	0,0000	0,8779	0,6189	0,6158	18,5759	0,1404
371	0,0000	0,8731	0,0000	0,9000	0,8015	0,5860	21,0861	0,1265
372	0,0000	0,9169	0,0000	0,9309	0,6379	0,7064	0,0000	0,9169
373	0,0000	0,8520	0,0000	0,8686	0,8189	0,5469	18,2060	0,1311
374	0,0000	0,9189	0,0000	0,9379	0,6563	0,6255	24,5286	0,1431
375	0,0000	0,9580	0,0000	0,9705	0,4676	0,7868	32,0040	0,1565

376	0,0000	0,9107	0,0000	0,9157	0,5657	0,6720	20,6935	0,1579
377	0,0000	0,8520	0,0000	0,8573	0,5400	0,6554	13,0163	0,1754
378	0,0000	0,9078	0,0000	0,9186	0,5369	0,6687	22,4714	0,1443
379	0,0000	0,8987	0,0000	0,9214	0,5417	0,5513	25,0923	0,1242
380	0,0000	0,8956	0,0000	0,8975	0,5716	0,6869	19,3703	0,1533
381	0,0000	0,9028	0,0000	0,9223	0,7525	0,5865	23,9281	0,1324
382	0,0000	0,9352	0,0000	0,9541	0,5647	0,6732	23,8613	0,1648
383	0,0000	0,8720	0,0000	0,8889	0,5907	0,6092	18,8781	0,1393
384	0,0000	0,8912	0,0000	0,9095	0,6545	0,5903	22,4474	0,1315
385	0,0000	0,8505	0,0000	0,8643	0,7544	0,5848	17,2750	0,1362
386	0,0000	0,7983	0,0000	0,8441	0,8987	0,3804	18,3919	0,1062
387	0,0000	0,9591	0,0000	0,9664	0,4455	0,8365	28,9586	0,1732
388	0,0000	0,7471	0,0000	0,8126	2,1593	0,2535	14,9545	0,1065
389	0,0000	0,9443	0,0000	0,9581	0,5690	0,7696	26,6891	0,1606
390	0,0000	0,8960	0,0000	0,9289	0,7604	0,5688	23,6916	0,1285
391	0,0000	0,8886	0,0000	0,8999	0,7267	0,5230	21,0610	0,1377
392	0,0000	0,9227	0,0000	0,9378	0,5640	0,7147	23,6386	0,1515
393	0,0000	0,9149	0,0000	0,9152	0,6276	0,7175	0,0000	0,9150
394	0,0000	0,9693	0,0000	0,9779	0,4726	0,7911	39,7391	0,1485
395	0,0000	0,9575	0,0000	0,9679	0,4401	0,8114	21,9702	0,2168
396	0,0000	0,9650	0,0000	0,9749	0,4615	0,8204	23,8884	0,2202
397	0,0000	0,8425	0,0000	0,8756	0,9417	0,4778	19,7913	0,1171
398	0,0000	0,8279	0,0000	0,8646	1,0806	0,4311	19,0074	0,1145
399	0,0000	0,8821	0,0000	0,9148	0,4899	0,6483	21,2695	0,1314
400	0,0000	0,9551	0,0000	0,9667	0,5241	0,7549	34,8359	0,1403
401	0,0000	0,9372	0,0000	0,9572	0,5641	0,7123	28,3985	0,1431
402	0,0000	0,9371	0,0000	0,9523	0,5396	0,7321	28,5169	0,1424
403	0,0000	0,7586	0,0000	0,8061	1,2238	0,3348	14,8183	0,1120
404	0,0000	0,9630	0,0000	0,9711	0,4377	0,7974	24,8456	0,2077
405	0,0000	0,9569	0,0000	0,9681	0,5387	0,7955	25,6989	0,1879
406	0,0000	0,9199	0,0000	0,9337	0,5577	0,6912	19,4017	0,1773
407	0,1774	0,9227	0,0000	0,9349	0,4158	0,8635	0,1829	0,9224
408	0,0000	0,9305	0,0000	0,9494	0,4995	0,7177	28,0049	0,1376
409	0,0000	0,9101	0,0000	0,9217	0,6462	0,6598	22,8588	0,1440
410	0,0000	0,7540	0,0000	0,8198	1,4335	0,2963	16,6136	0,0998
411	0,0000	0,9180	0,0000	0,9327	0,5434	0,7034	26,2674	0,1335
412	0,0000	0,8485	0,0000	0,8801	0,9715	0,4630	19,5293	0,1217
413	0,0000	0,8544	0,0000	0,8738	0,6840	0,5709	18,6695	0,1296
414	0,0000	0,9494	0,0000	0,9628	0,5049	0,7779	23,2335	0,1905
415	0,0000	0,9206	0,0000	0,9405	0,6340	0,6911	19,5738	0,1766
416	0,0000	0,9383	0,0000	0,9481	0,4823	0,7397	27,5399	0,1484
417	0,0000	0,8397	0,0257	0,8365	0,7070	0,6153	12,8538	0,1681
418	0,0000	0,9444	0,0000	0,9616	0,6158	0,7274	27,7961	0,1551
419	0,0000	0,8515	0,0000	0,8805	0,8581	0,4948	19,4274	0,1239
420	0,0000	0,8369	0,0000	0,8535	0,9005	0,5251	15,9978	0,1378
421	0,0000	0,7711	0,0000	0,8212	1,2394	0,3542	14,7024	0,1173
422	0,0000	0,8554	0,0000	0,8850	0,7541	0,5442	16,6533	0,1443
423	0,0000	0,8917	0,0000	0,9015	0,7326	0,6715	0,0001	0,8917

424	0,0000	0,9084	0,0000	0,9115	0,6112	0,6622	19,8640	0,1616
425	0,0000	0,7343	0,0791	0,7223	0,9648	0,4244	10,9900	0,1325
426	0,0000	0,9211	0,0000	0,9428	0,6892	0,6731	21,9862	0,1599
427	0,0000	0,7752	0,0000	0,8047	1,0200	0,4054	14,1051	0,1230
428	0,0000	0,9038	0,0000	0,9337	0,7702	0,5633	25,8023	0,1244
429	0,0000	0,9395	0,0000	0,9544	0,5125	0,7827	22,8737	0,1768
430	0,0000	0,9561	0,0000	0,9677	0,5247	0,7913	31,2024	0,1566
431	0,0000	0,8610	0,0000	0,8917	0,8490	0,5119	20,5303	0,1228
432	0,0000	0,9509	0,0000	0,9643	0,5050	0,7594	31,2010	0,1485
433	0,0000	0,9227	0,0000	0,9305	0,5121	0,7203	23,0492	0,1550
434	0,0000	0,9626	0,0000	0,9716	0,5326	0,8109	23,9826	0,2130
435	0,0000	0,8939	0,0000	0,9157	0,6284	0,6792	0,0001	0,8939
436	0,0000	0,9399	0,0000	0,9632	0,5773	0,6888	29,4070	0,1418
437	12,9569	0,0488	6,3842	0,0807	63,5739	0,0130	5,3910	0,0991
438	0,0000	0,8664	0,0000	0,8828	0,8600	0,5634	19,5225	0,1313
439	0,0000	0,8015	0,0000	0,8198	0,9115	0,4737	15,3891	0,1245
440	0,0000	0,9649	0,0000	0,9746	0,4498	0,8188	32,4920	0,1677
441	0,0000	0,8094	0,0000	0,8394	0,9894	0,4877	15,6214	0,1261
442	0,0000	0,8543	0,0000	0,8794	0,9176	0,5523	15,8099	0,1498
443	0,0000	0,9511	0,0000	0,9664	0,5422	0,7658	30,7505	0,1507
444	0,0000	0,8301	0,0000	0,8550	0,8620	0,4809	17,8782	0,1220
445	0,0000	0,5684	0,0000	0,6136	0,7347	0,2774	10,1442	0,0898
446	0,0556	0,8912	0,0000	0,8997	0,6854	0,7121	0,0473	0,8920
447	0,0000	0,9615	0,0000	0,9771	0,5965	0,7783	31,4175	0,1655
448	0,0000	0,8854	0,0000	0,9106	0,7576	0,5964	21,1045	0,1345
449	0,0000	0,9227	0,0000	0,9378	0,5711	0,7390	20,5849	0,1712
450	0,0000	0,9109	0,0000	0,9352	0,7242	0,6502	0,0001	0,9109
451	0,0000	0,9029	0,0000	0,9213	0,7930	0,6066	21,0275	0,1486
452	0,0000	0,8233	0,0000	0,8507	0,9169	0,4707	17,5795	0,1204
453	0,3183	0,1348	0,2675	0,1303	91,2214	0,0028	2,0926	0,0590
454	0,0000	0,9439	0,0000	0,9578	0,5018	0,7321	26,7369	0,1600
455	0,0000	0,8924	0,0000	0,9258	0,8050	0,5158	24,5408	0,1223
456	0,0000	0,8918	0,0000	0,9145	0,7294	0,6243	22,3999	0,1318
457	0,0000	0,9288	0,0000	0,9535	0,6207	0,6875	27,1893	0,1395
458	0,0000	0,8399	0,0000	0,8713	0,9025	0,4463	16,1938	0,1391
459	0,0000	0,9480	0,0000	0,9600	0,4932	0,7923	23,7005	0,1847
460	0,0000	0,9373	0,0000	0,9568	0,6620	0,7275	26,5877	0,1519
461	0,0000	0,9549	0,0000	0,9735	0,5168	0,7205	32,1020	0,1508
462	0,0000	0,9367	0,0000	0,9566	0,5719	0,6983	28,6008	0,1416
463	0,1184	0,9313	0,0000	0,9471	0,5186	0,8301	26,8412	0,1518
464	0,0000	0,9537	0,0000	0,9664	0,4472	0,7984	21,5354	0,2121
465	0,0000	0,8893	0,0000	0,9115	0,6494	0,6506	18,0001	0,1581
466	0,0000	0,9427	0,0000	0,9636	0,5142	0,7404	29,1444	0,1464
467	0,0000	0,9089	0,0000	0,9231	0,6281	0,6971	0,0000	0,9089
468	0,0000	0,9303	0,0000	0,9586	0,7094	0,6444	31,4697	0,1235
469	0,0001	0,9291	0,0000	0,9330	0,4949	0,7849	19,8309	0,1851
470	0,0000	0,8703	0,0000	0,8955	0,7680	0,6094	16,1822	0,1579

CAPÍTOL 6:

Bondat d'ajust de la distribució de Sichel als capítols del *Tirant*

L'objecte del nostre estudi és trobar una distribució que ajusti bé a les nostres dades per poder determinar amb seguretat, si existeix un canvi d'estil en el *Tirant lo Blanc*. S'intuïa una distribució que probablement ajustava, però els paràmetres no eren coneguts i per tant s'han hagut d'estimar. Per a fer-ho hem emprat diferents mètodes d'estimació que ja s'han explicat amb tot detall en el capítol anterior, on també s'hi mostren els resultats obtinguts. Ara com ara, ja sabem quina seria la distribució i ja tenim els paràmetres, però els valors d'aquests canvien en funció del mètode emprat. És conegut, que el mètode de màxima versemblança és consistent i per tant, és un dels mètodes que ens plantegem d'entrada.

En aquest capítol comprovarem quin dels mètodes emprats en la distribució dels paràmetres és el que s'ajusta millor a les nostres dades. Es vol contrastar si les dades provenen d'una família de variables amb distribució Sichel de paràmetres α i θ estimats per cadascun dels mètodes. Per fer això apliquem una prova de bondat d'ajust, concretament fem servir el contrast khi-quadrat (χ^2).

6.1 Test Khi-quadrat

Les proves de bondat d'ajust consisteixen bàsicament en comparar els valors observats d'una variable d'interès donada amb els valors que se n'esperen a partir del model estimat. Si aquests valors s'assemblen podem dir que el nostre model formulat és correcte per ajustar les nostres dades, però si difereixen considerablement, vol dir que el model utilitzat és erroni per representar-les i per tant, hauriem d'ajustar-les amb altres models.

En el nostre cas, el model teòric que tenim és la distribució de Sichel. El que volem comparar és la freqüència de l'aparició de les paraules que observem en el text amb la freqüència esperada per la distribució de Sichel a partir dels paràmetres que ja hem estimat en l'anterior capítol. Un dels tests que es poden emprar per dur a terme aquesta comparació amb una cert confiança és el de la khi-quadrat (χ^2), el qual passem a descriure en les següents línies.

Suposem que una població gran es pot classificar en k categories, i cadascun dels individus tenen una probabilitat p_i de pertànyer a la categoria i $(i=1,...,k)$. Hem de suposar que p_i 0 per $i=1,...,k$ i que $p_i=1$. Siguin $\pi_1,...,\pi_k$ valors teòrics de la distribució, que compleixen que $\pi_i>0$ per $i=1,...,k$ i que $\pi_i=1$ i suposem que contrastem la següent hipòtesis:

$$\begin{cases} \boldsymbol{H_0}: \pi_i = p_i, \quad per\ i=1,...,k \\ \boldsymbol{H_1}: desmenteix\ H_0 \end{cases}$$

Suposem que es selecciona una mostra aleatòria de grandària n d'una població determinada, de tal manera que s'obtenen n observacions independents, les quals tenen una probabilitat π_i de pertànyer a una categoria i $(i=1,...,k)$. Partint d'aquesta mostra es contrasta la hipòtesis. Denotem n_i com el número d'individus que observem i pertanyen a la categoria i, així doncs $n_i=n$.

Així doncs, sota la hipòtesis nula, el nombre esperat d'individus en cadascuna de les categories és $n\pi_i$ *(i=1,...,k)*. La diferència entre el nombre observat n_i i el nombre esperat $n\pi_i$ tendirà a ésser més petit, quan H_0 sigui certa, que quan no sigui certa. És a dir, no podrem rebutjar la hipòtesis nula quan els valors esperats i els observats no divergeixin, del contrari, si les magnituds de les diferències entre observats i esperats siguin relativament grans, la rebutjarem.

L'any 1900, Karl Pearson va proposar el següent estadístic:

$$Q = \sum_{i=1}^{k} \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

Pearson va demostrar, que sota H_0, quan la grandària de mostra $n \to \infty$, la f.d. de Q convergeix a la f.d. de la distribució χ^2 amb *k-1* graus de llibertat. L'exposició que s'ha presentat indica que H_0 s'hauria de rebutjar quan $Q>c$, en el qual c és una constant apropiada. Si es desitja realitzar el contrast a un nivell de significació α_0, llavors c hauria d'escollir-se de forma que es compleixi $Pr(Q>c)=\alpha_0$ quan Q tingui una distribució χ^2 amb *k-1* graus de llibertat. Aquest contrast s'anomena *contrast* χ^2 *de bondat d'ajust.*

Sempre que el valor de cadascuna de les esperances $n\pi_i$ *(i=1,...,k)* no sigui molt petit, la distribució χ^2 serà una bona aproximació a la distribució real de Q. Concretament, l'aproximació serà molt bona si $n\pi_i>5$ per $i=1,...,k$, però també resultaria satisfactòria si $n\pi_i>1.5$ per $i=1,...,k$.

6.2 Resultats del contrast Khi-quadrat als ajustos del capítol 5

Després d'haver estimat els paràmetres pels quatre mètodes explicats en la secció 5.1 de l'anterior capítol, passem a contrastar si les nostres dades provenen realment d'aquest model, és a dir, de la distribució de Sichel amb els paràmetres α i θ, corresponents a l'estimació per cadascun dels mètodes. Per esbrinar-ho, utilitzem la prova de bondat d'ajust khi-quadrat.

Considerem una agrupació en el nombre de repeticions formada per les següents categories:

-*Categoria 1*: nombre de paraules que només apareixen una vegada
-*Categoria 2*: nombre de paraules que es repeteixen dues vegades
-*Categoria 3*: nombre de paraules que es repeteixen tres vegades
-*Categoria 4*: nombre de paraules que es repeteixen quatre vegades
-*Categoria 5*: nombre de paraules que es repeteixen cinc vegades
-*Categoria 6*: nombre de paraules que es repeteixen sis vegades
-*Categoria 7*: nombre de paraules que es repeteixen set o més vegades

A mesura que augmentem el número de categoria, el nombre de paraules que conté és cada cop menor. Per no tenir una categoria amb un número molt petit de paraules, hem decidit formar només aquestes set categories, on a la última hi afegim també les paraules que es repeteixen més de set cops. La raó per fer-ho és la d'augmentar $n\pi_i$ i per tant, millorar la potència del test χ^2.

Recordem que estem treballant amb la distribució de Sichel havent fixat el paràmetre γ igual a –1/2, la qual, com ja hem vist, és la següent:

$$\Phi(r|N)=\frac{\left(\frac{2\alpha}{\pi}\right)^{1/2}\exp(\alpha)}{\exp\left[\alpha\left(1-(1-\theta)^{1/2}\right)\right]-1}\frac{(0.5\alpha\theta)^r}{r!}k_{r-1/2}(\alpha),$$

A partir d'aquesta funció calculem les probabilitats de cadascuna de les classificacions, i posteriorment els valors esperats, que es calculen a partir del producte de les probabilitats per V (número de paraules diferents que surten en el text).

A continuació adjuntem un exemple perquè es vegi més clarament com calculem la prova de bondat d'ajust. Aquest exemple pertany al primer capítol del llibre i el mostrem en les taules 6.1, 6.2, 6.3 i 6.4 per cadascun dels mètodes emprats per l'estimació dels paràmetres. En aquest cas, hi ha un total de 255 paraules, de les quals 142 són diferents. Així doncs, el tamany de mostra és 142. Es calculen les probabilitats marginals, els valors esperats de la distribució de Sichel (α, θ) i l'estadístic χ^2.

Mètode 5.1.1				
Aparicions	**Funció densitat**	**E[ϕ(α,θ)]=ϕ(r=k/α,θ)*V**	**Observats**	$\frac{(n_i - n\hat{p}_i)}{n\hat{p}_i}$
1	ϕ(r=1/α,θ)= 0.6943	ϕ(r=1/α,θ)*V= 98.5872	107	0.7179
2	ϕ(r=2/α,θ)= 0.1474	ϕ(r=1/α,θ)*V= 20.9258	16	1.1595
3	ϕ(r=3/α,θ)=0.0626	ϕ(r=1/α,θ)*V= 8.8833	6	0.9358
4	ϕ(r=4/α,θ)= 0.0332	ϕ(r=1/α,θ)*V= 4.7138	2	1.5624
5	ϕ(r=5/α,θ)= 0.0197	ϕ(r=1/α,θ)*V= 2.8015	2	0.2293
6	ϕ(r=6/α,θ)=0.0126	ϕ(r=1/α,θ)*V= 1.7839	2	0.0262
7	$1-\sum_{k=1}^{6}\phi(r=k)$ =0.0303	$[1-\sum_{k=1}^{6}\phi(r=k)]$*V= 4.3044	7	1.6881
Total	1	142	142	χ^2=6.3192 p_valor=0.3884

Taula 6.1: *adjuntem un exemple de càlcul de bondat d'ajust, que correspon al primer capítol del llibre. A la taula s'hi adjunta, la probabilitat teòrica, el valor esperat per la distribució, el valor observat i el valor de l'estadístic χ^2. L'ajust correspon als valors α=0.000, θ=0.8490 estimats pel mètode 5.1.1.*

Mètode 5.1.2				
Aparicions	**Funció densitat**	**E[φ(α,θ)]=φ(r=k/α,θ)*V**	**Observats**	$\frac{(n_i - n\hat{p}_i)}{n\hat{p}_i}$
1	φ(r=1/α,θ)= 0.6851	φ(r=1/α,θ)*V=97.2812	107	0.9709
2	φ(r=2/α,θ)= 0.1478	φ(r=1/α,θ)*V=20.9880	16	1.1854
3	φ(r=3/α,θ)=0.0638	φ(r=1/α,θ)*V=9.0562	6	1.0313
4	φ(r=4/α,θ)= 0.0344	φ(r=1/α,θ)*V=4.8846	2	1.7035
5	φ(r=5/α,θ)= 0.0208	φ(r=1/α,θ)*V=2.9507	2	0.3063
6	φ(r=6/α,θ)=0.0134	φ(r=1/α,θ)*V=1.9098	2	0.0042
7	$1-\sum_{k=1}^{6}\phi(r=k)$ =0.0347	$[1-\sum_{k=1}^{6}\phi(r=k)]$*V= 4.9295	7	0.8696
Total	1	**142**	**142**	**χ^2=6.0714 p_valor=0.4152**

Taula 6.2: *adjuntem un exemple de càlcul de bondat d'ajust, que correspon al primer capítol del llibre. A la taula s'hi adjunta, la probabilitat teòrica, el valor esperat per la distribució, el valor observat i el valor de l'estadístic χ^2. L'ajust correspon als valors α=0.000, θ=0.8630 estimats pel mètode 5.1.2.*

Mètode 5.1.3				
Aparicions	**Funció densitat**	**E[φ(α,θ)]=φ(r=k/α,θ)*V**	**Observats**	$\frac{(n_i - n\hat{p}_i)}{n\hat{p}_i}$
1	φ(r=1/α,θ)= 0.3358	φ(r=1/α,θ)*V=47.6825	107	73.7915
2	φ(r=2/α,θ)= 0.2682	φ(r=1/α,θ)*V=38.0864	16	12.8079
3	φ(r=3/α,θ)=0.1736	φ(r=1/α,θ)*V=24.6558	6	14.1159
4	φ(r=4/α,θ)= 0.1014	φ(r=1/α,θ)*V=14.3958	2	10.6736
5	φ(r=5/α,θ)= 0.0562	φ(r=1/α,θ)*V=7.9764	2	4.4778
6	φ(r=6/α,θ)=0.0303	φ(r=1/α,θ)*V=4.3054	2	1.2344
7	$1-\sum_{k=1}^{6}\phi(r=k)$ =0.0345	$[1-\sum_{k=1}^{6}\phi(r=k)]$*V= 4.8976	7	0.9025
Total	1	**142**	**142**	**χ^2=118.0040 p_valor=0**

Taula 6.3: *adjuntem un exemple de càlcul de bondat d'ajust, que correspon al primer capítol del llibre. A la taula s'hi adjunta, la probabilitat teòrica, el valor esperat per la distribució, el valor observat i el valor de l'estadístic χ^2. L'ajust correspon als valors α=0.7728, θ=0.5133 estimats pel mètode 5.1.3.*

Mètode 5.1.4				
Aparicions	**Funció densitat**	**$E[\phi(\alpha,\theta)]=\phi(r=k/\alpha,\theta)*V$**	**Observats**	$\frac{(n_i - n\hat{p}_i)}{n\hat{p}_i}$
1	$\phi(r=1/\alpha,\theta)$= 0.4931	$\phi(r=1/\alpha,\theta)*V$= 70.0255	107	19.5230
2	$\phi(r=2/\alpha,\theta)$= 0.3072	$\phi(r=1/\alpha,\theta)*V$=43.6158	16	17.4852
3	$\phi(r=3/\alpha,\theta)$=0.1347	$\phi(r=1/\alpha,\theta)*V$=19.1246	6	9.0069
4	$\phi(r=4/\alpha,\theta)$= 0.0468	$\phi(r=1/\alpha,\theta)*V$= 6.6394	2	3.2418
5	$\phi(r=5/\alpha,\theta)$= 0.0137	$\phi(r=1/\alpha,\theta)*V$= 1.9458	2	0.0015
6	$\phi(r=6/\alpha,\theta)$=0.0035	$\phi(r=1/\alpha,\theta)*V$= 0.5011	2	4.4835
7	$1-\sum_{k=1}^{6}\phi(r=k)$=0.0010	$[1-\sum_{k=1}^{6}\phi(r=k)]*V$=0.1478	7	317.6768
Total	**1**	**142**	**142**	**χ^2=371.5194 p_valor=0**

Taula 6.4: *adjuntem un exemple de càlcul de bondat d'ajust, que correspon al primer capítol del llibre. A la taula s'hi adjunta, la probabilitat teòrica, el valor esperat per la distribució, el valor observat i el valor de l'estadístic χ^2. L'ajust correspon als valors α=17.8149, θ=0.1324 estimats pel mètode 5.1.4.*

El valor de l'estadístic Khi-quadrat l'hem de comparar amb el valor d'una χ^2 amb 6 graus de llibertat, donat que tenim set categories i a un nivell de significació del 0.05. Aquest valor és 12.59. Així doncs, podem veure en les taules 6.1, 6.2, 6.3 i 6.4 respectivament, que les nostres dades del capítol 1 del *Tirant lo Blanc* s'ajusten bé pel mètode basat en el valor esperat i la proporció de paraules que apareixen una vegada (mètode 5.1.1) i també per el mètode dels moments (mètode 5.1.2).

Per els mètodes basats en la proporció de paraules que apareixen una i dues vegades (5.1.3) i pel de màxima versemblança (5.1.4), obtenim un estadístic Khi-quadrat molt superior a 12.59 i per tant, els paràmetres estimats per aquests dos procediments no se'ns ajusten bé a les dades amb una confiança del 95%.

Aquests càlculs es repeteixen per cadascun dels capítols i per cadascun dels mètodes emprats en l'estimació dels paràmetres, a continuació es presenta una taula resum

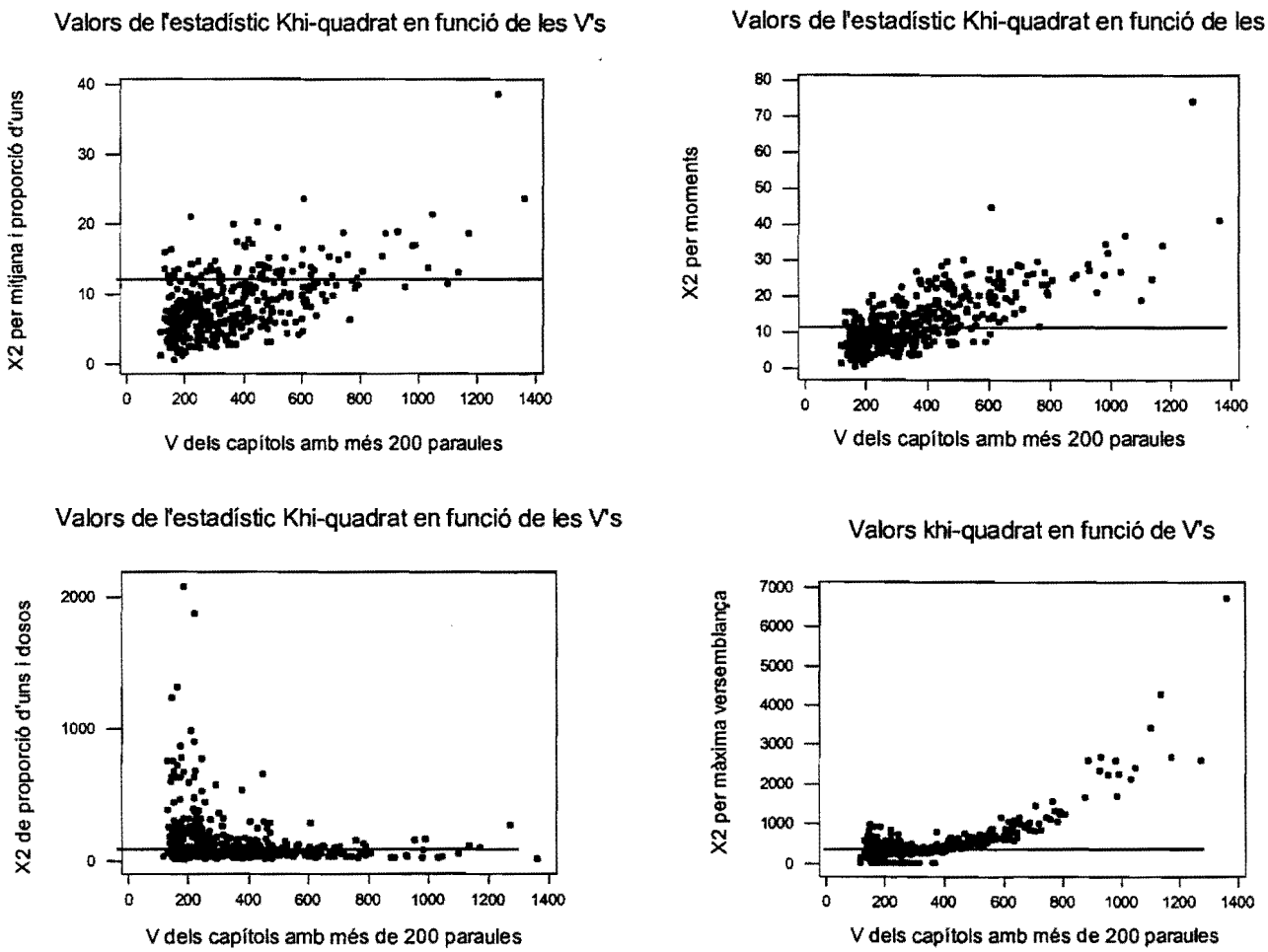
adjuntant el percentatge dels capítols que provenen de la distribució que a priori haviem estimat, és la taula 6.5.

Mètodes	Total de capítols ben ajustats de 425	Percentatge de capítols ben ajustats de 425
5.1.1	313	73.6%
5.1.2	224	52.7%
5.1.3	11	2.6%
5.1.4	31	7.3%

Taula 6.5: *adjuntem el número i percentatge de capítols, amb més de 200 paraules, que s'ajusten bé a la distribució amb els paràmetres ajustats per cada mètode.*

En la taula 6.5 es veu clarament, que el mètode que estima millor els paràmetres de la distribució és el 5.1.1, el que utilitza el moment de primer ordre i la freqüència de paraules que apareixen una sola vegada, shi ajusten bé el 73.6% dels capítols de més de 200 paraules. Tot i amb això, resten 112 capítols que no s'ajusten bé. Per altra banda, només hi ha 224 capítols que s'ajusten bé a la distribució amb els paràmetres estimats pel mètode dels moments (5.1.2). Pel mètode que utilitza la probabilitat que una paraula surti una i dues vegades només s'ajusten bé 11 capítols, que representa el 2,6% del contingut de tots els capítols el llibre. Així doncs, els paràmetres estimats per aquest mètode no els podem utilitzar per esbrinar l'existència d'un canvi d'estil. Els que ens validaran el canvi d'estil amb més fiabilitat són els paràmetres estimats pel mètode 5.1.1. En l'annexe del capítol mostrem els valors observats i esperats en capítols de llargada diferent, així com també el valor χ^2 i el p_valor que n'hem calculat per cadascún dels mètodes.

A continuació estudiem la bondat d'ajust en funció del total de paraules diferents que poden aparèixer en un capítol (V). Per fer-ho, grafiquem els valors de l'estadístic χ^2 per cadascun dels capítols en funció de les paraules que hi surten.



Gràfics 6.1: *es representen els estadístics χ^2 en funció de V (paraules diferents que surten en un text), pels capítols que tenen més de 200 paraules i per cadascun dels mètodes emprats. La línia indica el valor de la Khi-quadrat a partir del qual no s'ajusten bé les dades, 12.59.*

S'observen dos fets significativament diferents en els gràfics 6.1. Primer de tot, cal dir que el comportament dels valors χ^2 en funció de V, calculats quan s'han estimat els paràmetres igualant el valor mitjà i la proporció de paraules que no es repeteixen (mètode 5.1.1) i calculats a partir del mètode dels moments (mètode 5.1.2) tenen un comportament molt semblant. En ambdós observem una correlació positiva, és a dir, a valors grans χ^2 li corresponen valors grans de V. Aquest fet ens indica que per mostres petites la distribució ens ajusta millor que per mostres grans, cosa choerent tenint en compte que per poques dades és més fàcil ajustar, no aquesta sinó qualsevol distribució.

Segurament per mostres grans, al tenir més informació, aquesta no la podem acabar de recullir amb la distribució que hem ajustat.

En canvi, pel mètode de probabilitats de una i dues ocurrències (mètode 5.1.3), la correlació és negativa. Tanmateix, el comportament dels valors χ^2 havent estimat els paràmetres per el mètode 5.1.3 és molt diferent que pels mètodes restants. Però també és ben sabut, que aquest mètode no obté bons resultats en la bondat d'ajust.

A la vista dels resultats decidim que ens els capítols que segueixen a continuació, treballarem amb els estimadors calculats a partir del mètode de la mitjana i proporció de paraules que surten una vegada i el mètode dels moments. S'ha comprovat que són els paràmetres que s'ajusten millor a les nostres dades i per tant, els més consistents.

ANNEXE Capítol 6:

6.1 Valors observats i esperats per cadascun dels mètodes en el capítol 1 (N=255)

Freqüència de les paraules	*Valors observats*	*Valors esperats segons 5.1.1*	*Valors esperats segons 5.1.2*	*Valors esperats segons 5.1.3*	*Valors esperats segons 5.1.4*
1	107	98.5858	97.2812	47.6825	70.0255
2	16	20.9259	20.9880	38.0864	43.6158
3	6	8.8835	9.0562	24.6558	19.1246
4	2	4.7141	4.8846	14.3958	6.6394
5	2	2.8017	2.9507	7.9764	1.9458
6	2	1.7841	1.9098	4.3054	0.5011
7	7	4.3050	4.9295	4.8976	0.1478
χ^2		**6.3192**	**6.0714**	**118.0040**	**371.5194**
p_valor		**0.3884**	**0.4152**	**0**	**0**

6.2 Valors observats i esperats per cadascun dels mètodes en el capítol 3 (N=1174)

Freqüència de les paraules	*Valors observats*	*Valors esperats segons 5.1.1*	*Valors esperats segons 5.1.2*	*Valors esperats segons 5.1.3*	*Valors esperats segons 5.1.4*
1	299	285.3131	276.0141	111.7853	123.3008
2	70	67.1097	66.1690	106.7790	133.0139
3	32	31.5703	31.7255	82.3135	99.2751
4	16	18.5644	19.0139	57.0221	57.6643
5	10	12.2265	12.7630	37.3763	27.8012
6	5	8.6276	9.1791	23.8121	11.5866
7	27	35.5884	44.1354	39.9116	6.3581
χ^2		**5.1445**	**11.7697**	**425.5656**	**438.0676**
p_valor		**0.5254**	**0.0673**	**0**	**0**

6.3 Valors observats i esperats per cadascun dels mètodes en el capítol 47 (N=37)

Freqüència de les paraules	*Valors observats*	*Valors esperats segons 5.1.1*	*Valors esperats segons 5.1.2*	*Valors esperats segons 5.1.3*	*Valors esperats segons 5.1.4*
1	33	33.0577	33.2528	17.0261	33.0984
2	2	1.8463	1.6110	9.9925	1.8081
3	0	0.0913	0.1232	4.7109	0.0890
4	0	0.0044	0.0116	1.9879	0.0043
5	0	0.0002	0.0012	0.7912	0.0002
6	0	0.0000	0.0001	0.3053	0.0000
7	0	0.0000	0.0000	0.1860	0.0000
χ^2		**0.1088**	**0.2321**	**29.3608**	**0.1142**
p_valor		**1.0000**	**0.9998**	**5.1980e-005**	**1.0000**

6.4 Valors observats i esperats per cadascun dels mètodes en el capítol 179 (N=6521)

Freqüència de les paraules	*Valors observats*	*Valors esperats segons 5.1.1*	*Valors esperats segons 5.1.2*	*Valors esperats segons 5.1.3*	*Valors esperats segons 5.1.4*
1	732	731.9817	746.4910	150.0451	70.5480
2	217	198.9880	184.9822	191.1768	148.8863
3	118	98.3098	91.6780	193.0399	215.7289
4	76	60.5579	56.7950	172.4299	241.4222
5	42	41.7624	39.4070	143.8634	222.5659
6	31	30.8534	29.2955	115.4796	176.0514
7	149	202.5468	216.3512	398.9653	289.7974
χ^2		**23.6699**	**41.1112**	**2.6343e+003**	**6.7249e+003**
p_valor		**6.0053e-004**	**2.7533e-007**	**0**	**0**

6.5 Valors observats i esperats per cadascun dels mètodes en el capítol 337 (N=2269)

Freqüència de les paraules	*Valors observats*	*Valors esperats segons 5.1.1*	*Valors esperats segons 5.1.2*	*Valors esperats segons 5.1.3*	*Valors esperats segons 5.1.4*
1	498	455.7899	439.3627	156.4987	149.1911
2	108	109.3426	107.1076	161.0249	191.1060
3	40	52.4619	52.2213	133.2424	170.2498
4	25	31.4636	31.8262	98.7817	118.6508
5	24	21.1345	21.7240	69.1253	68.9865
6	10	15.2103	15.8876	46.9272	34.8478
7	54	73.5971	90.8707	93.3998	25.9680
χ^2		**15.6051**	**29.5379**	**958.1589**	**1102.5**
p_valor		**0.0160**	**4.8110e-005**	**0**	**0**

CAPÍTOL 7:

Mesures de diversitat de vocabulari d'un autor

L'anàlisi estadística de l'estil literari busca característiques quantificables d'un text i les aprofita per caracteritzar l'evolució de la riquesa d'un text i estudiar aspectes d'estil que l'autor rarament controla conscientment. En el nostre cas, caracteritzem l'estil del *Tirant lo Blanc* a través de la riquesa i diversitat del seu vocabulari, mesurada a partir d'un inventari de totes les paraules que hi surten i del nombre d'ocurrències de cada paraula.

En el capítol 3, on haviem començat a mirar les dades, ja hem intuït uns possibles indicadors de la riquesa de vocabulari, aquests eren el nombre de paraules diferents d'un text, V, i també el nombre de paraules que apareixen r vegades en el text, V_r. El principal problema d'aquests índexos és que el seu valor esperat depen molt de la llargada del text N. Per tant, a l'hora de mesurar la riquesa del *Tirant lo Blanc*, si només emprem V i V_r podem arribar a certes confusions a l'hora de decidir si realment existeix un canvi de riquesa en algun capítol del llibre, ja que cada capítol té una llargada N diferent. Per exemple, si observéssim un canvi en V_1 en un cert capítol, podria ser degut a un canvi de N i no a un canvi de riquesa que és el que ens interessa detectar.

Així doncs, en aquest capítol es fa un recull dels diferents indicadors que ens podrien ser útils per estudiar la riquesa i autoria del *Tirant lo Blanc*. Els que més ens interesseran són els que tinguin un valor esperat que no depen d'N. Aquests permeten comparar la riquesa en textes de diferents llargades, com és el cas quan comparem capítols del *Tirant lo Blanc*.

7.1 Indicadors per quantificar l'estil d'un autor

Per caracteritzar l'estil d'un autor a través de la diversitat del seu llenguatge escrit cal suposar que l'escriptor disposa d'un vocabulari composat per la llista de paraules que pot fer servir quan escriu, i que per cada paraula i, existeix un valor π_i que és la proporció de vegades que la paraula i apareixeria en un text seu de llargada infinita. Implícitament es suposa que tant la llista de paraules com els valors π_i són constants al llarg de l'obra de l'autor en estudi (hipòtesis que pot ser discutible quan els textos abasten èpoques o gèneres molt diferents).

Els textos d'un autor són mostres del seu vocabulari i per tant, la freqüència d'aparició de les paraules reflecteix les característiques d'aquest vocabulari. Denotem N al nombre total de paraules que hi ha en un text i per n_i al nombre de vegades que hi apareix la paraula i. D'aquesta manera, si calculem el quocient entre n_i i N, trobem la proporció de vegades que apareix la paraula i en el text, π_i , i si la hipòtesis d'estacionarietat és vàlida, quan N augmenta, π_i tendeix a aproximar-se a π_i.

Tal com hem anat veient, donat un text de llargada N, denotem per V, al nombre de paraules diferents i per V_r, al nombre de paraules diferents que hi apareixen r vegades. El quocient $p_r=V_r/V$ és la proporció de paraules que hi apareixen r vegades, que estima la probabilitat que una paraula es repeteixi r vegades.

La majoria de mesures de diversitat del vocabulari es basen en la proporció de paraules diferents que hi apareixen r vegades i s'estimen a través de p_r observades. Donat un text amb un total de N paraules, quant més gran sigui el nombre de paraules diferents V, més ric i divers és el llenguatge, mentre que per unes N i V donades, quant més gran és pr per r petites i més petit pr per r grans, més divers és el text. Qualsevol índex de diversitat ha de respectar aquestes dues ordenacions.

L'índex més simple per quantificar la diversitat del llenguatge d'un text, compta el nombre de paraules diferents que s'utilitzen i l'anomenem V. En el gràfic 3.2 s'observava,

que V depèn de N, de forma que N pot créixer indefinidament, mentre V queda limitada pel nombre de paraules que composen el vocabulari de l'autor. Així doncs, V només servirà per a comparar textos de la mateixa llargada.

Un segon índex de diversitat és V_1, (o *Hapax legomena*) nombre de paraules que només apareixen una vegada en un text. Donada una N constant, quant més gran sigui V_1, més divers serà el text de l'autor. En aquest cas també tenim que V_1 només ens és útil per comparar textos de la mateixa llargada.

Com alternatives a V i V_1, Simpson (1949) proposa l'índex D, que és la probabilitat que agafant dues paraules a l'atzar d'un text, aquestes dues siguin iguals, i es calcula de la següent forma:

$$D = \frac{\sum_r r(r-1)V_r}{N(N-1)} = \frac{\sum_i n_i(n_i-1)}{N(N-1)} \tag{7.1}$$

Quant més petit sigui l'índex D, proposat ja en un altre context per Gini (1912), més divers serà el lèxic del text (menor és la probabilitat de repetició de dues paraules qualsevols). El complementari de l'índex Simpson seria la probabilitat que dues paraules siguin diferents i es calcularia 1-D.

L'índex de Simpson és més útil que V i V_1 gràcies a una particularitat, que el seu valor esperat depen de la diversitat de vocabulari de l'autor, però no depen de la llargada del text N. Tanmateix, la variabilitat d'aquest índex si que depèn de N. Els valors esperats de Simpson són molt més variables quan la N és petita. Així doncs, D és un indicador que ens permet estimar la diversitat de lèxic d'un autor "sense biaix" i permet comparar les diversitats de textos de llargades diferents.

Una altra mesura alternativa és l'entropia, que mesura el grau de concentració de la probabilitat en unes poques paraules de totes les que composen el text i es calcula:

$$H = -\sum_i \pi_i \log \pi_i = -\sum_i \frac{n_i}{N} \log \frac{n_i}{N} = -\sum_i \frac{rV_r}{N} \log \frac{r}{N} \tag{7.2}$$

Quant més gran és l'entropia H, més "desordenat" és el text, i és el que ens indicarà la riquesa del lèxic. El problema d'aquest indicador és que el seu valor esperat no coïncideix amb l'entropia teòrica del vocabulari de l'autor; en augmentar *N*, l'entropia de qualsevol text tendeix a disminuir i per tant *H*, tampoc serveix per comparar textos de llargades diferents.

De totes aquestes mesures, les més útils són el nombre de paraules diferents, *V*, gràcies a la seva simplicitat, i l'índex Simpson *D*, perquè no depèn de la llargada del text. Ambdues són fàcils d'interpretar. Nosaltres hem fet èmfasi en aquests índexs de riquesa perquè són els que hem calculat, però n'existeixen d'altres els quals llistem i expliquem breument a continuació:

-*V/N*: quocient forma/ocurrència (o quocient tipus/token). És un índex molt senzill, però la seva distribució segueix depenent de la llargada del text, *N*, ja que no és proporcional a *N* perquè creix més lentament. A la pràctica, si es fixa el nombre d'ocurrències, *N*, el qüocient forma/ocurrència és exactament igual que *V*, i com més gran és *V/N*, més ric és el llenguatge.

-V_2: *Hapax dislegomena* o número de paraules que apareixen dues vegades en un text. Valen les mateixes consideracions que per V_1.

-***R***: índex d'*Honoré* (Honoré(1979)), mesura la propensió d'un autor a escollir entre emprar una paraula ja usada prèviament o utilitzar-ne una de nova. Es calcula mitjançant l'expressió:

$$R = \frac{100 \log N}{1 - \frac{V_1}{V}}$$

Com més gran és *R*, més ric és el llenguatge, perquè un nombre major de paraules són usades poc freqüentment. La seva distribució també depen de la llargada del text, *N*, però Honoré comprova empíricament com, a partir d'una certa llargada del text (*N>1300*), el seu valor esperat es satura.

-***V₂/V***: Relació entre els *Hapax dislegomena* i el nombre de paraules diferents. Holmes i Forsyth (1995) l'usen en l'anàlisi d'alguns texts.

-***M***: índex de *McIntosh* (McIntosh, 1967), basat en consideracions geomètriques. Es calcula:

$$N = \frac{N - \sqrt{\sum_{r=1}^{V} V_r^2}}{N - \sqrt{N}}$$

Per una *N* donada, com més gran é *M*, més ric és el llenguatge.

-***W***: índex de *Brunet* (Brunet, 1978), utilitzat per Holmes i Forsyth (1995). Es defineix com:

$$W = N^{V^{-a}}$$

on *a* és una constant que pren un valor comprès en l'interval [0.165, 0.172]. Brunet assegura que la distribució de *W* depen molt poc de la llargada del text i que és específic de cada autor. Com en el cas de l'índex de Simpson, com més petit és *W*, més ric és el llenguatge.

Sichel (1986) i Yule (1944) plantegen el debat sobre si aquestes mesures de diversitat s'han de calcular sobre paraules de tota mena o bé, és millor restringir-se a paraules d'alguna classe especial com substantius, advervis, verbs o preposicions i conjuncions. Nosaltres no desagreguem el vocabulari segons les funcios de les paraules. Un segon dilema és el de si

lematitzar prèviament el text, és a dir, reduïr totes les seves paraules a diccionari. Algunes regles de lematització proposades converteixen les formes verbals en el seu infinitiu, els substantius en substantius singulars, els adjectius en adjectius masculins singulars etc. Per textos petits es pot lematitzar a mà, però fins i tot en aquest cas alguns dels criteris de lematització són discutibles. Seguint l'exemple de Mosteller i Wallace(1984), nosaltres no lematitzem el *Tirant*.

Per més detalls sobre l'ús d'aquestes mesures podeu veure Good(1953), Good i Toulmin (1956), Margalef (1958), Efron i Thistedt (1979, 87), Paril i Taillie (1982), Harris (1982), Manly (1994), Holmes (1985), i Sichel (1986 a,b). Part d'aquesta literatura està escrita pensant en la caracterització de la diversitat d'un ecosistema on el nombre d'espècies correspon al nombre de paraules diferents i l'àrea de l'ecosistema a la llargada del text.

*7.2 L'índex de riquesa K**

Com ja hem vist, en l'apartat 4.3 del projecte s'explica amb detall la distribució que proposa Sichel, definida com una barreja de distribucions en la qual la distribució inicial és una Poison de paràmetre $N^*\pi$ i la distribució de π és una Gausiana inversa generalitzada. Recordem que els moments de primer i segon ordre d'aquesta distribució, fixant el paràmetre γ a $-1/2$ es defineixen com:

$$\mu_1(r|N) = E(r|N) = \frac{1}{2}bcN\left[1 - \exp(-b(\{1+cN\}^{1/2} - 1))\right]^{-1}$$

$$\mu_2(r|N) = E(r^2|N) = \frac{1}{2}bcN\left[1 + \frac{1}{2}(1+b)cN\right]\left[1 - \exp\left(-b\left(\{1+cN\}^{1/2} - 1\right)\right)\right]^{-1}$$

Així doncs, el coeficient de dispersió es defineix com el quocient entre $\mu_2(r|N)$ i $\mu_1(r|N)$. Desenvolupant aquest quocient obtenim el coeficient de dispersió, el qual depen dels paràmetres de la distribució de Sichel fixant el paràmetre γ a $-1/2$:

$$\varpi(r|N) = \frac{\mu_2(r|N)}{\mu_1(r|N)} = 1 + \frac{1}{2}(1+b)cN - \mu_1(r|N)$$

Quan N tendeix a un valor molt gran aquest coeficient tendeix a incrementar-se linealment i es pot aproximar amb la següent relació, la qual depen de N i del paràmetre c:

$$\varpi(r|N) \sim 1 + \frac{1}{2}cN$$

Definim el coeficient de variació com:

$$CV(r|N) = \frac{\mu_2(r|N)^{\frac{1}{2}}}{\mu_1(r|N)} = \left[\frac{1 + \frac{1}{2}(1+b)cN}{\mu_1(r|N)} - 1 \right]^{\frac{1}{2}}$$

Aquest coeficient s'incrementa a mesura que s'incrementa N. Assimptòticament, quan N tendeix a infinit, podem establir la següent relació:

$$\lim[CV(r|N)] = b^{\frac{-1}{2}} \quad \text{quan } N \to \infty$$

A partir d'aquests indexos, els quals ens mesuren d'alguna manera la dispersió i la variació en un text, és interessant fer referència a l'índex de riquesa que va proposar Yule (1944), definit com:

$$K^* = \frac{10^4}{N}\left[\frac{\mu_2(r|N)}{\mu_1(r|N)} - 1 \right]$$

Desenvoluant K^* a partir de les equacions dels dos primers moments, obtenim l'índex de riquesa de Yule en funció dels paràmetres de la distribució de Sichel, b i c, fixant el paràmetre γ a $-1/2$. Així doncs:

$$K^* = \frac{10^4 c}{N}[1+b] \tag{7.3}$$

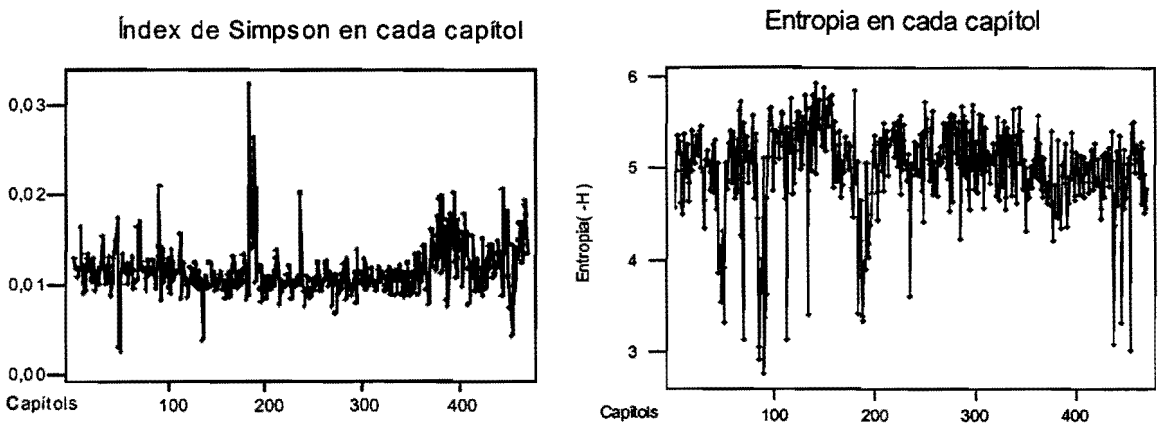
Per tant, fent referència als resultats obtinguts en l'estimació dels paràmetres b i c, que es presenten en el capítol 5, observem que l'índex de riquesa K^*, no és més que el mateix paràmetre c de la distribució de Sichel canviat d'escala, donat que b és molt proper a 0. Com l'índex de Simpson D, en aquest cas la distribució de K^* depen de c i de b i per tant, no hauria de dependre de N.

7.3 Índexos de riquesa en el Tirant lo Blanc

En aquest apartat del capítol ens centrem a descobrir el comportament dels indicadors de riquesa definits en l'anterior secció, l'índex de Simpson, l'Entropia i l'índex K^*.
Com ja s'havia avançat en estudis anteriors, en el *Tirant lo Blanc*, s'observa un canvi de patró en l'índex de Simpson a partir del capítol 380, aproximadament. Recordem que l'índex de Simpson estimava la probabilitat que, agafant dues paraules a l'atzar, aquestes siguin iguals. Observi's que a partir d'aquest punt, els índexos es distribueixen en l'entorn d'una constant més elevada, ens trobem en el cas d'un canvi en la constant. Per altra banda, la variància també sembla que augmenti en aquest moment del llibre.

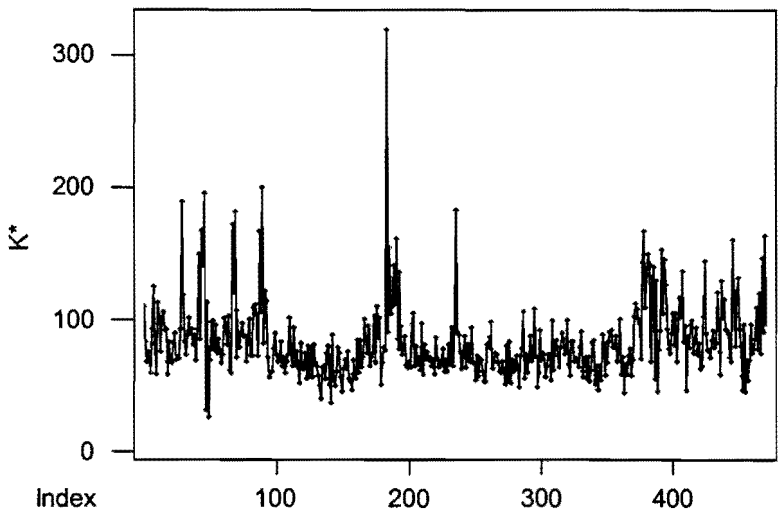
Cal destacar també, que els capítols 183, 187 i 188 tenen un índex de Simpson molt gran. Aquests capítols es destaquen per ser molt curts ja que tenen 68, 54 i 43 paraules, respectivament. Així doncs, intuïm que aquests índexos de riquesa es poden relacionar amb la grandària del text. Ens interessa estudiar un índex de riquesa que sigui comparable per tots els capítols, és a dir, un indicador que no depengui de la llargada del text. Hi ha precedents, que ens indiquen que Simpson és un bon candidat, perquè el seu valor esperat no en depen. Passem a estudiar la relació entre les paraules d'un text i els índexos d'Entropia, Simpson i K^*. El primer que presentem, són aquests indicadors per cadascun dels capítols del *Tirant lo Blanc*.

Avancem dient que l'index Entropia *H*, com ja veurem en els gràfics 7.5 depen del tamany del text *N*. Per tant, representar aquest índex en funció de tots els capítols només té sentit per veure'n la seva evolució i no per intentar trobar algun canvi en la riquesa del vocabulari en el *Tirant lo Blanc*.



***Gràfics 7.1:** es presenten els indexos Simpson(D) i Entropia(H), que fan referència a la riquesa del text, en el desenvolupament de tot el llibre.*

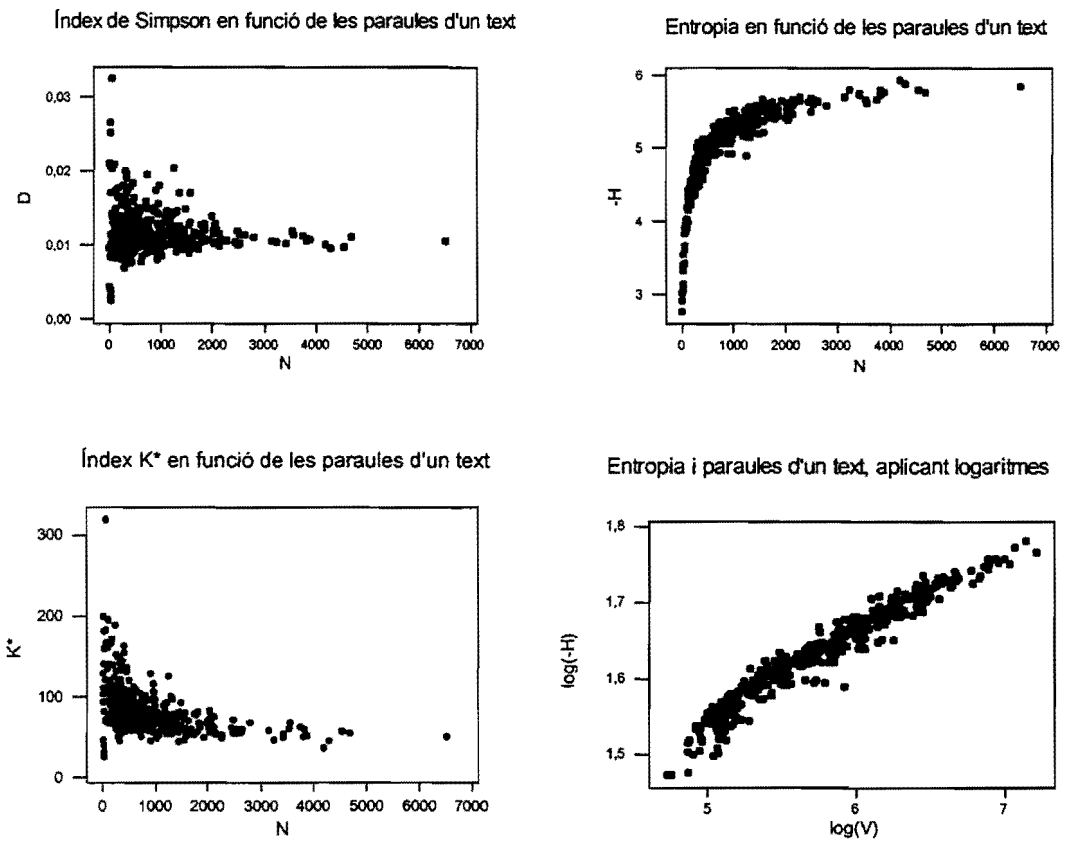
Tal i com ens indiquen els gràfics 7.1, el valor esperat de l'índex de Simpson presenta un cert increment i canvi de variabilitat al voltant del capítol 380. Això doncs, ja ens pot fer pensar amb un canvi d'estil i riquesa en el vocabulari del *Tirant lo Blanc* a partir d'aquest punt. En l'Entropia aquest canvi no s'observa i en general, presenta una variabilitat força més elevada que la de *D*. En *H* es pot veure un petit canvi al voltant del capítol 150 tot i que no es pot assegurar amb prou fermesa degut a la forta variabilitat. A més, això no ens indica cap canvi d'estil perquè aquest index depen de la llargada dels capítols.



***Gràfic 7.2:** index de riquesa K* al llarg de tots els capítols del Tirant lo Blanc.*

En el gràfic 7.2 podem observar l'índex de riquesa K^* per tots els capítols del llibre. Veiem calarament que el seu comportament és molt semblant al de l'índex de Simpson, D, ja que els gràfics tenen pràcticament la mateixa forma. Així doncs, l'índex K^*, igual que ens passa amb D, també ens marca un canvi en la riquesa del vocabulari del *Tirant lo Blanc* al voltant del capítol 380. En els primers capítols també hi podem veure una variabilitat més gran.

A continuació passarem a explorar la relació d'aquests índexos amb N, que és el número total de paraules del capítol. Observi's, en els gràfics 7.3, com els índexos de Simpson i K^* es distribueixen al voltant d'una constant. En augmentar el tamany d'un text els valors esperats de D i K^* no canvien, és la seva variabilitat el que canvia, ja que per capítols curts els índexos es troben molt dispersats i, a mesura que augmenten les paraules d'un text disminueix la variabilitat.



Gràfics 7.3: *Índexos de riquesa (Simpson a dalt a l'esquerra, Entropia a dalt a la dreta i K* a sota a l'esquerre), en funció del nombre de paraules dels capítols del Tirant lo Blanc. A sota a la dreta hi presentem l'Entropia linearitzada.*

Pel què fa a l'Entropia, s'observa que es correlaciona positivament amb la grandària d'un text. Aquesta relació, però, no és lineal, com es demostra en els gràfics 7.3, la relació passa a ser lineal prenent logaritmes en les dues variables. Així doncs, l'Entropia només es pot utilitzar si es comparen textos amb el mateix tamany. Cal destacar un fet curiós, el comportament de la V en funció de les paraules d'un text N, mostrat en el capítol 3 en els gràfics 3.2 és molt semblant al comportament de l'Entropia en funció de les paraules diferents que apareixen al text (V).

Així doncs, vistos aquests resultats, els índexos més interessants per treballar la riquesa de vocabulari d'un text, i en el nostre cas concret del *Tirant lo Blanc*, són l'índex de Simpson i l'índex K^*. Recordem que K^* depen dels paràmetres b i c, i per tant, per poder calcular aquest índex, ens ha estat necessari ajustar les nostres dades a la distribució de Sichel. D'aquesta manera hem pogut veure que el paràmetre c també està relacionat amb la riquesa i la diversitat de vocabulari d'un text. A continuació passarem a veure quines relacions hi ha entre els paràmetres estimats, b, c, α i θ amb l'Entropia i Simpson.

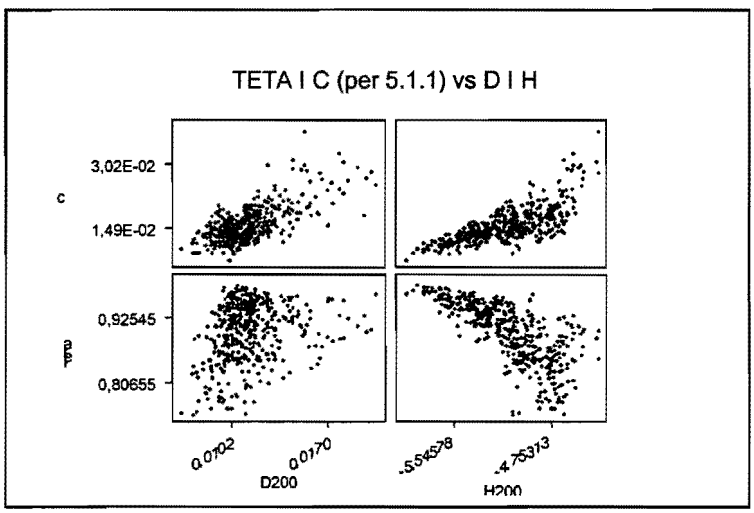
7.4 Relació entre els paràmetres de la distribució i altres indexos de riquesa

Anteriorment, hem comprovat que les dades s'ajusten millor a la distribució amb els paràmetres estimats pel mètode basat en el valor esperat i la proporció de paraules que apareixen una vegada en el text (5.1.1). També hem vist, que els paràmetres estimats pel mètode que proposàvem, "igualar les probabilitats teòriques a les observades, que una paraula surti un i dos cops" (5.1.3) i també el mètode de la màxima versemblança (5.1.4), no s'adequaven a les nostres dades. Així doncs, aquests mètodes els descartem i treballem només amb els paràmetres estimats pels dos primers mètodes, els 5.1.1 i 5.1.2.

Cal tenir en compte, també, que els paràmetres α's i b tenen valors molt propers a 0 o 0, i per tant, no ens aporten informació. És per això, que decidim explorar la informació dels paràmetres c i θ. Per tant, a partir d'ara, els gràfics que presentem es relacionaran amb aquests dos paràmetres.

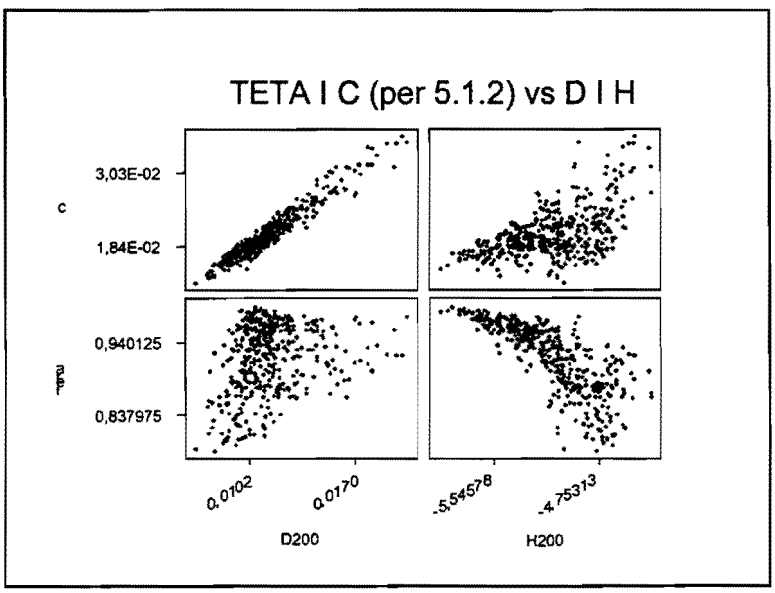
En aquest apartat busquem relacions entre aquests paràmetres i els altres indexos de riquesa que hem presentat en les seccions anteriors, l'índex de Simpson D i l'entropia H. Les correlacions es poden observar en els gràfics 7.4 i 7.5. Hi ha correlació positiva entre el paràmetre c, tant en l'índex de Simpson D, com en l'entropia H. Mentre que la correlació entre el paràmetre θ i l'entropia és negativa. La correlació entre θ *i* D és menys significativa.

Per altra banda, si comparem els mètodes emprats per estimar els paràmetres, ens n'adonem que la correlació entre c i D és molt més acusada en el mètode dels moments (5.1.2), en canvi, la correlació entre H i c, és més gran en el mètode (5.1.1).



	Correlacions amb *D*	Correlacions amb *H*
c	0,676	0,730
θ	0,252	-0,719

Gràfic 7.4:*Presentem els paràmetres θ i c, estimats per el mètode definit en 5.1.1 en funció dels indexos de riquesa Simpson , que l'anotem D, i entropia, que l'anotem H, juntament amb les seves correlacions Pearson.*



	Correlacions amb *D*	Correlacions amb *H*
c	0,958	0,538
θ	0,260	-0,766

Gràfic 7.5:*Presentem els paràmetres θ i c, estimats per el mètode definit en 5.1.2 (moments) en funció dels indexos de riquesa Simpson , que l'anotem D, i entropia, que l'anotem H, juntament amb les seves correlacions Pearson.*

Recordem que l'índex de Simpson estimava la probabilitat que agafant dues paraules a l'atzar, aquests fossin iguales. L'entropia, mesura com "de desordenat" és un text. Així doncs, un valor alt d'entropia es relaciona amb un text més ric, mentre que un valor petit de *D*, s'interpreta que el text és més ric. Tenint en compte els resultats en les correlacions que s'adjunten en els gràfics 7.4 i 7.5, d'alguna manera també podem relacionar el paràmetre *c* de la distribució amb la riquesa del text. És a dir, un capítol que té un valor alt de *c*, s'associa amb un índex de Simpson alt, i per tant, s'intueix que el vocabulari és poc ric. Curiosament, observem també que la correlació entre *D* i *c* és força més elevada en el mètode dels moments (5.1.2), el qual, tal i com hem vist en el capítol 6, no ens ha ajustat les dades tan bé com el mètode 5.1.1.

Pel que fa a l'Entropia *H*, a simple vista observem que es relaciona positivament amb el paràmetre *c*, la qual cosa ens indicaria que com més gran sigui *c* més ric és el vocabulari. Però aquesta relació, a part de ser contradictòria amb la relació que acabem d'explicar entre *c* i *D*, ens pot dur a certes confusions ja que *H* depen de *N* i *c* no en depen. De la mateixa manera la correlació negativa que observem entre *H* i θ és dubtosa.

Per acabar, hem d'esmentar un altre indicador interessant, el qual també podem relacionar amb la riquesa del text, són totes les paraules que l'autor coneix. Aquest valor el podem estimar de la següent manera:

$$V_a = V_0 + V = \left(\frac{2}{b*c}\right),$$

On V_0 són les paraules que l'autor coneix però no escriu i V les paraules diferents que apareixen en el text de l'autor. Així doncs podem estimar el valor V_0:

$$V_0 = \left(\frac{2}{b*c}\right) - V$$

Recordem que el paràmetre b el calculàvem a partir de α i θ de la següent forma:

$$b = \alpha(1-\theta)^{\frac{1}{2}}$$

Per tant, en el nostre cas, com que els valors estimats α són 0 o molt propers a 0, la b també és 0 i per tant, no hem pogut estimar les paraules que l'autor coneix i que no escriu, V_0.

CAPÍTOL 8:

Ajust de la distribució de Sichel per blocs del Tirant lo Blanc

En el capítol 5 hem vist que l'estimador α que hem calculat, gairebé sempre té un valor nul. Això ens fa pensar que, degut al reduït tamany mostral que tenim quan treballem per capítols, no tenim prou informació per estimar correctament α i θ. També hem vist en l'exemple proposat per Sichel, que la mostra que utilitza per estimar els paràmetres conté més de vuit mil paraules. En el nostre cas, disposem de pocs capítols que sobrepassin les tres mil paraules. És per això que decidim ajuntar les nostres dades formant blocs de capítols que contenen un número de paraules equiparable a l'exemple proposat.

Anàlogament amb el capítol 5 es volen estimar els paràmetres α i θ de la distribució de Sichel al text del *Tirant lo Blanc*, però en aquest cas, en comptes de fer-ho per capítols, ho farem per blocs de capítols. És a dir, ajuntarem les paraules del text obtenint 34 blocs de 14 capítols i un bloc de 13. Per fer això, partirem d'una base de dades diferent de la que hem utilitzat fins ara, la qual expliquem en el primer punt i que ja ha estat comentada en el capítol 3. Tot seguit presentarem els resultats de l'estimació dels paràmetres pels dos primers mètodes ja explicats en el capítol 5, els quals són els que ens han ajustat millor.

Finalment calculem l'estadístic Khi-quadrat per comprovar si les nostres dades s'ajusten bé a la distribució de Sichel per els paràmetres estimats per ambdós mètodes.

8.1 Base de dades utilitzada

La base de dades de la qual partim en aquest capítol ja les hem mostrat en la taula 3.2 del capítol 3. Hem vist que disposem de tot el llistat de les diferents paraules que conté el *Tirant lo Blanc*, un total de 13828. Per cada paraula en tenim la seva freqüència dins de cada capítol. És a dir, partim d'una matriu de dades de dimensió 13828 x 489. Recordem que el llibre té 487 capítols, però el 71 i el 107 estan subdividits en 71a, 71b, 107a i 107b, i és per això que tenim 489 columnes i no 487. De fet, la base da dades que hem utilitzat en el capítol 5 prové d'aquesta, ja que d'aquí en podem extreure quantes paraules apareixen una, dues, tres, etc vegades en cada capítol.

Tal com ja hem explicat en el capítol 3, per formar els blocs sumem les freqüències de les paraules en grups de 14 capítols i un últim de 13. D'aquesta manera, en comptes de tenir 489 capítols, passem a tenir 35 blocs de capítols i per tant, una matriu de dimensió 13828 x 35. Aquestes dades les hem mostrat de forma clara en la Taula 3.2. del capítol 3 que ens ilustra la freqüència de les paraules per blocs de capítols. Així doncs, tenint en compte la divisió del llibre per blocs de capítols, observem que la paraula més freqüent apareix un total de 1542 vegades en el bloc 12. Això ens porta a definir la base de dades que ens serà útil per estimar els paràmetres de la nostra distribució de vocabulari. Aquestes dades consten d'una primera columna que ens indica el bloc i 1542 columnes que ens indiquen el nombre de paraules que apareixen repetides r vegades, des de $r = 1$ fins a $r = 1542$, que corresponen al comptatge del nombre de paraules en cada una de les aparicions en els 35 blocs que hem format del llibre. Per entendre millor aquestes dades en presentem una part en la taula 9.1. També hi mostrem la llargada dels blocs N, per fer èmfasi en els variables tamanys d'aquests.

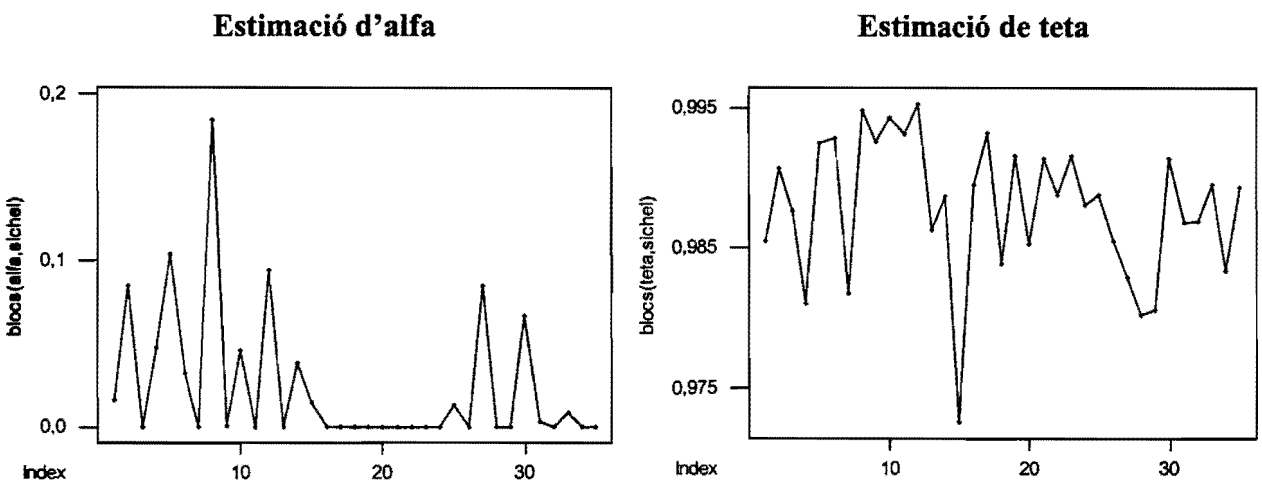
Observem, per exemple, que en el primer bloc, format pels 14 primers capítols del *Tirant* conté 945 paraules que no es repeteixen, 268 que surten 2 vegades i així succesivament. Concretament aquest bloc està format per 7957 paraules. A partir d'aquí procedirem a calcular els 35 estimadors corresponents a cada bloc.

Blocs	N_i	r																	
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	...
1	7957	945	268	126	71	57	45	26	21	11	18	15	9	13	6	3	4	4	...
2	11629	1038	311	137	87	80	46	32	28	21	18	14	14	12	13	8	8	8	...
3	8234	922	259	113	73	47	36	21	27	21	13	7	10	10	6	7	2	0	...
4	5131	679	212	91	62	30	30	12	16	9	7	9	6	2	1	3	6	2	...
5	14233	1122	352	155	86	81	50	27	34	23	22	18	18	8	14	12	13	6	...
6	12761	1052	289	142	84	77	43	26	28	21	16	14	10	12	10	18	4	11	...
7	5036	718	163	89	61	29	20	23	12	8	9	5	6	4	4	1	1	4	...
8	24986	1514	517	241	141	92	84	60	32	37	37	23	22	20	20	12	13	6	...
9	16093	1388	373	187	121	82	51	38	32	23	29	19	8	16	17	8	9	9	...
10	22212	1613	470	216	141	104	75	46	44	33	25	20	16	17	14	9	15	15	...
11	20601	1731	489	223	130	75	67	67	35	36	29	28	11	20	16	12	9	7	...
12	28976	1825	563	267	159	120	81	47	55	46	27	37	21	16	8	14	19	14	...
13	8234	986	254	123	77	52	31	25	20	27	16	8	6	11	5	4	6	5	...
14	11141	1143	337	152	107	63	54	37	31	20	13	10	11	4	7	10	3	9	...
15	3441	564	158	80	37	31	24	16	12	3	5	5	4	1	5	1	1	4	...
16	11384	1179	313	154	91	66	34	41	34	19	17	15	5	6	9	5	10	5	...
17	15180	1259	318	175	106	69	56	32	34	23	13	20	17	8	8	13	8	13	...
18	6249	827	193	90	61	54	29	18	13	7	9	5	14	7	8	7	9	5	...
19	12795	1184	313	159	96	66	39	37	28	21	16	11	12	7	17	10	13	14	...
20	8482	1060	280	131	73	52	38	25	20	16	14	19	6	15	3	4	8	5	...
21	13948	1305	365	155	120	73	45	32	32	21	19	18	11	15	12	5	12	7	...
22	11169	1202	335	148	86	58	36	45	23	21	13	17	13	11	9	9	7	7	...
23	14220	1322	360	165	106	83	52	24	28	21	24	20	12	10	14	14	8	8	...
24	10199	1131	314	159	66	54	41	31	22	17	8	18	16	8	13	9	6	9	...
25	10209	1066	315	145	64	65	40	32	27	14	18	14	16	12	9	15	4	2	...
26	10180	1264	374	134	82	61	24	29	28	23	14	19	16	10	3	6	6	7	...
27	7891	959	308	150	71	58	32	24	22	17	7	11	14	7	4	3	5	2	...
28	6391	976	226	98	57	45	30	21	17	7	8	11	8	8	5	9	2	4	...
29	5997	898	209	111	50	31	31	22	21	13	3	7	7	5	7	2	4	4	...
30	10808	946	277	126	88	58	49	36	20	13	21	13	11	6	13	10	5	7	...
31	7606	872	248	104	81	45	22	25	19	22	16	12	10	11	7	4	8	5	...
32	7520	867	243	109	58	46	40	25	19	16	17	11	10	4	6	7	8	3	...
33	9599	978	275	125	83	56	45	26	23	26	10	12	11	7	3	11	4	6	...
34	7924	1070	303	116	63	45	32	23	26	16	12	9	8	6	12	8	7	4	...
35	9826	1052	284	138	74	34	33	19	25	27	13	14	6	11	10	11	2	4	...

Taula 8.1: *Exemple de la base de dades on presentem la seqüència V, per r=1 fins a r=17 dels 35 blocs de capítols del Tirant lo Blanc. Els blocs marcats amb gris, corresponen als blocs que s'ajusten bé per la distribució Sichel a través del mètode 5.1.1.*

8.2 Estimació basada en el mètode 5.1.1

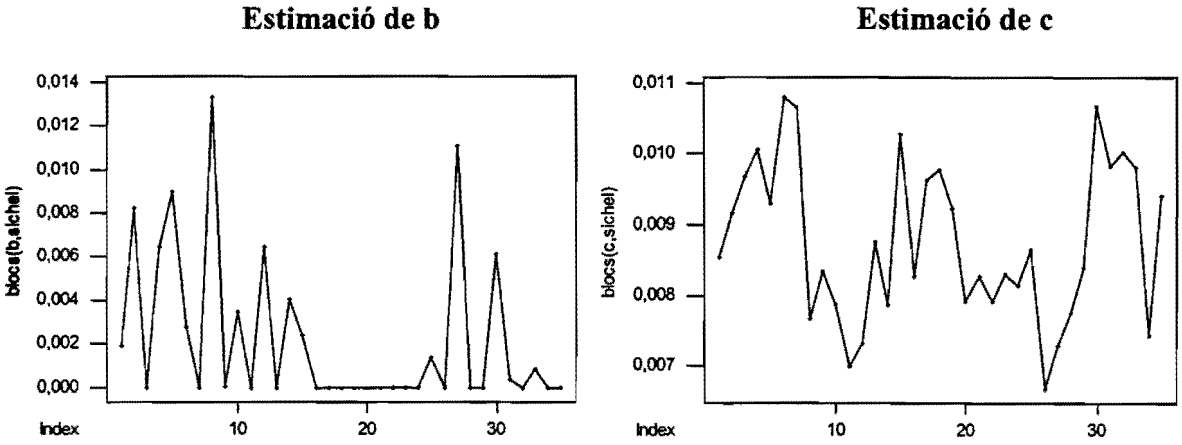
En l'annexe 8.1 presentem una taula dels estimadors d'α i θ per cadascun dels blocs que hem format a partir dels 489 capítols del *Tirant lo Blanc*, calculats a partir d'aquest primer mètode. A continuació, en els gràfics 8.1 observem l'evolució dels estimadors a través dels blocs.



Gràfic 8.1: *Paràmetres estimats per α i θ ordenats pels 35 blocs*

Observem que els valors d'α estimats continuen essent molt propers a zero i que a partir del bloc 15 són més baixos. En el gràfic de les θ's observem que el bloc 15 presenta també un valor més baix. Aquest bloc es correspon als capítols que van entre el 209 i el 223, els quals tenen un total de 3441 paraules. En el bloc 26, que consta de 10180 paraules i el comprenen els capítols del 374 al 388, es pot veure que l'α mostra un petit increment i la θ disminueix lleugerament. Però a la vista d'aquest resultats no podem assegurar que hi hagi un canvi d'estil en cap dels blocs, ja que tot i veure algun canvi de tendència sobre aquests valors no ens determinaran un canvi en la riquesa del vocabulari, perquè és sabut que α i θ depenen de la llargada del text (N).

A continuació presentem els gràfics pels paràmetres b i c reparametritzats a partir dels α i θ.

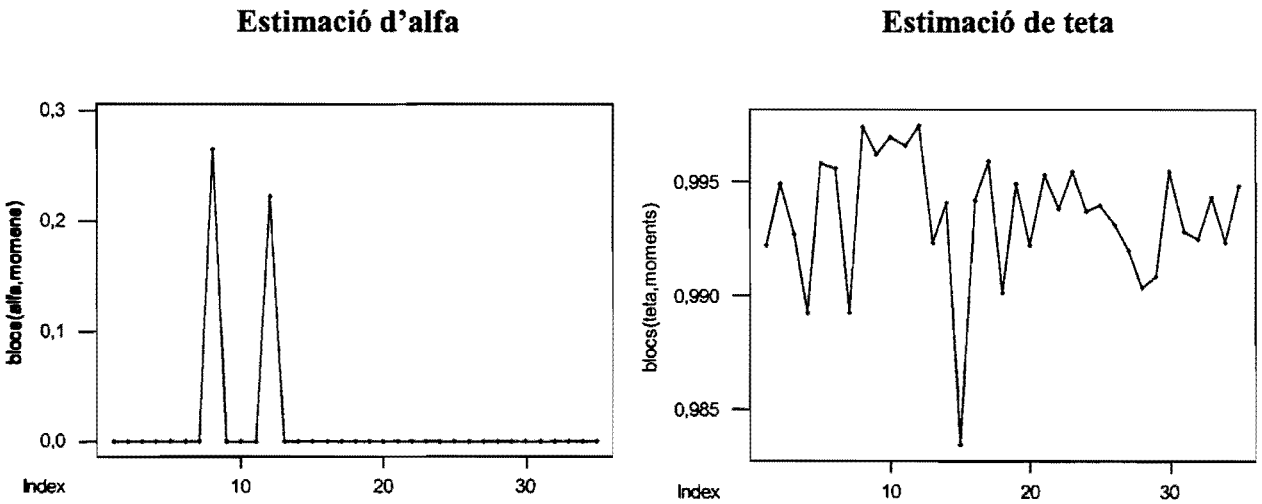


***Gràfic 8.2:** Paràmetres estimats per α i θ ordenats pels 35 blocs*

En els gràfics 8.2 veiem que el paràmetre b té el mateix comportament que l'α, presentant un valor més baix a partir del bloc 15. En l'estimació de c observem dos canvis, un en el bloc 26 (capítols del 374 al 388) igual que en les θ's, i l'altre a partir del bloc 10, el qual està comprès pels capítols 134 a 148. Com ja hem vist en el capítol 7, el paràmetre c està relacionat amb la riquesa del vocabulari i a més, el seu valor esperat no depen de N, la qual cosa ens pot fer pensar en un canvi de riquesa en el bloc 10 i en el 26.

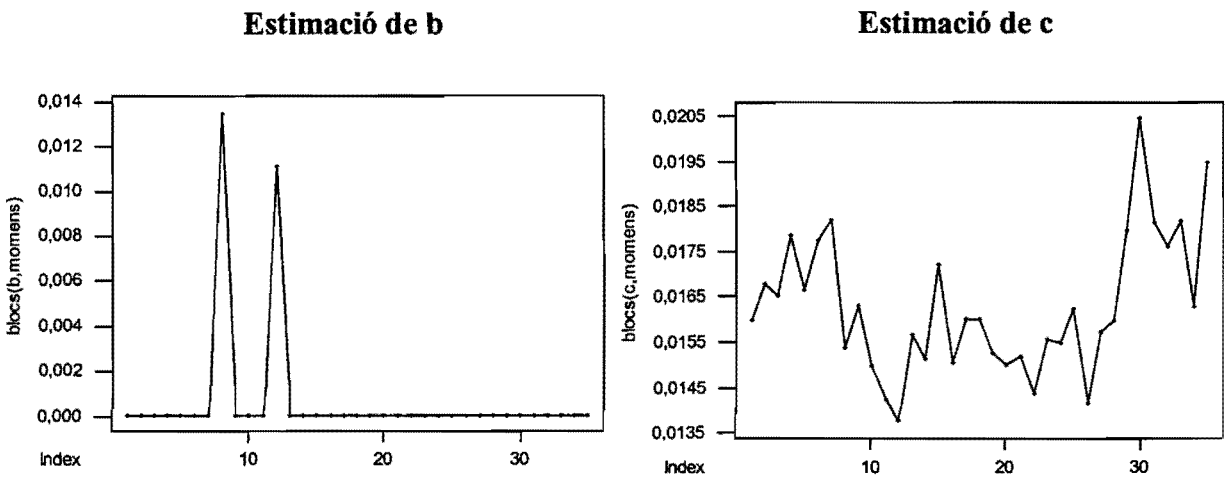
8.3 Estimació basada en el mètode 5.1.2

En l'annexe 8.1 presentem una taula dels resultats amb els estimadors d'α i θ per cadascun dels blocs que hem format a partir dels 489 capítols del *Tirant lo Blanc*, calculats a partir del mètode dels moments. A continuació, en els gràfics 8.3 i 8.4 observem l'evolució dels estimadors a través dels blocs.



Gràfic 8.3: *Paràmetres estimats per α i θ ordenats pels 35 blocs*

Observem que els valors d'α són 0, tret dels blocs 8 i 12. En l'estimació de les θ's tornem a veure pel mètode dels moments que a partir del bloc 26 disminueix lleugerament el seu valor i a partir del 30, corresponent als capítols del 434 al 448, es torna a produir un cert increment. Igual que abans, no podem veure cap canvi amb prou claredat, i si l'observéssim no podriem afirmar un canvi de riquesa amb aquests estimadors. A continuació grafiquem els estimadors *b* i *c*.



Gràfic 8.4: *Paràmetres estimats per α i θ ordenats pels 35 blocs*

En el gràfic 8.4 veiem que el paràmetre b té el mateix comportament que l'α, presenta valors nuls llevat dels blocs 8 i 12. En els valors de c tornem a veure-hi l'increment que es produeix a partir del bloc 30.

8.4 Resultats de la bondat d'ajust

En els dos punts anteriors hem tornat a veure que, tot i agrupar les dades per blocs per tal de tenir un major tamany mostral, seguim tenint unes estimacions dels paràmetres α i b molt properes a zero, especialment en el mètode dels moments. En aquest punt veurem si els paràmetres estimats s'ajusten bé a les dades emprant el test Khi-quadrat, anàlogament amb el capítol 6. Recordem que aquesta prova de bondat d'ajust compara la discrepància entre els valors observats i els valors esperats per la distribució teòrica, que en aquest cas es tracta de la distribució mixtura d'una Poisson i una distribució Gaussiana inversa generalitzada.

Alhora de fer el test hem agrupat les dades en classes diferents de les que havíam agafat en el capítol 5, ja que no és el mateix treballar per capítols que per blocs de capítols. Així doncs, hem escollit 14 categories de la següent manera:

- Les 12 primeres categories estan formades per les paraules que es repeteixen de una fins a dotze vegades respectivament.
- La categoria 13 està formada per les paraules que es repeteixen entre 13 i 20 vegades.
- Finalment la categoria 14 la formen totes les paraules que apareixen 21 o més vegades.

Escollim les classes d'aquest tamany per obtenir un valor esperat major que 2 i d'aquesta manera complir amb els requisits teòrics del test Khi-quadrat definits en la secció 5.5.1 del capítol 5. Els valors calculats pel test, els compararem amb el valor fronterís de la zona d'acceptació, que en el nostre cas és 21,03, prenent un nivell de significació de 0,05 amb 12 graus de llibertat. Si el valor de discrepància Khi-quadrat és més petit que 21,03, no podem rebutjar que la distribució ajusti bé en aquell bloc.

Els valors observats i esperats els mostrem en l'annexe 9.3 d'aquest capítol, com a exemple ensenyem els del bloc 15.

Observant l'annexe 8.2 podem veure que no hem pogut obtenir un bon ajust de les nostres dades a la distribució de Sichel. En el mètode dels moments no obtenim ajustos satisfactoris. Sembla que el mètode que ajusta millor és el basat en la mitjana i la proporció de paraules que apareixen una vegada en el text. La distribució ajusta bé a sis dels 35 blocs, que corresponen al 6, 7, 15, 17, 19 i 32, i en percentatges representa el 17% de tots els blocs. En qualsevol cas, les discrepàncies entre els valors observats i esperats de la resta de blocs, estan relativament a prop del llindar de l'acceptació de la hipòtesis de bondat d'ajust.

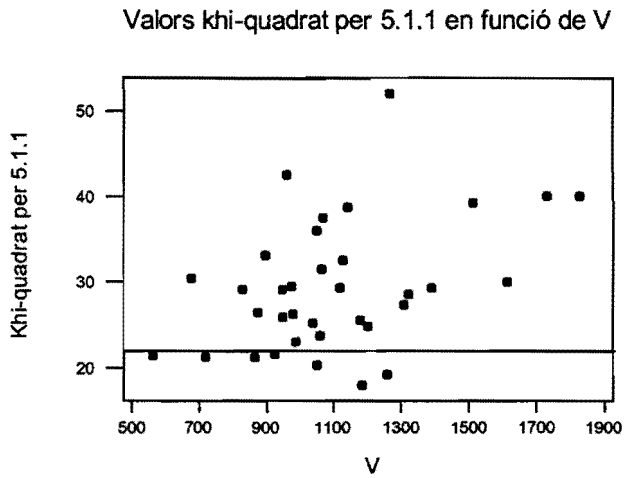
Les característiques dels blocs 6,7,15,19 i 32 pel què fa a la grandària de text en paraules i nombre de paraules que no es repeteixen, es presenten en la taula 8.2. Observant la taula, els valors de la Khi-quadrat més petits, corresponen als blocs més llargs, en quan a paraules, però no a vocabulari.

Blocs	*N*	*V*	*(V/N)*100*	χ^2
6	12761	1052	8,244	20.276
7	5036	718	14.257	21.150
15	3441	1143	33.217	21.321
17	15180	1259	8.294	19.171
19	12795	1184	9.254	17.856
32	7520	867	11.529	20.198

Taula 8.1: *es presenta la grandària de text N, el nombre de vocabulari V, el quocient entre aquests dos índexs i el valor de discrepància Khi-quadrat dels blocs ben ajustats per la distribució mixtura.*

A continuació explorem el comportament de les discrepàncies entre valors observats i esperats per la distribució mixtura, en funció del nombre de paraules diferents que apareixen en cada bloc. A valors petits de les discrepàncies li corresponen valors petits de vocabulari diferent, mentre que a valors grans de les discrepàncies es corresponen

valors grans de vocabulari diferent. Així doncs, hi ha correlació significativament positiva (0,412) entre el tamany de mostra V i la discrepància Khi-quadrat.



Gràfic 8.5: *Discrepàncies Khi-quadrat entre valors observats i valors esperats per la distribució mixtura, en funció del vocabulari que utilitza l'autor.*

Amb tot això, comprovem que malgrat dispondre de tamanys de mostra més gran, no millorem l'ajust de la distribució mixtura als textos del Tirant.

Recordant el capítol quatre, la distribució que proposàvem té tres paràmetres, α, θ i γ, tots tres provenen de la distribució barreja, que és la distribució de la probabilitat d'aparició de les paraules π: una distribució Gausiana inversa generalitzada. Fins a aquest punt de projecte hem estimat els paràmetres α i θ d'aquesta distribució, deixant el valor de gamma constant en -1/2.

Vistos els resultats dels ajustos de la distribució de Sichel als blocs del Tirant, se'ns planteja la possibilitat que no estiguem ajustant la distribució pel valor adequat del paràmetre gamma. En estudis precedents, s'havia fixat la gamma a -1/2, -3/2 i 1/2. En el punt següent es descriuen les noves equacions del sistema per estimar els paràmetres α i θ fixant la Gamma a –3/2 i 1/2, i es presenten els resultats.

8.5 Distribució de Sichel fixant la Gamma a –3/2 i 1/2

Estimem els valors per α i θ, fixant la γ a –3/2 i 1/2, a través d'una generalització de la distribució que depén de tres paràmetres. Això ho farem mitjançant el mètode 5.1.1. Recordem que aquest mètode consisteix en igualar la probabilitat que una paraula aparegui una sola vegada amb el valor observat en el text, i igualar el valor esperat amb el promig del nombre de vegades d'aparició de les paraules. Per fer-ho, s'implementa un sistema, que resulta de l'equació 4.10 reparametritzada a b i c, i de l'equació 4.12.

Les estimacions pels paràmetres α *i* θ i els resultats de la bondat d'ajust a la distribució basada amb aquestes estimacions es poden trobar en l'annexe 8.4 i 8.5, respectivament. Els valors de les discrepàncies χ^2 no són satisfactoris. L'alternativa a aquest fet és ajustar la generalització de la distribució depenent de tres paràmetres. En aquest cas no fixem la Gamma i estimem α, θ *i* γ.

8.6 Distribució de Sichel en funció de α, θ *i* γ

En aquest punt, considerem el paràmetre γ desconegut i el passem a estimar juntament amb α *i* θ igualant els valors teòrics de la distribució als valors observats del text. Per fer-ho hem implementat un sistema de tres equacions no lineals, que calculen la probabilitat que qualsevol paraula que coneix l'autor aparegui un cop, el valor esperat i el moment d'ordre 2. Els programes que resolen aquest sistema es troben en l'annexe 2.5 i els resultats que obtenim es mostren en la taula 8.5 de l'annexe d'aquest capítol. S'observa que els valors que estimen α són més grans que en els mètodes anteriors, els valors que estimen a θ es mantenen, i els valors de les noves gammes es distribueixen al voltant de –1.3. En qualsevol cas, els resultats de la bondat d'ajust continuen sense ser satisfactoris.

Així doncs, havent contemplat totes les possibilitats per estimar els paràmetres, s'obté que la distribució de Sichel més versemblant pel vocabulari del Tirant, depén dels paràmetres α *i* θ estimats a partir del mètode 5.1.1 i fixant el paràmetre Gamma a –1/2.

ANNEXE Capitol 8:

8.1: estimació d'α i θ basada en els dos mètodes pels 35 blocs

Bloc	*5.1.1*		*5.1.2*	
	α	*θ*	*α*	*θ*
1	0.0157	0.9855	0.0000	0.9922
2	0.0850	0.9907	0.0000	0.9949
3	0.0000	0.9876	0.0000	0.9927
4	0.0470	0.9810	0.0000	0.9892
5	0.1037	0.9925	0.0000	0.9958
6	0.0322	0.9928	0.0000	0.9956
7	0.0000	0.9817	0.0000	0.9892
8	0.1846	0.9948	0.2640	0.9974
9	0.0002	0.9926	0.0000	0.9962
10	0.0459	0.9943	0.0000	0.9970
11	0.0000	0.9931	0.0000	0.9966
12	0.0940	0.9953	0.2219	0.9975
13	0.0000	0.9863	0.0000	0.9923
14	0.0382	0.9887	0.0000	0.9941
15	0.0144	0.9725	0.0000	0.9834
16	0.0000	0.9895	0.0000	0.9942
17	0.0000	0.9932	0.0000	0.9959
18	0.0000	0.9839	0.0000	0.9901
19	0.0000	0.9916	0.0000	0.9949
20	0.0000	0.9853	0.0000	0.9922
21	0.0000	0.9914	0.0000	0.9953
22	0.0000	0.9888	0.0000	0.9938
23	0.0000	0.9916	0.0000	0.9955
24	0.0000	0.9881	0.0000	0.9937
25	0.0131	0.9888	0.0000	0.9940
26	0.0000	0.9855	0.0000	0.9931
27	0.0846	0.9829	0.0000	0.9920
28	0.0000	0.9802	0.0000	0.9903
29	0.0000	0.9805	0.0000	0.9908
30	0.0664	0.9914	0.0000	0.9955
31	0.0032	0.9868	0.0000	0.9928
32	0.0000	0.9869	0.0000	0.9925
33	0.0079	0.9895	0.0000	0.9943
34	0.0000	0.9833	0.0000	0.9923
35	0.0000	0.9893	0.0000	0.9948

8.2: càlcul de la Khi-quadrat i el p-valor en els dos mètodes pels 35 blocs

Bloc	*5.1.1*		*5.1.2*	
	χ2	*p-valor*	*χ2*	*p-valor*
1	29.0817	0.0064	54.3467	0.0000
2	25.1694	0.0219	48.8144	0.0000
3	21.5448	0.0628	38.9785	0.0002
4	30.4834	0.0040	49.1024	0.0000
5	29.2540	0.0060	53.4169	0.0000
6	20.2758	0.0885	33.8345	0.0013
7	21.1496	0.0700	37.0319	0.0004
8	39.3432	0.0002	77.5106	0.0000
9	29.2852	0.0060	55.4415	0.0000
10	30.0648	0.0046	57.1709	0.0000
11	40.0019	0.0001	73.0443	0.0000
12	40.1052	0.0001	88.2823	0.0000
13	23.0700	0.0408	44.5174	0.0000
14	38.8226	0.0002	69.9903	0.0000
15	21.3210	0.0668	36.2329	0.0005
16	25.5352	0.0196	46.6633	0.0000
17	19.1715	0.1179	34.0456	0.0012
18	29.1778	0.0062	44.7103	0.0000
19	17.8560	0.1631	34.7150	0.0009
20	23.6377	0.0346	51.9024	0.0000
21	27.3095	0.0113	50.9387	0.0000
22	24.8450	0.0242	48.5170	0.0000
23	28.6874	0.0072	53.5551	0.0000
24	32.5668	0.0020	58.0669	0.0000
25	31.5042	0.0028	55.6887	0.0000
26	52.0301	0.0000	91.8695	0.0000
27	42.5130	0.0001	78.3414	0.0000
28	29.5743	0.0054	64.4738	0.0000
29	33.2144	0.0016	66.7225	0.0000
30	25.9461	0.0173	47.3830	0.0000
31	26.3695	0.0152	47.2759	0.0000
32	20.1981	0.0691	40.5338	0.0001
33	26.2401	0.0158	48.1061	0.0000
34	37.4638	0.0004	78.0180	0.0000
35	36.1099	0.0006	63.6009	0.0000

8.3: càlcul dels valors observats i esperats en els tres mètodes pel bloc 15

Freqüència de les paraules	*Valors observats*	*Valors esperats segons 5.1.1*	*Valors esperats segons 5.1.2*
1	564	601.4305	585.8502
2	158	148.3310	144.0326
3	80	72.1311	70.8215
4	37	43.8430	43.5291
5	31	29.8464	29.9648
6	24	21.7694	22.1007
7	16	16.6342	17.0768
8	12	13.1437	13.6447
9	3	10.6519	11.1819
10	5	8.8052	9.3469
11	5	7.3954	7.9384
12	4	6.2930	6.8309
13-20	15	28.8142	32.6649
21	19	28.9110	43.0166

8.4: estimació d'α i θ fixant gamma a –3/2 i 1/2, basada en el mètode 5.1.1

Bloc	*Gamma=-3/2*		*Gamma=1/2*	
	α	*θ*	*α*	*θ*
1	2.693	0.99982	$0.0001*10^{-3}$	0.8518
2	3.1945	0.9998	$0.0008*10^{-3}$	0.8868
3	2.8218	0.99984	$0.4707*10^{-3}$	0.8625
4	2.4844	0.99977	$0.0338*10^{-3}$	0.8321
5	3.4343	0.99979	$0.0026*10^{-3}$	0.8997
6	3.4001	0.9998	$0.0009*10^{-3}$	0.8982
7	2.4703	0.9998	$0.0258*10^{-3}$	0.8311
8	3.9222	0.99981	$0.1776*10^{-3}$	0.9210
9	3.3343	0.99982	$0.0000*10^{-3}$	0.8951
10	3.6506	0.99982	$0.1375*10^{-3}$	0.9103
11	3.3914	0.99986	$0.1714*10^{-3}$	0.8986
12	3.9097	0.99986	$0.6699*10^{-3}$	0.9211
13	2.7284	0.99983	$0.0002*10^{-3}$	0.8551
14	2.9485	0.99982	$0.0019*10^{-3}$	0.8716
15	2.1487	0.99971	$0.0049*10^{-3}$	0.7933
16	2.9832	0.99983	$0.0019*10^{-3}$	0.8741
17	3.4278	0.99981	$0.0000*10^{-3}$	0.8998
18	2.5814	0.99984	$0.1146*10^{-3}$	0.8423
19	3.2037	0.99981	$0.0022*10^{-3}$	0.8875
20	2.6556	0.99986	$0.0004*10^{-3}$	0.8493
21	3.1789	0.99982	$0.0000*10^{-3}$	0.8862
22	2.9187	0.99983	$0.0023*10^{-3}$	0.8694
23	3.2004	0.99982	$0.0000*10^{-3}$	0.8875
24	2.8579	0.99984	$0.0011*10^{-3}$	0.8653
25	2.9388	0.99982	$0.0007*10^{-3}$	0.8706
26	2.6724	0.99986	$0.0035*10^{-3}$	0.8508
27	2.6059	0.99981	$0.0003*10^{-3}$	0.8440
28	2.3941	0.99984	$0.3153*10^{-3}$	0.8241
29	2.412	0.99983	$0.1771*10^{-3}$	0.8258
30	3.2411	0.99983	$0.0003*10^{-3}$	0.8899
31	2.7687	0.99982	$0.3095*10^{-3}$	0.8580
32	2.7765	0.9998	$0.2722*10^{-3}$	0.8583
33	2.9836	0.99984	$0.0002*10^{-3}$	0.8742
34	2.5408	0.99985	$0.0004*10^{-3}$	0.8390
35	2.9598	0.99982	$0.0003*10^{-3}$	0.8722

8.5: càlcul de la Khi-quadrat i el p-valor havent fixat gamma a –3/2 i 1/2, basat en el mètode 5.1.1

Bloc	*Gamma=-3/2*		*Gamma=1/2*	
	χ^2	*p-valor*	*χ^2*	*p-valor*
1	427.17	0.000000	825.41	0.000000
2	824.3	0.000000	1144.6	0.000000
3	527.95	0.000000	915.95	0.000000
4	219.52	0.000000	543.85	0.000000
5	1162.7	0.000000	1404.8	0.000000
6	1118.5	0.000000	1357.4	0.000000
7	299.86	0.000000	653.27	0.000000
8	2442.2	0.000000	2343.1	0.000000
9	1424	0.000000	1797.9	0.000000
10	2199.2	0.000000	2377.1	0.000000
11	1940.8	0.000000	2387.5	0.000000
12	3105.8	0.000000	2980	0.000000
13	527.94	0.000000	957.66	0.000000
14	700.99	0.000000	1150.1	0.000000
15	104.65	2.22e-016	374.41	0.000000
16	844.36	0.000000	1305.8	0.000000
17	1441.8	0.000000	1715.8	0.000000
18	405.85	0.000000	753.83	0.000000
19	1070.5	0.000000	1437.6	0.000000
20	523.79	0.000000	995.7	0.000000
21	1151	0.000000	1585.1	0.000000
22	809.92	0.000000	1299.5	0.000000
23	1206.2	0.000000	1633.2	0.000000
24	710.09	0.000000	1187.6	0.000000
25	686.23	0.000000	1102.5	0.000000
26	668.25	0.000000	1258.1	0.000000
27	353.9	0.000000	794.83	0.000000
28	409.78	0.000000	886.59	0.000000
29	371.56	0.000000	814.43	0.000000
30	808.84	0.000000	1098.2	0.000000
31	453.83	0.000000	809.92	0.000000
32	462.82	0.000000	818.49	0.000000
33	656.6	0.000000	1032	0.000000
34	480.43	0.000000	981.78	0.000000
35	783.85	0.000000	1198.7	0.000000

8.6: Estimació d'α, θ i γ de la generalització de la distribució Sichel a través del el mètode 5.1.1

Bloc	Generalització de la distribució Sichel		
	α	*θ*	γ
1	2.6819	0.99861	-1,2912
2	2.8048	0.999	-1,2312
3	2.7885	0.99866	-1,2971
4	2.6671	0.99784	-1,3270
5	2.8779	0.99916	-1,2105
6	2.8584	0.99907	-1,2070
7	2.6677	0.99783	-1,3294
8	2.941	0.99959	-1,1667
9	2.7318	0.99929	-1,1939
10	2.9056	0.9995	-1,1961
11	2.7778	0.99944	-1,2036
12	3.0038	0.99963	-1,1862
13	2.7177	0.99863	-1,2950
14	2.7087	0.99893	-1,2501
15	2.7996	0.99729	-1,4775
16	2.7526	0.99895	-1,2571
17	2.8991	0.99922	-1,2202
18	2.7695	0.99817	-1,3430
19	2.862	0.99907	-1,2491
20	2.6257	0.99857	-1,2791
21	2.7394	0.99911	-1,2185
22	2.7508	0.99892	-1,2710
23	2.7254	0.99917	-1,2121
24	2.6707	0.99883	-1,2542
25	2.7021	0.99887	-1,2485
26	2.5808	0.99881	-1,2657
27	2.6165	0.99859	-1,2900
28	2.5461	0.99829	-1,3097
29	2.4742	0.99827	-1,2763
30	2.6448	0.99899	-1,1706
31	2.6544	0.99858	-1,2619
32	2.7409	0.99861	-1,2921
33	2.6198	0.99879	-1,2059
34	2.463	0.99856	-1,2473
35	2.4646	0.99887	-1,1672

8.7: càlcul de la Khi-quadrat i el p-valor per la generalització de la distribució Sichel, basat en el mètode 5.1.1

Bloc	*Generalització de la distribució*(α,θ,γ)	
	χ^2	*p-valor*
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		
21		
22		
23		
24		
25		
26		
27		
28		
29		
30		
31		
32		
33		
34		
35		

CAPÍTOL 9:
Conclusions

Com ja s'ha anat explicant, al voltant de l'autoria del *Tirant lo Blanc* s'han fet diferents hipòtesis. Les principals són l'autoria única atribuïda a Joanot Martorell i la hipòtesi de la doble autoria entre Martorell i Martí Joan de Galba.

L'objectiu de l'estudi era el d'intetar esbrinar si existeix un canvi d'estil en el decurs de tot el llibre a través de l'estudi de la distribució del vocabulari al llarg del llibre. Les eines amb les quals hem treballat per intentar aclarir-ho ens han permès descobrir un canvi en la riquesa de vocabulari; sembla que a partir del capítol 380, el llenguatge utilitzat és molt més pobre que a la resta del llibre. Tot i que aquesta frontera es podria atribuir a altres factors, el fet que coïncideixi amb el que diuen alguns experts, fa que sigui la hipòtesi més creïble.

9.1 Resum dels resultats

Per intentar demostrar aquest canvi vam ajustar la distribució de Sichel. Aquesta distribució, que és una barreja de Poissons, varia en funció dels paràmetres alfa i teta, els quals depenen del nombre de paraules d'un text, *N*. La reparametrització d'aquesta distribució en funció dels paràmetres de la distribució de barreja, *b* i *c*, elimina aquesta dependència de *N* i per tant, va funcionar millor en el nostre cas, perquè els capítols tenen llargades molt diferents.

Analitzant l'evolució dels valors estimats del paràmetre *c*, hem assolit el nostre objectiu principal. El comportament de *c* indica, que hi ha dos canvis en la riquesa del vocabulari del *Tirant lo Blanc*.

El primer d'aquests canvis, el trobem a l'inici de l'obra, entre els 100 primers capítols. De fet, ja és sabut que el començament del *Tirant* és una traducció d'un llibre de cavalleries anglès i aquesta primera frontera possiblement detecta aques fet.

El segon punt de canvi, molt més rellevant, es troba al capítol 382, en el punt on alguns entesos en literatura medieval creuen que hi ha un canvi d'autor. Aquest canvi és més evident que el primer perquè el paràmetre *c*, a més de patir un canvi de tendència, també incrementa considerablement la variabilitat.

A l'hora d'interpretar el significat del paràmetre *c*, hem trobat que aquest està molt relacionat amb la riquesa del vocabulari. Això ens ha portat a analitzar el comportament de la riquesa del vocabulari del *Tirant* mesurat a través d'altres índexs coneguts, com el de Simpson i l'entropia. Hem vist que els resultats fent servir aquests indicadors coïncideixen amb els de *c*.

Com a resultat rellevant doncs, hem obtingut que *c* presenta una forta correlació positiva amb l'índex de Simpson, *D*, el qual calcula la probabilitat que en extreure dues paraules a l'atzar d'un text, aquestes siguin iguals. Per tant, valors alts d'aquest indicador ens

mostren un vocabulari poc ric, en canvi, valors baixos de D indiquen una major riquesa en el text.

9.2 Possibles extensions futures

Vistos els resultats, introduïm algunes propostes per intentar trobar aquesta frontera a partir d'altres mètodes. Les resumim en quatre punts:

1. Ajustar la distribució de Sichel per blocs de llargada idèntica i així, eliminar la dependència de N dels índexos de diversitat.

2. Ajustar la distribució de Waring Heradan al *Tirant* , la qual descriu breument en el capítol 4 del projecte. Es proposa intentar trobar la frontera a partir dels paràmetres d'aquesta distribució.

3. Veient que els valors estimats pel paràmetre b de la distribució barreja, Gaussiana inversa generalitzada, tendeixen a 0, proposem canviar-la per una Gamma.

4. Ajustar la mateixa distribució de Sichel des d'un punt de vista Bayesià. S'estimarien models geràrquics bayesians que tracten la distribució de barreja cam a distribució a priori.

Bibliografia

A. C. Atkinson and LAM YEH (1982). Inference for Sichel's Compound Poisson Distribution. *Journal of the American Statistical Association*

Andrew L. Rukhin and William E. Strawderman (1982). Estimating a Quantile of an Exponential Distribution. *Journal of the American Statistical Association*

Gillian Z. Stein, Walter Zucchini, and June M. Juritz (1987). Parameter Estimation for the Sichel Distribution and Its Multivariate Extension. *American Statistical Association*

Ginebra, J. and Cabos, S (1998). Anàlisi estadística en l'estil literari; aproximació a l'autoria del *Tirant lo Blanc, Afers, vol.* 29, *pp.* 185-206

Morris H. Degroot. Probabilidad y Estadística. *Segunda Edición*

Sichel, H.S. (1974). On a distribution representing sentence-length in written prose. *J.R. Satist. Soc.* A, **137**, 25-34

Sichel, H.S(1975). On a distribution law for words frequencies. J. Amer. Statist. Ass. , **70**, 542-547

Sichel, H.S. (1982). Asymptotic efficiencies of three methods of estimation for the inverse Gaussian-Poisson distribution. *Faculty of Industrial Engineering and Management, Technion, Haifa, Israel*

Sichel, H.S. (1986a). Word frequency distributions and type-token characteristics. Math Scientist, **11**, 45-72

Sichel, H.S. (1986b). Parameter Estimation for a Work Frequency Distribution based on Occupnacy Theory. *Communication in Statistics, Part A – Theory and Methods*, **15**, 935-949

Riba, A. (2003). Homogeneïtat d'estil en el *Tirant lo Blanc. Unpoblished phD thesis, UPC,* Barcelona.

Riba, A. (2004). Diversity of Vocabulary and Homogeneity of Style in *Tirant lo Blanc. JADT. Dep. Estadística, UPC.*

Riba, A. and Ginebra, J. (2003). "Change-point estimation in a multinomial sequence and homogeneity of literary style, *Doc. de recerca 2003-***, Dept. d'estadística i I.O., UPC, Barcelona.

Annexes:

Programes implementats

Annexe 1:

Programes de Matlab utilitzats pel tractament de dades

Adjuntem totes les funcions que hem necessitat pel tractament i manipulació dels fitxers de dades, així com programes per fer càlculs puntuals necessaris per l'estimació dels paràmetres.

```
function Z=ajuntar(X,Y)

%Els paràmetres d'entrada són els dos fitxers que volem agrupar en una matriu de dades Z.

[n,p] = size(X);
[n2,p2] = size(Y);

for i = 1:n
   for j = 1:p
      Z(i,j)=X(i,j);
   end
end

for i = 1:n
   for j=1:p2
      Z(i,p+j)=Y(i,j);
   end
 end
```

```
function [resul]=paraules(X)

%Aquesta funció calcula el vocabulari de l'autor utilitzat  V,  les paraules que surten un cop V1,
%i la llargada del text N.

[n,p] = size(X);
for i = 1:6
   for j = 2:p
      resul(j-1,i)=X(i,j);  %num. de paraules que surten una vegada a cada capitol
   end
end

%V=sum(X);          %num. de paraules diferents per cada capitol

V=0;
k=6;
for j = 2:p
   for i=1:n
      V=V+X(i,j);          %num. de paraules totals per cada capitol
   end
   resul(j-1,k+1)=V;
   V=0;
end;
N=0;
for j = 2:p
   for i=1:n
    N=N+(X(i,1)*X(i,j));
   end
  resul(j-1,k+2)=N;
  N=0;
end
```

```
function [pv1,q,m2,pv2]=termes(dades,V,N)

%funció que calcula els termes independents dels sistemes que estimen els paràmetres.

    %pv1= és la proporció de paraules que s'utilitzen una vegada i es calcula: V1/V
    %pv2= és la proporció de paraules que s'utilitzen dues vegades i es calcula: V2/V
    %q=és el promig de vegades que surt una mateixa paraula i es calcula: N/V
    %m2=calculem el moment d'ordre dos

    [n,p]=size(dades);

    %m2
    for j=1:p
        m2(j)=0;
      for i=1:n
        m2(j)=m2(j)+((i*i*dades(i,j))/V(j));
      end
    end

    %pv1 i pv2 i q
    for j=1:p
      pv1(j)=dades(1,j)/V(j);
      pv2(j)=dades(2,j)/V(j);
      q(j)=N(j)/V(j);
    end
```

```
function sumcap=sumaCap(dades)

    %Funció que agrega les freqüències de les paraules de capítols a blocs.

    [n,p]=size(dades);

    k=1;
    suma=0;
    for i=1:n
      for j=1:34
        while k<=14
          suma=suma+dades(i,14*(j-1)+k);
          k=k+1;
        end
        k=1;
        sumcap(i,j)=suma;
        suma=0;
      end
      while k<=13
        suma=suma+dades(i,14*(35-1)+k);
        k=k+1;
      end
      sumcap(i,35)=suma;
      suma=0;
      k=1;
    end
```

function Vblocs=calculvs(sumacaps)

```
%Aquesta funció calcula el nombre de paraules que apareixen r vegades per cada bloc, de r=1
%fins al número màxim de repeticions.

[n,p]=size(sumacaps);

for i=1:p
  for j=1:max(max(sumacaps))
   Vblocs(i,j)=0;
  end
end


for j=1:p
   for i=1:n
    if sumacaps(i,j)~=0
      h=sumacaps(i,j);
      Vblocs(j,h)=Vblocs(j,h)+1;
    end
   end
end

Vblocs=Vblocs';
```

function N=calculN(Vblocs)

```
%Funció que calcula la llargada de text de cadascun dels blocs.

[n,p]=size(Vblocs);

for j=1:p
N(j)=0;

 for i=1:n
    N(j)=N(j)+i*Vblocs(i,j);
 end

end
```

function alte=repalte(bc,N)

```
%funció que reparametritza b i c a alfa i teta, per blocs

for i=1:35;
   alte(i,1)=bc(i,1)*sqrt(1+bc(i,2)*N(i));
   alte(i,2)=bc(i,2)*N(i)/(1+bc(i,2)*N(i));
end
```

```
function bc=repabc(alte,N)

%funció que reparametritza alfa i teta a b i c, per blocs

for i=1:35;
   bc(i,1)=alte(i,1)*sqrt(1-alte(i,2));
   bc(i,2)=alte(i,2)/((1-alte(i,2))*N(i));
end
```

Annexe 2:

Programes de Matlab per estimar α i θ de la distribució Sichel

Adjuntem les funcions implementades en Matlab, que es necessiten per estimar els paràmetres de la distribució de Sichel. La funció de Matlab que resol sistemes d'equacions no lineals és "*lsqnonlin*". Aquesta funció crida el sistema que hem de solucionar indicant els punts inicials, el domini, i els termes independents de cada equació del sistema. Implementem un blucle per què estimi els paràmetres pels 470 capítols.

Annexe 2.1:

Programes en Matlab per estimar α i θ a través del mètode 5.1.1

Adjuntem les funcions que implementem per estimar els paràmetres de la distribució de Sichel segons el mètode basat en el valor esperat i la probabilitat que qualsevol paraula que coneix l'autor surti una vegada. En el cas de l'estimació dels paràmetres per blocs, utilitzem la generalització de la distribució de Sichel amb tres paràmetres, fixant gamma a –3/2 i a 1/2.

```
function resul=parametres(t,q)

%Programa que executa lsqnonlin per cada capítol (470). Aquesta funció crida el sistema
%d'equacions que ha de solucionar i indica els valors inicials, el domini i els termes
%independents t i q, calculats anteriorment.

    for i=1:470

    x=lsqnonlin(@sistema1,[1.33 0.77],[0 0],[Inf 1],[],t(i),q(i));
    resul(i,1)=x(1);
    resul(i,2)=x(2);

    end
```

```
function F=sistema1(x,t,q)

%sistema d'equacions que correspon al mètode 5.1.1 (cas particular de Gamma=-1/2)

F=[(0.5*x(1)*x(2))/(exp(x(1)*(1-sqrt(1-x(2))))-1)-t,
  (0.5*x(1)*x(2))/(sqrt(1-x(2))*(1-exp(-x(1)*(1-sqrt(1-x(2))))))-q];
```

```
function resul=pparametres(pv1,e,N)

%Programa principal generalitzat per les gammes;

for i=1:35
  Ns=N(i);
  save fitxerN Ns;
  x=lsqnonlin(@psichel,[1.33 0.77],[0 0],[Inf 1],[],pv1(i),e(i));
  resul(i,1)=x(1);
  resul(i,2)=x(2);
end
```

```
function F=psichel(x,t,q)
%sistema pel metode 5.1.1: t=V1/V i q=N/V
%generalitzat per les gammes;

load fitxerN;

%Gamma1:
%gamma=-0.5;

%Gamma2:
%gamma=-1.5;

%Gamma3:
gamma=0.5;

F=[((0.5*x(1)*x(2)*Ns*besselk(gamma+1,x(1)*sqrt(1+x(2)*Ns)))/(sqrt(1+x(2)*Ns)*((((1+x(2)*Ns)^
(gamma/2))*besselk(gamma,x(1)))-besselk(gamma,x(1)*sqrt(1+x(2)*Ns)))))-t,
((0.5*x(1)*x(2)*Ns*besselk(gamma+1,x(1)))/(besselk(gamma,x(1))-(((1+x(2)*Ns)^((-
1)*(gamma/2)))*besselk(gamma,x(1)*sqrt(1+x(2)*Ns)))))-q];
```

Annexe 2.2:

Programes en Matlab per estimar α i θ a través del mètode 5.1.2

Adjuntem les funcions que implementem per estimar els paràmetres de la distribució de Sichel segons el mètode dels moments, que es basa el valor esperat i el moment de segon ordre.

```
function resul=parametres(t,q)

%Programa que executa lsqnonlin per cada capítol (470). Aquesta funció crida el sistema
%d'equacions que ha de solucionar i indica els valors inicials, el domini i els termes
%independents e i m2, calculats anteriorment.

    for i=1:470

    x=lsqnonlin(@sistema2,[1.33 0.77],[0 0],[Inf 1],[],e(i),m2(i));
    resul(i,1)=x(1);
    resul(i,2)=x(2);

    end
```

```
function F=sistema2(x,e,m2)

%sistema d'equacions que correspon al mètode 5.1.2 (mètode dels moments)
% (cas particular de Gamma=-1/2)

F=[(0.5*x(1)*x(2))/(sqrt(1-x(2))*(1-exp(-x(1)*(1-sqrt(1-x(2))))))-e,
((0.5*x(1)*x(2)*((1-x(2)))^(-0.5))*(1+(0.5*(1+x(1)*sqrt(1-x(2)))*(x(2)*((1-x(2)))^(-1)))))/(1-exp(-
x(1)*sqrt(1-x(2))*(((1+(x(2)*(1-x(2))^(-1)))^(0.5))-1)))-m2];
```

Annexe 2.3:

Programes en Matlab per estimar α i θ a través del mètode 5.1.3

Adjuntem les funcions que implementem per estimar els paràmetres de la distribució de Sichel segons el mètode basat en la probabilitat que l'autor utilitzi les paraules del vocabulari una i dues vegades.

```
function resul=parametres(t,q)

%Programa que executa lsqnonlin per cada capítol (470). Aquesta funció crida el sistema
%d'equacions que ha de solucionar i indica els valors inicials, el domini i els termes
%independents pv1 i pv2, calculats anteriorment.

    for i=1:470

    x=lsqnonlin(@sistema3,[1.33 0.77],[0 0],[Inf 1],[],pv1(i),pv2(i));
    resul(i,1)=x(1);
    resul(i,2)=x(2);

    end
```

```
function F=sistema3(x,pv1,pv2)

%sistema d'equacions que correspon al mètode 5.1.3 (cas particular de Gamma=-1/2)
%sistema per pv1=V1/V i pv2=V2/V

F=[(0.5*x(1)*x(2))/(exp(x(1)*(1-sqrt(1-x(2))))-1)-pv1,
((((x(1)^2)*(x(2)^2))/(1-x(2)))/(8*(1+(x(2)/(1-x(2))))))*(1+((1/(x(1)*sqrt(1-x(2))))*(1/(sqrt(1+(x(2)/(1-
x(2))))))))*(1/(exp(x(1)*sqrt(1-x(2))*(sqrt(1+(x(2)/(1+x(2))))-1)-1)))-pv2];
```

Annexe 2.4:

Programes en Matlab per estimar α i θ a través del mètode 5.1.4

Adjuntem les funcions que implementem per estimar els paràmetres de la distribució de Sichel segons el mètode de màxima versemblança.

```
function sol=crida()

%Instruccions per cridar el mètode de la màxima versemblança 470 vegades

        clear sol;
        for i=1:470

           clear rmax,vsi;
           rmax=vultimr(i);
           vsi=vs(i,:);
           save versem1 rmax vsi;

           rVri=rVr(i);
           Vri=v(i);
           saveversem2 rVri Vri;
           t1=ti1;
           t2=ti2;
           x=tparamversem(t1,t2);
           sol(i,1)=x(1);
           sol(i,2)=x(2);
        end
```

```
        function x=tparamversem(ti1,ti2)

%Programa que executa lsqnonlin una sola vegada. Aquesta funció crida el sistema
%d'equacions que ha de solucionar per màxima versemblança indica els valors inicials, el
domini i els termes %independents ti1 i ti2, calculats anteriorment.

        x=lsqnonlin(@tsistema3,[1.33 0.77],[0 0],[Inf 1],[],ti1,ti2);

        resul(1,1)=x(1);
        resul(1,2)=x(2);
```

```
function F=tsistema3(x,ti1,ti2)

%sistema pel màtode de màxima versemblança
%ti1 i ti2 conte un vector de termes independents

load versem2;

F=[(1/x(1))*Vri+Vri-(((exp(x(1)*(1-sqrt(1-x(2)))))*(1-sqrt(1-x(2))))/((exp(x(1)*(1-sqrt(1-x(2)))))-
1))*Vri-rVrbessel(x(1))-ti1,
  (1/x(2))*rVri-(((x(1))*exp((x(1))*(1-sqrt(1-x(2)))))/((2*sqrt(1-x(2)))*(exp(x(1)*(1-sqrt(1-x(2)))))-
1)))*(Vri)-ti2];
```

```
function vbes=rVrbessel(alfa)

%calcula el sumatori de les Vr's per el quocient del bessel necessari per el sistema màxim
%versemblant
% capitol 1 r fins a 15 (vs0(1,r)), capitol 2 r fins a 28 (vs0(2,r)),...

load versem1;   %vconte rmax i vsi que es la fila corresponent a vs.
vbes=0;
    for r=1:rmax
        vbes=vbes+((vsi(r))*(besselk(r-3/2,alfa))/(besselk(r-1/2,alfa)));
    end
```

```
function rVr=vrs(vs)

%calcula el sumatori de les r*Vr

sumrVr=0;

for i=1:470
  for r=1:354
    sumrVr=sumrVr+(r*vs(i,r));
  end
  rVr(i)=sumrVr;
  sumrVr=0;
end
```

```
function vultimr=ultimr(vs0)

%construim un vector que ens indicara el maxim(rmax) de repeticions per capitol

for i=1:470
  for j=1:355
    if vs0(i,j)>0
      r=j;
    end
  end
  vultimr(i)=r-1
end
```

Annexe 2.5:

Programes en Matlab per estimar α, θ i γ

Adjuntem les funcions que implementem per estimar els paràmetres de la generalització de la distribució de Sichel amb tres paràmetres. Això ho fem per blocs de capítols.

```
function resul=p3parametres(pv1,e,m2,N)

% funció que estima els paràmetres de la distribució Sichel generalitzada en funció de α, θ i γ
%per blocs.
%Executa lsqnonlin per cada bloc (35). Aquesta funció crida el sistema d'equacions que ha de
%solucionar i indica els valors inicials, el domini i els termes independents pv1, e i m2, calculats
%anteriorment.


for i=1:35
   Ns=N(i);
   save fitxerN Ns;
   x=lsqnonlin(@p3sichel,[1.33 0.77 1],[0 0 -Inf],[Inf 1 Inf],[],pv1(i),e(i),m2(i));
   resul(i,1)=x(1);
   resul(i,2)=x(2);
   resul(i,3)=x(3);
end
```

```
function F=p3sichel(x,t,q,m2)

% distribució generalitzada per les gammes;
%sistema per estimar tres parametres:alfa, teta i gamma.
%t=V1/V, q=N/V=valor esperat, m2=moment d'ordre dos.

load fitxerN;
F=[((0.5*x(1)*x(2)*Ns*besselk(x(3)+1,x(1)*sqrt(1+x(2)*Ns)))/((sqrt(1+x(2)*Ns))*((((1+x(1)*Ns)^(x(
3)/2))*besselk(x(3),x(1)))-(besselk(x(3),x(1)*sqrt(1+x(2)*Ns))))))-t,
((0.5*x(1)*x(2)*Ns*besselk(x(3)+1,x(1)))/(besselk(x(3),x(1))-
((1+x(2)*Ns)^(x(3)/2))*besselk(x(3),x(2)*sqrt(1+x(2)*Ns))))-q,
((0.5*x(1)*x(2)*Ns*((1+(x(3)+1)*x(2)*Ns)*besselk(x(3)+1,x(1))+0.5*x(1)*x(2)*Ns*besselk(x(3),x(1)
)))/(besselk(x(3),x(1))-(((1+x(2)*Ns)^(((-1)*x(3))/2))*besselk(x(3),x(1)*(sqrt(1+x(2)*Ns))))))-m2];
```

Annexe 3:

Programes de Matlab per calcular la bondat d'ajust

Adjuntem els programes en Matlab, que ens permeten calcular la bondat d'ajust de la distribució de Sichel, tant per capítols, com per blocs. Per a aquest càlcul, s'han de reagrupar els valors observats, s'han de trobar els valors esperats per la distribució, i les discrepàncies entre ambdós calculant la Khi-quadrat.

Annexe 3.1:

Programes en Matlab per calcular la bondat d'ajust als capítols del *Tirant*

Adjuntem les funcions emprades pels càlculs necessaris per l'ajust de la distribució de Sichel als capítols del *Tirant lo Blanc*. Per fer-ho, cal executar un programa que reagrupa les observacions en set categories diferents.