Treball de Fi de Grau

Grau en Enginyeria en Tecnologies Industrials

Anàlisi Estadística de l'estil de J.K.Rowling

MEMÒRIA

Autor: Aleix Ortensi
Director: Josep Ginebra

Convocatòria: Juny 2017

Escola Tècnica Superior
d'Enginyeria Industrial de Barcelona

Resum

Aquest treball estudia l'estil literari de l'escriptora J.K. Rowling fent servir l'eina de l'estilometria, una ciència estadística d'anàlisi de textos. Utilitzant d'anàlisis gràfics, models lineals, anàlisis de correspondències i anàlisis discriminants, s'intenta respondre la principal pregunta del treball Hi han diferències d'estil entre els llibres de la seva famosa saga *Harry Potter* i les quatre novel·les posteriors de l'escriptora?

Com a part del treball s'ha escrit un programa en *python* per a poder tractar els textos i generar les dades desitjades sense dependre de les limitacions d'altres programes.

Per a caracteritzar l'estil literari s'usen variables tant dispars com la diversitat i la riquesa, la llargada de paraules i frases, i la freqüència d'ús de les paraules. El principal objectiu del treball era aprofitar l'estil de Rowling per aprendre a fer servir totes aquestes eines estadístiques.

Després de múltiples anàlisis, l'estudi troba diferències clares entre les novel·les de *Harry Potter* i les que no formen part de la saga, i posa de manifest les diferències existents entre riquesa i diversitat literària.

Sumari

1. INTRODUCCIÓ

1.1. L'estilometria

L'estilometria és una aplicació de l'estudi de l'estil lingüístic que utilitza anàlisis estadístics per a trobar diferències entre estils. S'usa principalment en el llenguatge escrit, però també s'ha aplicat en la música i en d'altres arts. Sol estar relacionada amb l'atribució d'autoria d'un document anònim. Té aplicacions legals, acadèmiques i literàries.

Quan vaig sentir parlar de l'existència de l'estilometria em vaig interessar a l'instant. Des de ben petit m'apassionen les històries i l'art d'explicar-les. I, com no, la literatura és un dels millors mitjans per a realitzar-ho. I si a més sumem el fet que estudio enginyeria, descobrir una ciència capaç d'analitzar l'estil lingüístic de texts d'una manera matemàtica i objectiva era mel pels meus llavis.

Notícies sobre novel·les anònimes que, un cop les havia visitat la ciència, es descobria el seu autor, no podien més que augmentar les meves ganes d'adquirir aquesta *arma*. Més endavant vaig descobrir el petit programa *Signature*, una eina que a partir de diversos texts d'autoria coneguda, si se li introdueix un de nou, és capaç de saber quin d'aquells autors l'ha escrit. La gran pregunta és "Com ho fa? Quines dades utilitza?". La resposta no pot ser més senzilla: utilitza dades com, per exemple, la freqüència d'ús de les paraules, o la simple longitud de les frases. Amb variables tant bàsiques com aquestes es poden obtenir grans conclusions. El mateix programa venia amb fragments dels *Federalist Papers* (col·lecció d'articles i assaigs per a ratificar la constitució americana el 1788. Els seus autors eren desconeguts, però existien sospites de qui eren). Amb dos simples clics, el programa era capaç de dir qui era l'autor del text. Fascinant.

L'estilometria es pot dir que va començar a aparèixer en el 1439, però no va ser fins el 1890 quan Wincenty Lutoslawski va escriure *Principes de stylométrie,* i definí els mètodes i les pautes a seguir. Els mètodes utilitzats han anat canviant, tornant-se cada vegada més sofisticats. Era molt difícil per mi saber per on tirar. Fins a quin nivell de complexitat estadística podia arribar? Vaig tenir la immensa sort que a la universitat hi hagués un expert en la matèria, el catedràtic Josep Ginebra, i que, tot i l'inacabable feina i la seva atapeïda agenda, va decidir ser el meu tutor del treball.

Llavors la meva ment va començar a volar. Se'm va ocórrer realitzar un model que detectés si una novel·la tindria èxit o no, trobar relacions entre gèneres literaris, o fins i tot gosar analitzar grans obres com *Los detectives salvajes.* Per sort, el meu tutor em va aturar. Eren preguntes massa àmplies, de difícil resposta. Apart, només havent cursat un any d'estadística en la carrera, i tenint un temps tant limitat per a fer el treball, havia de trobar un tema molt més concret.

Necessitava la obra d'un autor que fos força variada, amb facilitat per a trobar estils diferents. Una obra que em permetés jugar amb l'estilometria i aprendre nous conceptes estadístics. Em feia falta un camp de joc.

Dies després ho vaig veure clar. Analitzaria una, diem, marca registrada, que ha recaptat, i seguirà fent-ho, milers de milions d'euros. Una marca que, a partir de la narració i les idees intel·ligents de les que disposa, i el moment exacte en que va aparèixer, ha marcat per sempre més l'imaginari de la majoria dels joves (i no tant joves) del planeta. Estem parlant de J.K. Rowling, l'escriptora de la famosa saga *Harry Potter*.

1.2. L'obra de J.K. Rowling

El treball es basa en l'obra de Joanne Rowling. Nascuda al Regne Unit en el 1965, va estudiar filologia francesa i clàssica a la universitat d'Exeter. En un viatge en tren el 1990, se l'hi va acudir una idea que revolucionaria el món de la literatura juvenil. Una escola de màgia. El 1995 va acabar el manuscrit del primer llibre d'una suposada saga nombrada *Harry Potter*. Va ser rebutjat per 12 editorials. Els agents l'advertien que es busqués una altra feina, escriptora de novel·les infantils no li donaria per a menjar. Finalment, va aconseguir publicar la novel·la el 1997. Com els editors creien que el nom d'una dona no vendria suficients exemplars, va firmar com a J.K. Rowling, el nom amb el que se la coneixerà per sempre més.

El juliol del 2000, cinc anys després, la quarta novel·la de la saga batia rècords de venta als Estats Units i a Anglaterra. Poc més tard, Joanne Rowling es convertia en multimilionària, propietària d'una marca registrada anomenada *Harry Potter* valorada actualment en 7000 milions de lliures. La saga es va acabar amb el setè llibre el 2007. Posteriorment l'escriptora va escriure una novel·la adulta, *The Casual Vacancy*, la seva vuitena novel·la. Les crítiques, però, van ser dispars. Més endavant va escriure diversos llibres de relats sobre el món de *Harry Potter*.

El 2013, *The Cuckoo's Calling*, una novel·la de detectius escrita pel debutant Robert Galbraith, rebia bones crítiques i es defensava en vendes. Poc més tard es revelava que el nom de l'autor era un pseudònim. Darrere de la novel·la s'amagava Joanne Rowling. El 2014 i 2015 sortien al mercat les continuacions de la saga de detectius: *The Silkworm* i *Career of evil*. Més endavant, l'autora va col·laborar en una obra de teatre seqüela de l'escola de màgia, i va començar a escriure els guions d'una altra saga cinematogràfica sobre el món de *Harry Potter*.

S'ha d'admetre que J.K. Rowling disposa d'una llarga trajectòria com a escriptora. Ha treballat amb gèneres literaris ben diferents i ha aconseguit més ventes que la majoria d'empreses del planeta. Per aquestes raons s'ha escollit per a realitzar el treball. S'ha decidit estudiar només les set novel·les de *Harry Potter*, les tres detectivesques i *The Casual Vacancy,* degut a que són les úniques novel·les que té, i en les quals el seu estil és més depurat i exemplar. L'idioma escollit és l'original, el britànic. En la Taula 1.1 es llisten els llibres a analitzar.

Llibres estudiats
***Harry Potter and the Philosopher's Stone* (26/6/1997)**
***Harry Potter and the Chamber of Secrets* (2/7/1998)**
***Harry Potter and the Prisoner of Azkaban* (8/7/1999)**
***Harry Potter and the Goblet of Fire* (8/7/2000)**
***Harry Potter and the Order of Phoenix* (21/6/2003)**
***Harry Potter and the Half-Blood Prince* (16/7/2005)**
***Harry Potter and the Deathly Hallows* (21/7/2007)**
***The Casual Vacancy* (27/9/2012)**
***The Cuckoo's Calling* (18/4/2013)**
***The Silkworm* (19/6/2014)**
***Career of Evil* (20/8/2015)**

Taula 1.1 Novel·les a estudi

1.3. Descripció del treball

El problema de l'estadística, i en particular la estilometria, és que segons com siguin les dades, l'estudi pot canviar completament. Possiblement no es troba cap relació interessant, poder són caòtiques,... Per això, vaig decidir llençar-me a la piscina, començant fent proves (inspirades en estudis i eines vistos) i intentant veure per on tirar.

La gràcia de l'estudi, doncs, seria extreure respostes importants a partir de dades tant simples com l'ús de paraules o llargades de frases i paraules, projectades en gràfics i amb alguna eina més. També tenia l'oportunitat d'aprendre un tema molt interessant d'estadística que no entra en el grau: l'anàlisi multivariant, utilitzat força en l'àmbit laboral. Per a tornar més enginyeril el treball, vaig tenir la idea de programar un codi per a extreure la informació necessària a analitzar, i així posar en pràctica els coneixements informàtics apresos en els primers anys del grau, i adquirir-ne de nous.

A mesura que avançava amb l'estudi, jugant amb les dades vaig anar veient les direccions que podia prendre. La pregunta més general a la que aspirava respondre era: "Hi ha diferències significatives entre les novel·les de la saga de *Harry Potter* i i les que no?".

En el capítol 2 del treball es detallen les variables explicatives amb les que respondre aquesta pregunta, tant les que s'han acabat utilitzant com les que finalment no han tingut lloc en l'estudi definitiu.

En el capítol 3 s'exposen breument les eines estadístiques utilitzades durant el treball, i en el 4es detalla el procés per aconseguir les dades, i el codi programat per a extreure les variables explicatives del capítol 3.

Els capítols 5, 6 i 7 presenten les variables *longitud de les paraules, longitud de les frases* i *paraules més usades.* En el 8è capítol es presentarà la diversitat i riquesa de l'obra de Rowling, al 9 una aproximació a altres estudis possibles, al 10 es detallarà el pressupost i finalment es donarà pas a les conclusions del treball.

2. DESCRIPCIÓ DE LES VARIABLES

El més important un cop es té el material a analitzar, és saber quines variables interessen. Les eines més utilitzades en l'estilometria són: longitud de les paraules, longitud de les frases, paraules més utilitzades i la diversitat i riquesa d'un text. És important notar com els llibres s'han dividit en capítols independents. Per tant, com tots els capítols escrits per J.K. Rowling sumen 436, es tindran 436 observacions per a estudiar.

En un primer moment es va optar per les següents variables:

Eines Inicials
Longitud de les paraules
Longitud de les frases segons caràcters
Longitud de les frases segons paraules
Lletres
Les paraules més utilitzades
Variables de riquesa i diversitat

Taula 2.1 Eines plantejades a l'inici

A mesura que es va anar avançant en el treball, algunes d'elles es van suprimir. A continuació s'exposa cadascuna:

- **Longitud de les paraules:** mesura el nombre de lletres que conformen les paraules d'un text. Ho agrupa en paraules d'1 lletra, de 2, de 3,... Només interessa saber el nombre de paraules amb *i* lletres que apareixen en un text. Hi ha 10 categories diferents que mesuren les longituds de les paraules. La primera conté les paraules d'1 lletra, la segona la de dos lletres, i així fins arribar a la desena variable, que mesura les paraules de 10 o més lletres (a partir de 8 lletres el seu ús comença a disminuir).
  Aquestes 10 variables s'han utilitzat en l'estudi. Són discretes, però es poden

dividir pel nombre total de paraules que apareixen en un text, obtenint els perfils de probabilitats (continus).
Notar que els 10 perfils de probabilitats són redundants, ja que sumen 1. Amb només 9 perfils de probabilitat es poden fer segons quins estudis).

- **Longituds de les frases segons paraules:** no s'ha acabat utilitzant. La manera més normal de mesurar la longitud de les frases és segons el nombre de paraules que contenen. D'aquí va sorgir la pregunta: i per què no mesurar les frases segons els caràcters que contenen? Al no trobar-se cap estudi que ho fes, es va decidir provar les dues maneres. El problema era que s'havia d'escollir una de les dues. Es va decidir escollir la més nova, per esbrinar per a què no s'utilitzava en altres estudis. Tot i així es va comprovar més tard com mesurant per paraules, apareixien diferències majors en l'estudi que no mesurant caràcters.

- **Longituds de les frases segons els caràcters:** s'han acabat utilitzant en l'estudi final. Mesuren el nombre de lletres que apareixen en les frases i també les comes "," i el guionet llarg "–". L'autor del treball ha considerat que la coma representa una pausa en una frase, n'és un element important, i per tant mereixia ser comptada com a un caràcter. Havia de tenir un valor en la llargada. Pel que fa al guionet llarg, l'autora l'utilitza com a pausa enmig de la frase, representant que els personatges o el narrador s'entrebanquen o canvien de tema enmig d'una frase. És una unitat sonora com la coma, simplement a nivell de silenci (quan es llegeix la frase no s'emet cap so, però representa un temps). Pel que fa a la delimitació de frases, els elements ".", "?", "!", "..." i ";" marquen el fi d'una frase. S'ha considerat així degut a l'estil que utilitza J.K. Rowling. Si l'escriptor d'estudi fos un altre, probablement s'haurien canviat aquests límits.
  També s'han escollit 10 categories, que agrupen conjunts de frases. Per exemple, la primera variable podria incloure el nombre de frases de 1 a 3 caràcters, la segona de 4 a 5 caràcters, i així fins la desena variable, que inclou fins les frases de 200 caràcters. Les categories contenen nombres similars de frases. Les frases que en tenen més de 200 caràcters s'han considerat casos singulars que no entren dins l'estudi.
  Aquestes deu variables poden representar-se com a discretes o com a contínues (dividint cada variable pel nombre total de frases per a obtenir el segon cas).

- **Lletres:** no s'han utilitzat en l'estudi per falta de temps i per a la poca rellevància que es temia que aportessin. Tot i així s'ha dissenyat el programa que en recull la informació (explicat en el capítol 4).

- **Paraules més utilitzades:** s'ha acabat utilitzant en l'estudi. S'han escollit 10 categories. Cada una correspon a una de les 10 paraules més usades en total per J.K. Rowling. A cada text es comprova quantes vegades s'ha utilitzat cada una d'elles.

- **Diversitat:** els coeficients a estudiar són l'índex de Simpson (*D*) i el nombre total de paraules diferents que apareixen en un text (*V*). La *D* mesura més la diversitat, i la *V* la riquesa d'un text (la riquesa és un cas molt particular de diversitat). S'han utilitzat en l'estudi, així com també la V_1 i V_2, que mesuren les paraules que apareixen 1 o 2 vegades en el text. Aquestes últimes formen part de l'estudi de la diversitat.
  Estudis estilomètrics molt professionals, que combinen estadístics amb experts literaris, no escullen totes les paraules aparegudes en els textos, com s'ha fet en aquest treball, sinó que agrupen per famílies de paraules, o per exemple els verbs els ajunten sota l'infinitiu (un mateix verb en passat i en futur compta com el mateix),... En aquest treball, per manca de temps, de coneixement literari, i per estar més encarat a l'aprenentatge d'eines estadístiques, s'ha decidit comptar totes les paraules com a independents. Per exemple, una paraula en singular és diferent que en plural. Pel que fa al problema dels temps verbals, al treballar amb llengua anglesa, el problema és menor, ja que no tenen masses temps verbals i no varien segons la persona que els executa gairebé. En canvi, si fos, per exemple, el català, que disposa de múltiples temps verbals i variacions de cada verb segons la persona que executa o rep l'acció (1a singular, 2a singular, 3a singular, 1a plura, etc.), sí que podria representar un problema més important.
  En el seu apartat corresponent s'explicarà més sobre aquestes eines de mesura.

3. EINES ESTADÍSTIQUES

En aquest capítol s'exposaran en detall les eines estadístiques que s'han utilitzat en el transcurs del treball.

- **Gràfics:** tot i ser l'estudi més simple de tots, aporta molta informació. Degut a la quantitat d'observacions i a la dificultat de trobar distincions amb les eines estadístiques, té una funció primordial dins el projecte. Tant en les eines de mesura de longitud de paraules i de frases, com en d'altres, és molt útil graficar cada una de les variables al llarg dels capítols, distingint els llibres entre ells. Permet trobar diferències a simple vista que ajudaran a decidir com continuar amb l'estudi i a trobar petites relacions i detalls que es perdrien en estudis més concrets i numèrics. S'ha de mencionar que la distribució dels capítols no s'ha fet de forma temporal. *The Casual Vacancy* és el 8è llibre de J.K. Rowling, però en els gràfics s'ha situat al final, com si fos l'onzè i últim llibre escrit per l'autora. Aquest canvi no ha estat casual. L'obra es pot dividir en tres grups: els *Harry Potter*, els 3 thrillers de la saga de *Cormoran Strike,* i *The Casual Vacancy*. Per tant, un grup és juvenil fantàstic, un altre és novel·la negra detectivesca, i el darrer és una novel·la dramàtica social per adults. Després de fer els primers gràfics, l'autor del treball s'adonà que si el vuitè llibre es situava al final de tot (agrupant primer la saga de mags, després els thrillers i per últim el drama social), els gràfics seguien una mena d'evolució. Si es deixaven de forma temporal, *The Casual Vacancy* forçava un salt forca diferenciable al mig dels gràfics. Com durant l'estudi prima la diferenciació en grups i no l'evolució temporal, es va decidir fer el canvi i col·locar *The Casual Vacancy* a l'última posició.

- **Anàlisi de correspondències:** l'anàlisi de correspondències s'utilitza per a graficar una taula de contingència (amb **valors discrets**), quan les dades tenen 3 dimensions o més. En els casos que es tractaran, al haver-hi 10 categories, cada observació (capítol), es troba en un espai de dimensió 10. Això és un gran problema alhora de graficar, així que l'anàlisi de correspondències consisteix en trobar els dos eixos que més informació guarden quan se li projecten els punts multidimensionals. El que s'aconsegueix llavors és un gràfic on hi han els 436 capítols (de 10 dimensions cada un) projectats en dos dimensions, guardant la màxima informació possible de

l'espai de 10 dimensions. També en fa un amb les projeccions per les 10 categories. Si en els gràfics resultants els capítols d'una novel·la es troben majoritàriament més a prop d'una categoria *x* que d'una altra anomenada *y,* significa que en l'espai de 10 dimensions també estan més propers, i per tant que la novel·la que s'observa utilitza més la categoria *x* que la *y.* De fet, l'anàlisi dóna el tant per cent que expliquen els eixos principals de la informació total. Com més alt sigui aquest percentatge, més versemblant és la informació projectada, i més certes poden ser les interpretacions dels gràfics.
Hi ha dos tipus de gràfics de sortida, els simètrics i els asimètrics. Ambdós relacionen els dos tipus de gràfic: observacions (capítols) i les categories, projectant un sobre l'altre i viceversa. Els asimètrics ho fan mantenint les distàncies reals, i els simètrics no. El problema dels asimètrics, com ha ocorregut en els estudis d'aquest projecte, és que els capítols formen un conjunt de punts massa gran com per a trobar relacions amb el diminut espai que ocupen les categories. Per això s'han escollit els simètrics, que tot i no mantenir les distàncies, deixen entreveure les relacions entre observacions i categories.

- **Anàlisi clúster:** és una eina de classificació no supervisada que agrupa objectes de manera que els objectes d'un mateix grup són més semblants entre ells que amb objectes d'altres grups. Aquesta eina exploratòria s'utilitza en múltiples camps, com en l'anàlisi d'imatges, reconeixement de patrons, autoaprenentatge de màquines o de compressió de dades. Només es pot utilitzar per a dades contínues. Hi ha dos grans tipus de clúster, els de les observacions i els de les variables.
  Durant l'estudi s'han utilitzat, però al no haver diferències prou fortes entres grups, aquest procés de classificació no supervisada no ha funcionat correctament.

- **Anàlisi discriminant:** és una eina de classificació supervisada que serveix per a predir si una variable categòrica dependent pertany a un grup o a un altre gràcies a una o diverses variables independents (variables predictives). Aquest anàlisi busca un hiperplà (lineal o quadràtic) que separi les observacions en agrupacions el millor possible, i així permeti predir si una nova observació pertany a un grup o a un altre. L'anàlisi discriminant lineal va ser inventat per Roland Fisher.

En treball s'ha utilitzat com una de les eines principals, ja que al estar supervisat dóna bons resultats, no com el clúster.
Només accepta variables contínues.

- **Models lineals:** els models lineals ajustats intenten semblar-se als teòrics, explicant la variable resposta per mitjà de diferents variables explicatives, cada una d'elles multiplicades per diferents coeficients. Serveixen per a saber quina és la millor manera d'explicar a través d'unes variables explicatives una altra, que representa la resposta. Els models lineals segueixen les hipòtesis de linealitat, variància constant, normalitat i independència. Si les dades no són lineals, es pot intentar el truc de linealitzar-les mitjançant l'aplicació de logaritmes. Funciona força sovint, tal com es veurà més endavant. En el capítol 7 s'utilitzaran per a realitzar comparacions de mitjanes, per a saber si hi ha diferències entre els nivells d'una variable binària.

- **Models logístics:** són models lineals que expliquen una resposta categòrica (pot ser binària o nominal). Són útils per a modelar la probabilitat d'ocórrer un esdeveniment en funció d'altres factors. Es poden utilitzar també com una espècie d'anàlisi discriminant. No s'han acabat utilitzant en l'estudi final degut als millors resultats de l'anàlisi discriminant.

4. EINES COMPUTACIONALS PER A L'OBTENSIÓ DE LES DADES

4.1. Procés d'obtenció de dades

Un cop s'han decidit les 11 novel·les a estudiar i les eines de mesura que les conformaran, el pas a seguir és el d'extreure la informació de les novel·les. Per a realitzar-ho, primer es van obtenir els llibres en format *epub.* Mitjançant el programa gratuït *Hammster,* es van convertir els llibres al format *txt.*

La idea era treballar capítol a capítol. O sigui, disposar de 436 observacions (les 11 novel·les sumen aquest nombre de capítols). Per a fer-ho, es van obrir els *txt* en format *word* i, amb l'ajuda de l'eina de cerca, es van anar extraient els texts de cada capítol i creant un document *txt* per cada un. Cal mencionar que s'han guardat de la següent manera: primer la lletra que identifica el tipus de saga a la que pertany (*H* per *Harry Potter, R* pels thrillers, *C* per *The Casual Vacancy*); després s'afegeix el número de novel·la dins la saga (el primer llibre de *Harry Potter* queda com a *H1*, el tercer thriller com a *R3*, i *The Casual Vacancy es queda igual,* no se li suma cap número perquè és únic); posteriorment s'afegeix la lletra *C* i després el número de capítol dins la novel·la. Per exemple, el 5è capítol del 2n llibre de *Harry Potter* s'anomena *H2C5,* el 18è capítol del 2n thriller és el *R2C18*, i el 3r capítol de *The Casual Vacancy* rep el nom de *CC3.*

Un cop els 436 capítols estan ben classificats i ordenats, se'ls ha d'extreure la informació (les eines de mesura). Per a realitzar-ho, es podrien emprar eines prefetes, com per exemple el programa lliure anomenat *Signature,* que si se li insereixen textos els analitza i dóna informació sobre les llargades de frases, l'ús de paraules, i fins i tot permet predit si els textos els ha escrit el mateix autor o no. Però com l'autor del projecte volia treballar amb les eines descrites anteriorment, amb els requisits imposats per ell mateix, i a més creia convenient aprendre més programació, degut a l'elevat ús en l'àmbit laboral i al baix nombre d'assignatures que ho ensenyen durant la carrera, es va decidir programar unes línies de codi en *python* per a extreure aquesta informació dels documents *txt*. Es va utilitzar el programa lliure *IDLE* per a escriure el codi .

Un cop obtingudes les dades requerides per a començar l'estudi, es van passar en diferents documents *Excel.* S'ha de mencionar que les dades van ser recollides sense

saber ben bé com anava l'estudi. Un cop obtingudes es va anar provant fins que es va definir la línia que seguiria l'estudi. Per això el codi de *python* i l'*Excel* no han pres les dades seguint les pautes del capítol 2, sinó intentant agafar-ho tot, i així tenir un munt de possibilitats per a fer estudis diferents. Un cop a l'*Excel*, per exemple, es va procedir amb la creació de les 10 categories per eina de mesura, cosa que si s'hagués sabut des de l'inici del treball, el codi del programa s'hauria ocupat d'obtenir-ho.

Posteriorment, les dades es van transportar al *Minitab*, el programari d'estadística utilitzat en l'escola. Amb ell s'han realitzat tots els anàlisis estadístics. És important notar que les dades només es van poder extreure amb les observacions (capítols) en columnes, però que en el *Minitab* es necessiten en files. Mitjançant l'eina *Transponer* de l'*Excel* es va poder efectuar el canvi.

Per últim, el *Word* s'ha utilitzat per escriure la memòria del treball. A la Taula 4.1 s'indica un resum del programari utilitzat.

Programari utilitzat	
Python	Extracció de dades dels textos
Minitab	Anàlisi estadístic
Microsoft Excel	Tractament de les dades inicials
Hammster	Conversió de format *epub* a *txt*
Bloc de notas	Per a guardar les dades i capítols
Word	Redacció memòria del treball

Taula 4.1 Resum del programari emprat

4.2. Descripció del codi

A continuació es detallarà el programet realitzat per a l'extracció de la informació dels textos de J.K. Rowling. Cal recordar que al no saber en un primer moment com aniria l'estudi, es va intentar extreure el màxim d'informació, inclosa les eines de mesura que posteriorment no es van utilitzar en l'estudi definitiu, com les lletres i el nombre de paraules per frase. Tot i així, per si algú vol continuar amb l'estudi o vol utilitzar algunes línies de codi, s'indiquen a continuació tots els detalls. **A l'annex del treball es detalla el codi sencer.** Correspon al programa per a extreure informació de *H1,* la primera novel·la de Rowling.

Es requereixen sis funcions del programa. Extreure el nombre de lletres per paraula, el nombre de caràcters per frase, les paraules per frase, les paraules més usades en els textos, les eines de mesura de diversitat i les lletres. Cada un d'aquests requeriments del programa s'ha treballat en una funció diferent de les altres dins el mateix programa, excepte diversitat i paraules més usades, que ha estat més senzill treballar-les conjuntament.

Un lector atent, al llegir el codi del programa, s'adonarà que a cada capítol es criden les cinc funcions, i que cada funció llegeix repetides vegades el text, crea llistes llargues i fa forces bucles. Si es treballés amb codis molt llargs seria una pèrdua de temps enorme, però al tractar-se de documents *txt* petits, és més fàcil i intuïtiu treballar d'aquesta manera.

- **Cos del programa**: El programa treballa per una sola novel·la a la vegada, llegint els seus capítols. Per tant, s'haurà d'executar 11 vegades. Es podria fer tot alhora afegint unes quantes línies més de codi, però interessava fer-ho a poc a poc, controlant que no apareguessin errors i revisant els resultats.
  Les primeres línies executables del programa són un bucle, on s'obre cada document *txt* (cada capítol) d'una novel·la. Per exemple, pel primer llibre de tots, els arxius estan guardats com a *H1C1, H1C2,* i així fins l'últim capítol *H1C17.* En el bucle, s'uneix l'*string H1C* amb un altre *string* numèric que varia a cada passada del bucle (cada capítol). Gràcies a la creació dels noms dels capítols, s'obra l'arxiu *txt* que el conté, es llegeix i es guarda en una variable *f.* Llavors es criden les 5 funcions, i un cop acabades es suma una xifra a l'*string* numèric que porta el compte dels capítols i es continua el bucle.

- **Funció 1: ParaulesFrase().** És la primera funció en cridar-se. Per cada capítol calcula el nombre de paraules usades a cada frase (compta des de frases d'una paraula fins a frases de 200 paraules). Les frases de més de 200 paraules s'han considerat casos puntuals no avaluables en el treball. Per això crearà un bloc de notes amb 204 files i una columna. Les quatre primeres files contenen el nom del capítol, el nombre de paraules totals, les frases totals, i la mitjana de paraules per frases. Les 200 files següents contenen el nombre de frases de cada tipus (d'una paraula, de dues, fins a 200).
  La variable d'entrada és un text *f* que conté el capítol sencer. Abans de començar a treballar amb *f,* es creen dues llistes. Una és *Numerossols,* conté els números de l'1 al 200. L'altre és *Números,* i és el mateix que l'anterior però amb subllistes [número, 0]. El número 0 representa el comptador que indicarà més endavant el nombre total de frases del tipus indicat. Tot això en 200 subllistes.
  Un cop obtingudes, es fa un *re.split* (a l'inici del programa s'ha importat la biblioteca *re*), i es divideix el text en frases. El principal problema és definir què és una frase i què no. En aquest estudi s'han considerat els següents elements:

Elements divisors de frases	
.	**;**
!	**?**
\n	**\t**
...	

Taula 4.2 Elements que indiquen canvi de frase

El punt, el signe d'exclamació i l'interrogant són evidents. Els ***\n*** i ***\t*** són el canvi de línia i la tabulació. S'han inserit per si de cas, ja que alguns diàlegs de les novel·les s'interrompen de cop, sense cap signe de puntuació.
Els casos més punxeguts són els punts suspensius i el punt i coma. L'autora, enlloc d'utilitzar els punts suspensius per a mostrar dubitació o interrupció enmig d'una frase, usa "**–**" (guionet llarg, el que en català s'usa per acotar diàlegs). Els punts suspensius els utilitza més en finals de frases conclusius.

El problema radica en que els arxius *txt* no ho entenen com a tres caràcters (tres punts), sinó com un sol caràcter. Per tant, és necessari incloure aquest caràcter especial en la comanda de delimitació de frases.
Pel que fa al punt i coma, l'autora el sol utilitzar per a separar frases diferents però de temàtica semblant, pel que s'ha considerat com a un element delimitador de frases.
Un cop dividit el text en frases, es copia la llista *Numeros* i se li diu *l,* i s'entra en un *for* (for *e* in *Text*). Per tant, *e* és la frase *i* del text. Dins del bucle, es fa un *split* a la frase, obtenint així les paraules que la formen, i es guarda el nombre de paraules amb la funció predeterminada *len.* El problema és que aquest últim *split* sol guardar espais en blanc i signes de puntuació com a paraules individuals. Per això es fa un bucle que mira paraula per paraula i resta 1 a la longitud total de la frase si es troba amb una d'aquestes falses paraules (solen ser espais en blanc " "). Un cop acabat això, es té la longitud (nombre de paraules) de la frase *e.* Llavors es busca si la longitud pertany a la llista *Numerossols* (números de l'1 al 200), i si és així es suma 1 a la posició que li pertany a la llista *l.* Per exemple, si s'acaba de trobar la tercera frase de 5 paraules, es va a la subllista de *l* [5,2], i se li suma 1 a la segona posició, que significa el nombre d'aquest tipus de frases trobades en el text *f*: [5,3].
Quan s'ha fet això per totes les frases del text, s'acaba el bucle principal (el que llegeix totes les frases) i es crea la variable *q* i *longitud.* En un bucle que recorre *l,* es va calculant el valor definitiu de *q* (nombre total de paraules per capítol) i *longitud* (nombre total de frases). Posteriorment es calcula la *mitjana.*
Llavors es procedeix a l'escriptura. Es crea el títol del document *txt* de Paraules per frase pel capítol X de la novel·la Y, i es comença a escriure el que s'ha comentat en el primer paràgraf, tenint en compte que s'han de convertir algunes variables en *string* si es volen escriure en el document.

- **Funció 2: CaractersFrase().** El codi és pràcticament idèntic al de la funció anterior. L'únic que varia (apart dels noms dels arxius), és que en el bucle que llegeix cada frase, enlloc de buscar frases, busca caràcters.
  Per aconseguir-ho, es crea, a més a més de les dues llistes numèriques de la funció anterior, una llista *ll* que conté tots els caràcters de l'alfabet anglosaxó i els símbols "—" i ",", ja que es consideren com a caràcters en l'estudi, tal com s'ha explicat en el capítol 2.
  Llavors s'entra en un *for,* que equival a treballar amb cada frase del text.

Aquesta frase es converteix tota a minúscula amb la funció *lower,* i entra en un altre bucle que analitzarà cada caràcter de la frase. Però com no tots els caràcters interessen, es comprova un a un si es troben a la llista *ll* (la que conté tots els caràcters que l'estudi busca. No es desitgen ni els guions, ni els apòstrofs, ni els punts ni altres elements que van units a les paraules i que la funció bàsica *split* no pot eliminar). Si el caràcter és una lletra de l'alfabet anglosaxó o bé "—" o ",", un comptador suma 1. Quan s'han llegit tots els caràcters de la frase, es té un comptador que indica el nombre de *bons caràcters* que conté aquella frase.
La resta del programa és idèntica a la de la funció *ParaulesFrase().*

- **Funció 3: CaractersParaules().** Es pretén crear una llista amb els diferents tipus de paraules del text, entenent tipus com el nombre de lletres que formen la paraula. El màxim estudiat són 40 lletres per paraula.
  La funció s'inicia igual que l'anterior, amb la creació de 3 llistes: una numèrica amb nombres de l'1 al 40, una altra numèrica amb subllistes, i una última amb les lletres de l'alfabet anglosaxó (només lletres, cap símbol més). Llavors es fa un *split()* normal, que divideix el text *f* en paraules.
  Tota la resta és idèntic a la funció *CaractersFrase()*. Es converteix la paraula en minúscula, s'entra en un bucle que llegeix caràcter a caràcter, s'obté la longitud real de la paraula,... I s'acaba creant un arxiu que ordena el nombre total de paraules del text *f* segons els caràcters que conten. El resultat és una sola llista numèrica.

- **Funció 4: Lletres().** Al no haver-se utilitzat les dades en el treball, es resumirà el programa breument. Es creen dues llistes inicials: una amb les lletres de l'alfabet, i una altra nombrada *l,* que fa el mateix però amb subllistes [lletra, 0]. Es separa el text per paraules i es llegeixen els caràcters de cada una. Per cada lletra trobada, es suma 1 al comptador (el segon element de la subllista de *l)* que li correspongui a la lletra. Per finalitzar, es crea un arxiu *txt* que conté el nom del capítol, el nombre total de lletres del capítol (calculat amb un bucle molt simple) i el nombre de vegades que apareix cada lletra en el text.

- **Funció 5: Paraules().** En aquesta funció es desitja extreure les eines de mesura de diversitat *N, V, V1, V2* i *D* (s'explicaran en el capítol corresponent), i també el nombre de vegades que apareixen les paraules més usades.

Inicialment es creen dues llistes: una nombrada *ll* conté l'alfabet més el símbol "—", i l'altre que conté subllistes on a cada una hi ha una paraula i un comptador que comença a 0 (cada paraula és una de les escollides per l'estudi, les més usades. Per a saber quines eren les més usades, es va utilitzar una funció molt semblant a aquesta, que retornava una llista amb totes les paraules utilitzades per l'escriptora, ordenades de major a menor). Es divideix el text *f* en paraules i es crea la llista *l* buida.

Es continua amb un bucle (ara es treballa paraula per paraula). Com cada paraula té probablement elements que no interessen, no es pot guardar directament. Per exemple, es tindria "*Hola."* , però només es vol *hola* (python no entén que és la mateixa paraula). Per tant, s'ha de "netejar": posar-ho tot en minúscula i eliminar elements que tenen incorporats. S'ha decidit que l'apòstrof s'eliminarà. *I'm* passarà a ser *im.*

Per a realitzar aquesta neteja, es crea un *string q* buit (*q=' '*). Llavors es llegeix caràcter per caràcter, i es copia només els caràcters que interessen (els de la llista *ll)* en *q.* Al finalitzar, *q* serà la paraula netejada. Llavors es comprova que *q* no estigui buida i es guarda en la llista *l* que contindrà al final del procés totes les paraules que apareixen en el text. Com per a escollir les paraules més usades es va utilitzar aquesta funció amb petites variacions (es va crear una llista amb totes les paraules que apareixien en el text i el nombre de vegades que ho feia cada una, i després es va copiar en un arxiu), aquí s'ha creat aquesta llista *l* que conté totes les paraules, però no es crearà cap arxiu amb totes elles. Si es volgués, amb unes poques línies de codi s'obtindria.

Per a guardar aquesta informació en *l,* es comprova si la paraula llegida en el bucle ja pertany a la llista *l.* Si és així, es va a la posició pertinent i es suma 1 al comptador de la subllista [paraula, nombre vegades que apareix]. Si és una paraula nova, es crea una subllista amb ella i un comptador a 0. Posteriorment, es comprova si aquesta paraula és de les més usades (si pertany a la llista creada anteriorment amb els paraules més usades). Si és una d'elles, es suma 1 al comptador d'aquesta subllista.

Un cop acabat de llegir el text en el bucle, es creen els comptadors de diversitat N, V1 i V2 (tots enters iguals a 0), i es fan un seguit d'operacions que permeten calcular-les: N és el nombre total de paraules del text (s'obté sumant el segon element de cada subllista de *l),* V1 és el nombre de paraules que només apareix un cop (només es sumen les subllistes amb el comptador a 1), i V2 és el nombre de paraules que apareixen dues vegades (subllistes de comptadors

a 2). Per últim es calcula D (es detallarà més endavant) i V, que és igual a la longitud de *l,* és a dir, al nombre de paraules diferents del text.

Per acabar, es crea un arxiu que conté el nom del capítol, el valor de les eines de mesura de diversitat, i el nombre de vegades que apareix cada una de els paraules més usades.

Finalment es tanca l'arxiu i s'acaba la funció.

5. LLARGADA PARAULES

5.1. Presentació de les dades

Una manera de definir un estil d'escriure és per mitjà la llargada de les paraules, entesa com el nombre de lletres que contenen. Per a fer-ho possible, s'ha creat la funció *CaractersParaules()*, exposada en el capítol 4, on per cada capítol crea un document *txt* d'una columna on s'enumeren les vegades que apareixen els diferents categories de paraules, sent una categoria de paraula o una altra segons el nombre de caràcters que la formen (el primer tipus són les paraules d'una lletra, el segon les paraules de dues lletres,...).

	1	2	3	4	5	6	7	8	9	10+	mitjana	ParaulTot
H1C1	148	756	1082	858	491	460	369	167	120	130	4.3434	4581
H1C2	104	498	815	645	515	340	271	131	61	47	4.3031	3428
H1C3	92	594	884	697	549	395	302	149	102	52	4.3532	3816
H1C4	171	561	879	738	511	356	217	145	52	48	4.1384	3678
H1C5	241	963	1409	1289	915	689	489	279	153	118	4.3632	6545
H1C6	205	961	1552	1250	879	578	392	225	108	111	4.2086	6261
H1C7	156	650	1060	828	595	394	317	187	131	117	4.3759	4435
H1C8	93	423	684	628	424	288	245	123	84	51	4.3855	3043
H1C9	147	720	1090	961	666	490	375	192	156	103	4.409	4900
H1C10	114	622	1022	873	551	370	333	204	104	78	4.3559	4271
H1C11	80	495	790	603	461	327	236	145	93	82	4.4133	3312
H1C12	138	861	1334	1108	711	494	356	225	145	103	4.2966	5476
H1C13	68	478	736	627	452	271	234	146	81	84	4.4249	3177
H1C14	110	541	819	684	419	346	255	161	73	49	4.2875	3457
H1C15	144	798	1122	1040	669	525	365	200	120	99	4.3369	5082
H1C16	185	990	1493	1358	898	567	397	282	144	106	4.2688	6420
H1C17	193	878	1222	1106	702	505	343	211	138	141	4.3019	5439
H2C1	70	363	604	485	355	274	200	118	53	43	4.384	2565
...	...	...	...	...	...	...	...	...	...	...	...	...
CC79	24	117	248	170	78	57	50	9	21	15	4.0279	789
CC80	178	1035	2082	1410	822	687	621	293	150	261	4.4092	7541

Taula 5.1. Resum taula de caràcters per paraula per cada capítol

Per tant, s'obté una taula (parcialment representada en la Taula 5.1) on s'enumera, per cada capítol de les onze novel·les, les paraules que té de X lletres. Les paraules de 10 o més lletres s'han agrupat en un mateix grup (no són gaire usuals). També s'ha calculat la mitjana per a cada un i el nombre total de paraules.

Un cop obtinguda la taula, se n'ha creat una altra amb els perfils fila (dividint cada casella pel nombre de paraules totals del capítol corresponent). A partir d'aquestes noves dades, s'ha prosseguit a la creació de diferents gràfics. En el gràfic presentat en la Figura 5.1 es mostren les mitjanes de lletres per paraula al llarg dels capítols, i també la mitjana de cada novel·la. Pel que fa al gràfic de mitjana de lletres per paraula (el gràfic de l'esquerra de la figura), s'observa com la major part dels capítols oscil·len entre les 4,3 i 4,6 lletres per paraula. Tot i això, hi ha capítols de *The Casual Vacancy* (*CV* a partir d'ara*)* que superen les 5 lletres de mitjana, i d'altres que freguen les 4 per paraula. Cal destacar la significativa disminució de la mitjana en *H1* (*Harry Potter 1)* respecte als altres. Pel que fa al gràfic de la dreta de la Figura 5.1, s'aprecia com la llargada de les paraules va augmentant al llarg dels *Harry Potter*, amb algun lleuger descens, i arriba al seu màxim amb la sisena novel·la (*H6*). A la següent novel·la, la darrera dels *Harry Potter*, l'escriptora arriba al mínim de lletres per paraula. Pel que fa als llibres que no pertanyen a la saga de mags (*noHP* a partir d'ara), mantenen una mitjana semblant entre 4,45 i 4,5 lletres per paraula.

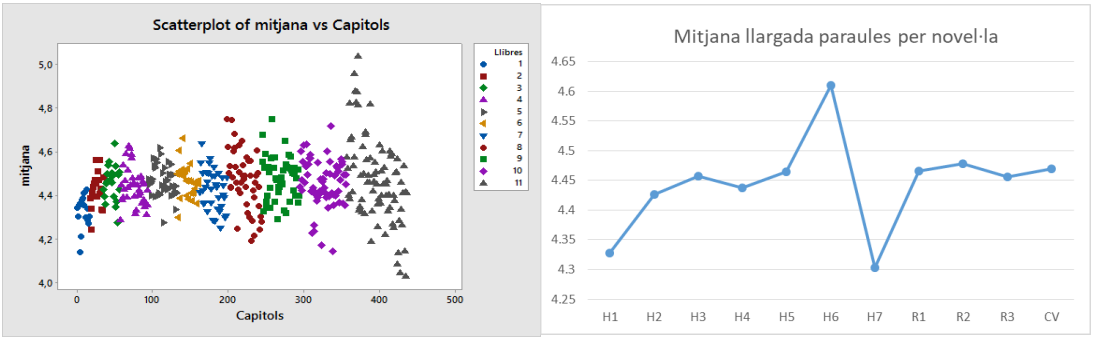

Figura 5.1.Evolució de la mitjana de caràcters per paraula al llarg dels 436 capítols

A la Figura 5.2 es presenten 10 gràfics. El primer representa els perfils de probabilitat dels capítols per paraules d'una lletra, el segon per les dos lletres, i així fins el desè, que engloba les paraules de 10 o més lletres.

En l'elaboració dels gràfics, els capítols es troben agrupats per llibre. Cal recordar que *The Casual Vacancy,* la vuitena novel·la de Rowling, s'ha situat al final, apareixent com l'onzè llibre.

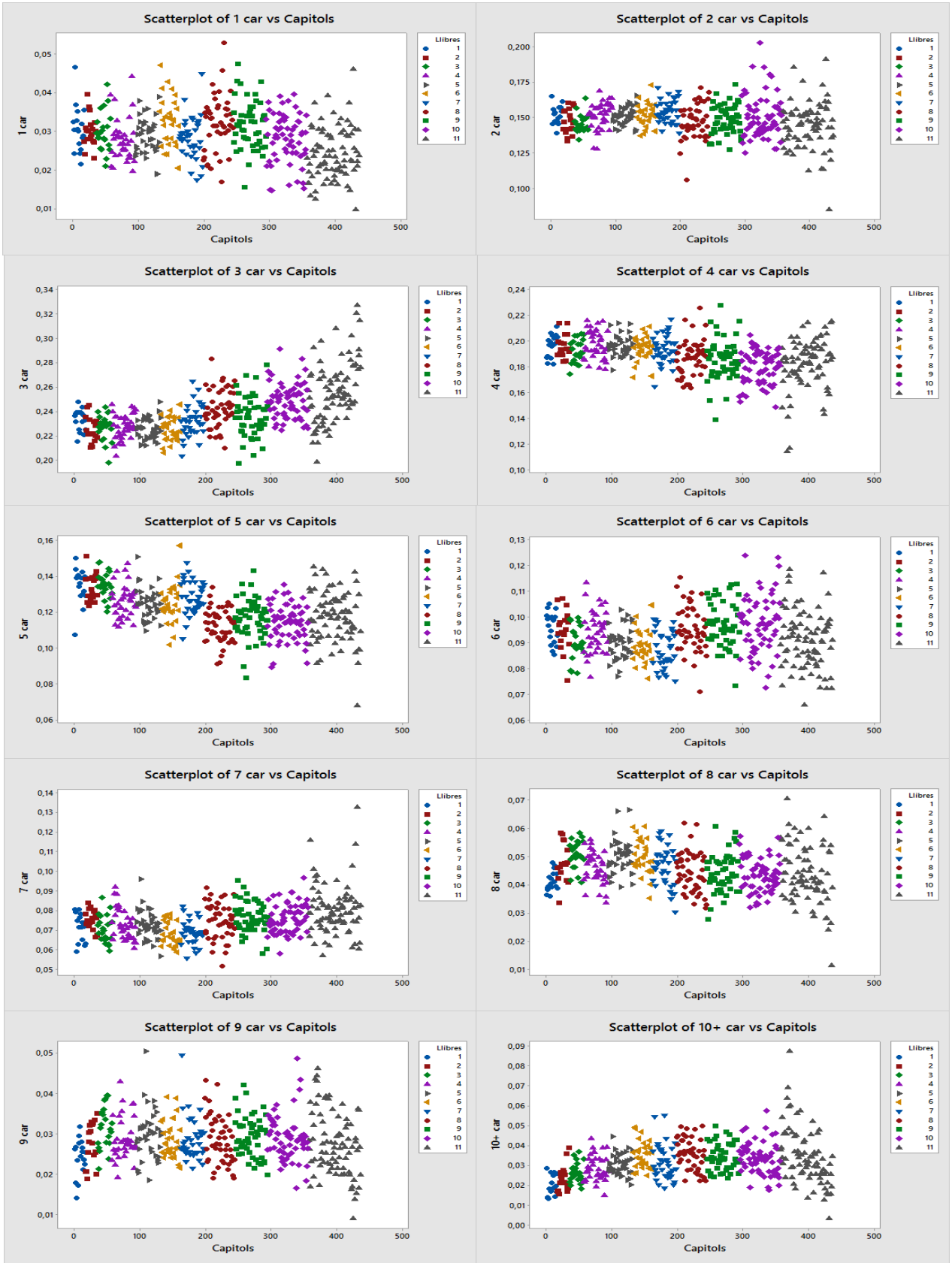

Figura 5.2. Probabilitats per tipus de paraules

Observant els gràfics (Figura 5.2), el que tenen en comú, és que les dades dels capítols de les novel·les que no són Harry Potter (*noHP*) tenen una variància més elevada que els Harry Potter (*HP*). D'entre ells, les dades dels capítols de *CV* tenen una tendència a ser les més disperses. Cal comentar que és el que disposa de més capítols (80), i que una bona part d'ells no arriben a les 1000 paraules.

Pel que fa a *noHP*, apart de tenir més variància, utilitza més que *HP* les paraules de 3, 6 i 7 lletres. En les paraules de 3 lletres la diferència és més clara. En les d'una lletra, *noHP* inicia una disminució en el seu ús de manera progressiva.

HP predomina en les paraules de 2, 4 i 5 lletres. És interessant fixar-se en les paraules de 5 lletres, on *HP* experimenta un creixement a mesura que avança la saga, però a partir dels *noHP* s'atura el creixement i l'ús es manté com a constant.

Per *H1*, a partir del gràfic de 8 lletres, es pot observar com el seu ús disminueix de manera dràstica. Té sentit, ja que en la Figura 5.1 es veu com la mitjana de la llargada de les paraules que utilitza és de les més baixes.

Les paraules de 1, 8, 9 i 10 o més caràcters no s'utilitzen gaire, tal com es pot comprovar en els valors de l'eix vertical del gràfic. En les de 8 lletres s'aprecia un decreixement en l'ús al llarg de les novel·les a partir de *H6.*

5.2. Diferències entre *HP* i *noHP*

Un cop realitzats els gràfics es procedeix a un anàlisi de correspondències a partir de la taula de contingència (Taula 5.1).

Analysis of Contingency Table

Axis	Inertia	Proportion	Cumulative	Histogram
1	0,0028	0,2763	0,2763	******************************
2	0,0018	0,1807	0,4570	******************
3	0,0015	0,1446	0,6016	***************
4	0,0010	0,0969	0,6985	**********
5	0,0008	0,0813	0,7798	********
6	0,0006	0,0639	0,8438	******
7	0,0006	0,0596	0,9033	******
8	0,0005	0,0534	0,9568	*****
9	0,0004	0,0432	1,0000	****
Total	0,0101			

Taula 5.2. Resultats Anàlisi de correspondències

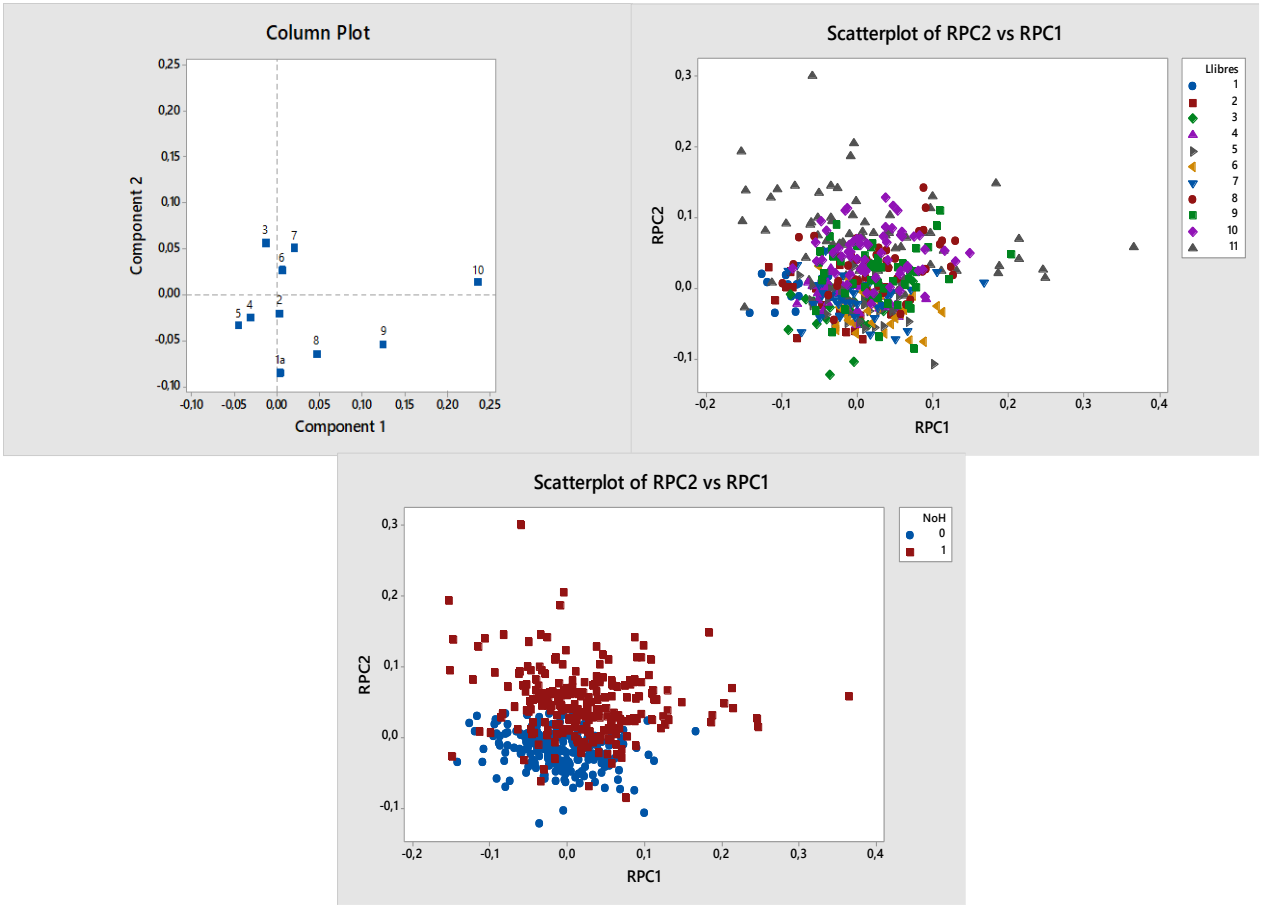

Figura 5.3.Gràfics Anàlisi de correspondències: per categories i per capítols, agrupats per llibres o *HP*

Entre les dues components principals expliquen un 45,7% de la informació (Taula 5.2). Això vol dir que gairebé la meitat de la informació de l'espai de 10 dimensions es manté al projectar-los en els dos eixos principals. La primera component explica un 27,63% de la informació, mentre que la segona un 18,07%.

En la tercera imatge de la Figura 5.3 es pot apreciar com els capítols de *HP* i els de *noHP* estan continguts dins dos núvols de punts separats, centrats en punts diferents. També es torna a notar com les dispersions dels *noHP* són més grans, en especial els capítols de *CV.*

Pel que fa als tipus de paraules, el gràfic de columnes separa tres tipus de categories: les paraules de 3, 6 i 7 lletres; les paraules de 2, 4 i 5 caràcters; i per últim les de 1, 8, 9 i 10. En concret, l'eix de la primera component (l'horitzontal), situa a l'esquerra les paraules de 3, 4 i 5 lletres, i la resta a la part dreta. Sembla ser que ordena,

aproximadament, les categories segons l'ús que se'n fa. A l'esquerra hi han les categories més utilitzades (paraules de 3, 4 i 5 lletres), i a la dreta els que tenen menor ús. Pel que fa a la segona component (eix vertical), separa les paraules de 3, 6, 7 i 10 lletres a dalt del gràfic, i la resta en la part inferior.

Tal com s'ha vist en l'anàlisi gràfic (apartat 5.1), els *noHP* utilitzen més les paraules de 3, 6 i 7 lletres, i els *HP* les de 2, 4 i 5. Això es pot comprovar en els gràfics del l'anàlisi de correspondències. Els capítols de *HP* estan centrats més avall que els *noHP*. També s'observa com les paraules de 2, 4 i 5 lletres cauen en la zona dels *HP,* mentre que per les de 3, 6 i 7 lletres, es troben més pròxims a la zona dels *noHP.* Pel que fa a les paraules de 1, 8, 9 i 10 o més lletres, com s'utilitzen menys que les altres, apareixen més separades.

Posteriorment s'ha intentat realitzar un anàlisi clúster, però tal com s'ha explicat en capítols anteriors, aquest no ha funcionat. Les diferències no són suficients per a un estudi no supervisat.

5.3. Es pot predir si és *HP* o *noHP*?

Pel que s'ha vist als apartats anteriors, s'aprecien diferències clares entre *HP* i *noHP*. Tot i això, falten resultats més conclusius. Sembla adequat doncs fer un anàlisi discriminant que intenti predir si un capítol és *HP* o *noHP,* i esperar que mostri resultats encara més concloents.

Per a fer l'estudi, s'ha de mencionar que, tot hi haver 10 variables explicatives, totes juntes sumen 1 per cada capítol, ja que estan en forma de perfil de probabilitat. Això implica que hi ha 10-1 dimensions (9 dimensions). En l'anàlisi no s'ha agafat la variable que inclou les paraules de 10 o més lletres.

L'anàlisi discriminant s'ha realitzat per mitjà d'una funció lineal, degut a que explicava millor que la quadràtica (en aquest cas), i s'ha emprat la tècnica de *cross-validation*. Aquest evita fer "trampes", ja que, per cada observació de les 436 existents, crea un nou pla lineal que no té en compte la observació. És a dir, intenta predir si el capítol que analitza en aquell moment pertany a *HP* o no mitjançant un pla que ha estat trobat a partir dels altres 435 capítols, sense usar les dades del capítol que s'està intentant predir. Això es fa per a que aquesta observació no influeixi en la predicció formant-hi part. El *cross-validation* és més important com menys observacions hi hagin. Tot i això, la diferència de previsió utilitzant-lo o no varia entre un 1% i un 2%.

```
Summary of Classification with Cross-validation

                 True Group
Put into Group       0      1
0                  178     36
1                   20    202
Total N            198    238
N correct          178    202
Proportion       0,899  0,849

N = 436            N Correct = 380           Proportion Correct = 0,872
```

Taula 5.3. Resultats Anàlisis discriminant lineal amb cross-validation

Els resultats són molt bons i conclusius. Més de l'esperat. En un 87,2% dels capítols el programa ha encertat si el capítol a classificar era *HP* o *noHP.* Concretament, un 89,9% d'encerts en capítols *HP*, i un 84,9% en *noHP.* A continuació la taula d'errors per llibre (percentatge de capítols erronis que presenta):

H1	5,88%	**H5**	10,52%	**R2**	22%
H2	11,11%	**H6**	10%	**R3**	3,22%
H3	0%	**H7**	22,22%	**CV**	18,75%
H4	8,1%	**R1**	15,21%		

Taula 5.4. Errors de predicció per novel·la

La novel·la amb pitjors classificacions és l'últim llibre de la saga de mags (*H7*). En alguns gràfics de l'apartat 5.1 ja es podia veure una tendència a semblar-se als *noHP.* Aquest fet pot implicar que l'anàlisi discriminant el confongui amb ells en alguns casos.

Pel que fa als *noHP,* el que millor es preveu és la tercera entrega del thriller (*R3)*. Amb la resta té més problemes. Destaca *H3* per a no tenir cap capítol mal classificat, i *H1* i *H4* per a tenir més d'un 90% de capítols ben predits.

Tot i això, sempre s'encerta més d'un 77% de les vegades els capítols d'una novel·la. Els màxim de capítols seguits mal classificats en l'estudi no supera els 3, i només ocorre una vegada.

6. LLARGADA DE FRASES PER CARÀCTERS

El mètode més usual per a mesurar la longitud d'una frase és comptant el nombre de paraules que conté. L'autor del treball, però, es va preguntar si es podia mesurar per nombre de caràcters. Durant el desenvolupament del treball es van estudiar els dos mètodes. Alhora de presentar-lo, s'havia d'escollir només un d'ells, per a que resultés repetitiu. L'elegit va ser la mesura de frases per caràcters, pel fet d'utilitzar-se menys en els altres estudis literaris.

També es vol recordar, tot i que s'ha detallat en el capítol 3, que els caràcters mesurats inclouen totes les lletres de l'abecedari, més la coma i el símbol "—", ja que es considera que en la lectura de la frase aquests dos símbols aporten una pausa.

6.1. Presentació de les dades

	H1C1	H1C2	H1C3	H1C4	H1C5	H1C6	H1C7	…
CaractersTotals	19475	14250	16388	14524	28363	25687	19022	…
mitjana	51.1155	58.1633	49.6606	40.6835	45.8207	43.4636	47.201	…
1	0	1	0	0	0	0	0	…
2	1	3	0	0	1	0	0	…
3	1	2	9	4	8	2	1	…
4	3	2	2	8	4	7	3	…
5	5	1	1	4	2	2	2	…
6	3	2	9	5	2	9	1	…
7	3	1	1	7	9	9	8	…
8	4	2	2	6	6	6	7	…
9	0	1	8	9	4	12	5	…
10	1	3	4	7	15	9	11	…
11	4	5	4	4	13	16	5	…
12	5	5	1	5	10	12	5	…
13	7	1	8	8	9	10	10	…
14	6	6	8	10	10	10	8	…
15	4	5	9	8	7	12	6	…
16	1	1	7	6	15	9	5	…
…	…	…	…	…	…	…	…	…
199	0	0	0	0	0	0	0	…
200	0	0	0	0	0	0	0	…

Taula 6.1 Caràcters per frase de cada capítol

Un cop recollides les dades en un document *txt,* de forma semblant al que s'ha fet en el capítol 5, s'han traslladat a un document *Excel.* La Taula 6.1 presenta una part d'aquestes dades. S'observa que la frase més curta mesurada és d'1 caràcter, i la més llarga de 200.

S'ha decidit prosseguir ajuntant frases en deu categories per a simplificar l'anàlisi. Les divisions en les 10 categories s'han fet procurant que tinguin una grandària semblant. Així, per exemple, es pot veure en la Taula 6.2 com el primer grup conté frases d'1 a 11 caràcters, el segon grup de 12 a 17, i així successivament fins al desè i últim grup, que engloba les frases de 110 fins a 200 caràcters. La Taula 6.2 mostra una part de la taula de dades obtinguda d'aquesta manera. A l'igual que en el capítol 5, alhora de realitzar els gràfics s'ha dividit cada valor pel nombre total de frases de cada capítol, convertint-los així en perfils de probabilitat.

	FraTot	mitjana	1-11	12-17	18-24	25-32	33-40	41-50	51-64	65-80	81-109	110-Fin
H1C1	381	51.1155	25	36	38	47	43	43	46	39	36	32
H1C2	245	58.1633	23	25	21	20	23	24	24	22	37	31
H1C3	330	49.6606	40	40	23	40	22	31	30	42	42	24
H1C4	357	40.6835	54	49	42	40	34	41	29	27	32	13
H1C5	619	45.8207	64	74	61	83	59	71	74	63	53	30
H1C6	591	43.4636	72	79	89	62	54	58	65	49	42	37
H1C7	403	47.201	43	52	52	48	31	30	45	38	39	30
H1C8	210	62.3048	12	15	10	22	19	21	27	27	28	30
H1C9	435	48.2322	50	51	40	55	40	38	51	48	48	27
H1C10	350	53.9314	22	32	35	35	36	38	51	39	29	37
H1C11	288	47.2153	31	36	34	27	31	29	35	30	22	20
H1C12	463	50.2354	46	47	57	41	51	44	56	40	46	42
H1C13	281	49.0534	37	30	33	20	29	26	28	32	33	19
H1C14	325	46.2862	36	30	44	30	31	38	41	30	32	17
H1C15	490	45.7327	37	53	60	64	56	57	53	43	53	20
H1C16	650	42.8892	70	78	81	78	61	80	71	63	47	26
H1C17	577	40.8354	77	72	85	68	55	63	66	42	36	24
H2C1	201	53.8259	20	14	17	23	18	20	24	22	21	22
H2C2	278	45.482	35	36	30	32	20	36	27	21	33	14
…	…	…	…	…	…	…	…	…	…	…	…	…
CC78	89	53.9551	9	10	11	8	8	4	9	13	11	9
CC79	63	51.5238	4	7	7	7	4	9	3	13	6	3
CC80	470	65.1617	33	44	34	45	31	46	44	52	56	90

Taula 6.2 Grups de frases

Es procedeix d'igual manera que en el capítol 5 en fer un gràfic amb les mitjanes de nombre de caràcters per capítol i per novel·la, i deu gràfics més amb la proporció de frases de cada categoria per capítol.

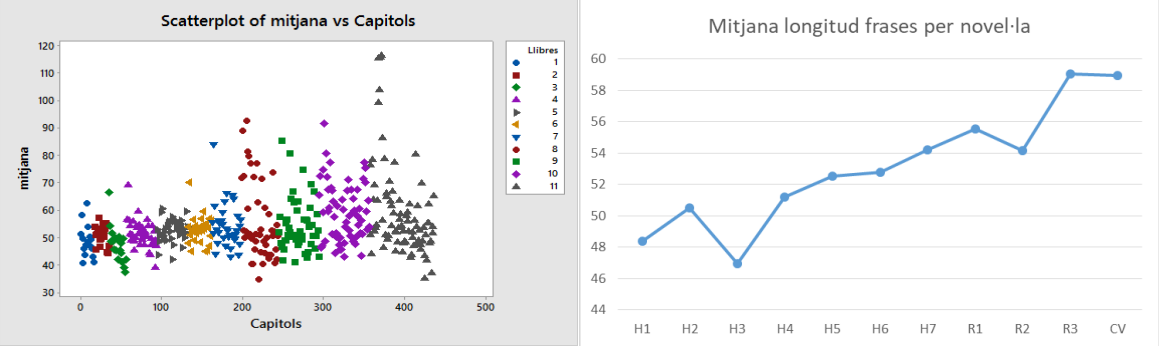

Figura 6.1 Gràfic de la mitjana de llargada frases per capítol i per novel·la

En ambdós gràfics de la Figura 6.1 s'aprecia com la longitud de les frases augmenta a cada llibre que escriu J.K. Rowling. Excepte per *H3* i *R2*, que marquen dos descens, la resta segueix una evolució creixent. En el gràfic de mitjanes per novel·la (el de la dreta), mostra que des de *H4* fins *R1* es pot dibuixar una línia gairebé recta. Pel que fa a mitjanes de les novel·les, les frases oscil·len entre els 47 caràcters i els 59. En les mitjanes per capítols, encara que hi ha un capítol amb 35 caràcters per frase de mitjana, i uns altres que s'apropen als 120 caràcters, la majoria dels capítols conté entre 40 i 65 caràcters per frase.

En els gràfics de la Figura 6.2, és més difícil intuir diferències entre *HP* i *noHP.* Les diferències són petites i costa de distingir degut a l'elevada variabilitat de *noHP*, que dificulta la comparació amb els capítols compactes de *HP.* Es pot intuir com *HP* utilitza una mica més les frases d'1 a 11 caràcters (primera categoria). En la cinquena categoria de frases, *HP* presenta un lleuger descens progressiu, mentre que els *noHP* es manté constant irregularment (un llibre puja, l'altre baixa, però no hi ha cap tendència de canvi general). En la tercera i setena categoria ocorre més del mateix. Pel que fa a la categoria amb les frases més llargues, augmenta l'ús a partir de *H4*, i s'experimenta una altra remuntada en *R3* (novel·la que destaca per a tenir la mitjana de frases més llargues per nombre de caràcters), però decau en *CV*.

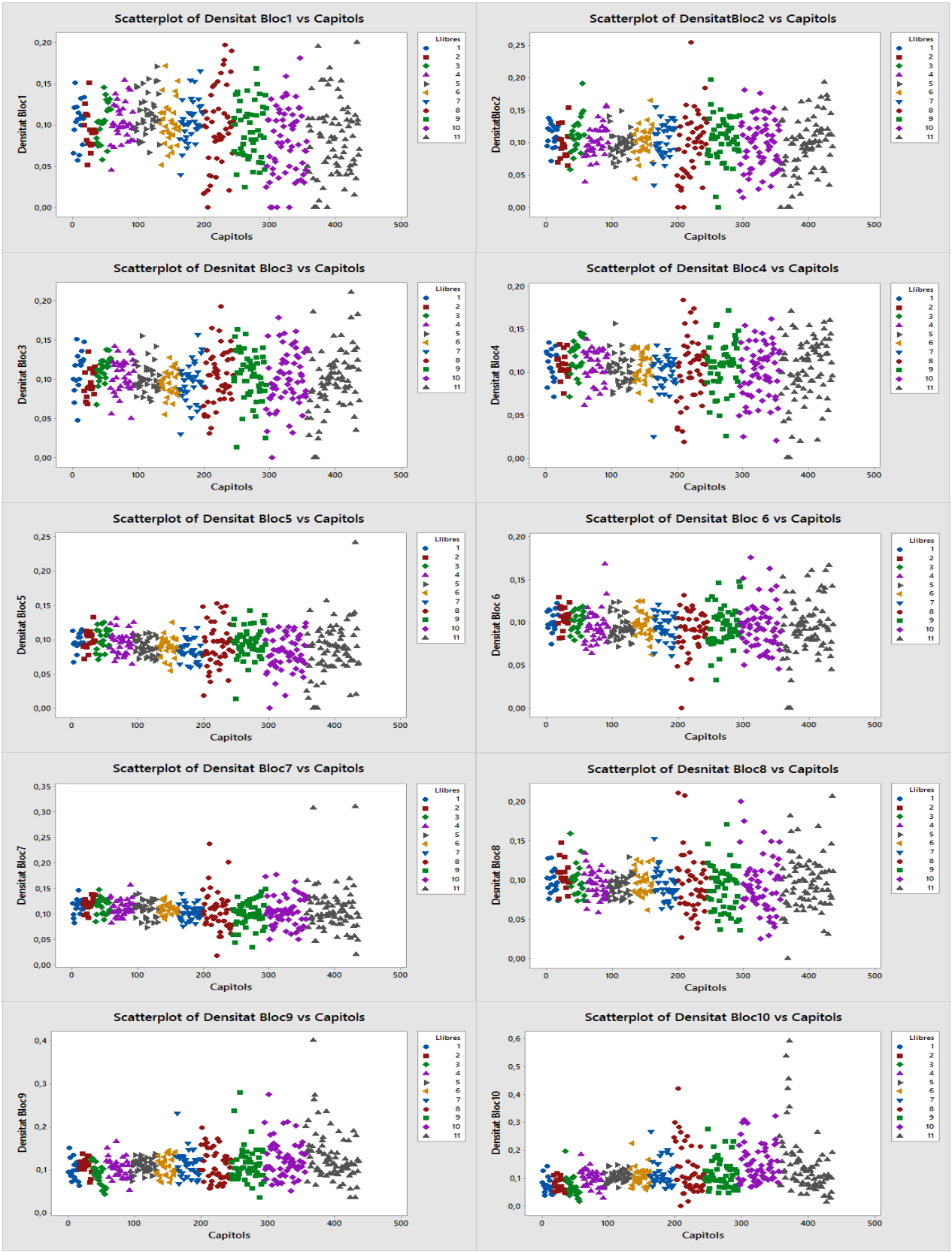

Figura 6.2 Gràfics dels 10 perfils de probabilitat

6.2. Diferències entre *HP* i *noHP*

Tal com s'ha fet en el capítol 5, es procedeix a fer un anàlisis de correspondències per a esbrinar més coses. En la Taula 6.3, es comprova que els dos eixos principals expliquen un 63,11% de la informació, molta. De fet, la primera component explica casi un 50% de la informació, i la segona només un 13,85%.

```
Analysis of Contingency Table

Axis  Inertia  Proportion  Cumulative  Histogram
    1   0,0317      0,4926      0,4926  ******************************
    2   0,0089      0,1385      0,6311  ********
    3   0,0054      0,0843      0,7154  *****
    4   0,0041      0,0632      0,7786  ***
    5   0,0035      0,0549      0,8334  ***
    6   0,0029      0,0458      0,8792  **
    7   0,0029      0,0445      0,9238  **
    8   0,0026      0,0399      0,9637  **
    9   0,0023      0,0363      1,0000  **
Total   0,0644
```

Taula 6.3 Sortida Minitab Anàlisi de correspondències

Observant el gràfic per les columnes de la Figura 6.3, es veu com la primera component ordena els tipus de frase de més petita (a l'esquerra de tot) a la més gran (a la dreta). La segona component només separa les categories 1, 2 i 10 (a la part superior), de la resta (zona inferior).

En els gràfics de la Figura 6.3 es posa de manifest com *HP* és més estable i els seus capítols es troben en un núvol de punts força compacte. Els *noHP* disposen d'una part de capítols centrats en el núvol de *HP,* però també presenta una desviació cap a la dreta en l'eix horitzontal. Tot i això, es comprova que la separació entre capítols de *HP* i capítols de *noHP* és menor que la observada al capítol 5 per llargada de paraula. Les observacions més distanciades de la resta pertanyen a *CV.*

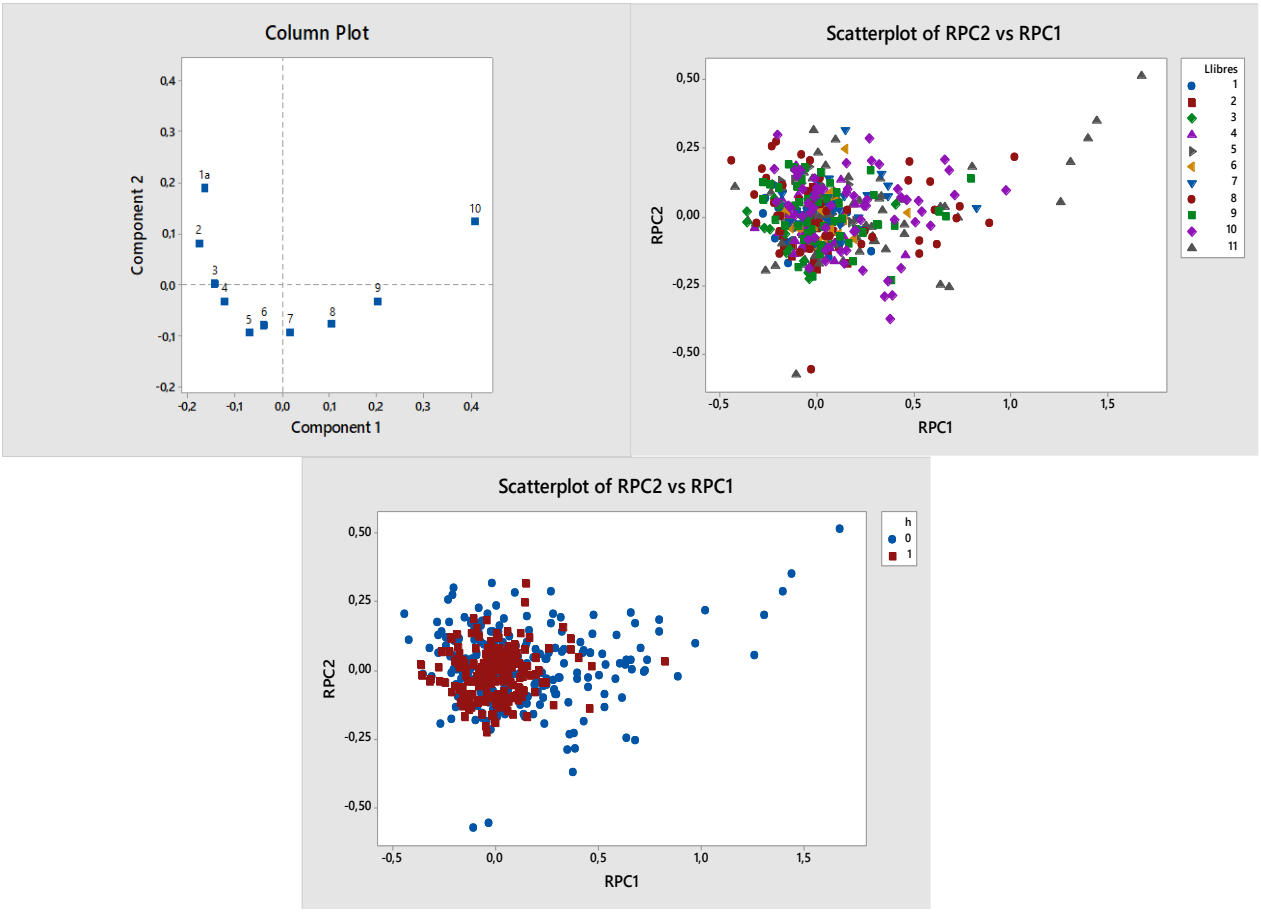

Figura 6.3 Gràfics Anàlisi de correspondències: per categories i per capítols, agrupats per llibres o *HP*

6.3. Es pot predir si és *HP* o *noHP*?

Com l'anàlisi de correspondències de la secció anterior sembla que no separa gairebé *HP* de *noHP*, un podria pensar que un anàlisi discriminant no acabés de distingir bé els capítols de *HP* dels de *noHP.* Altra vegada, però, l'anàlisi discriminant ha funcionat millor de l'esperat, classificant correctament un 73,9% dels capítols. Un 14% pitjor que per caràcters per paraula, però força bo al tenir en compte els gràfics observats. Això sí, la funció lineal era força pitjor, i s'ha hagut de realitzar un anàlisi discriminant quadràtic amb *cross-validation.* Els *HP* són molt més fàcils d'identificar (el seu núvol de punts és més compacte), per la qual cosa s'encerten un 88,4% de les vegades. Pel que fa als capítols de *noHP,* que estan més superposats amb els de *HP* que a l'estudi realitzat en el capítol 5, només es classifiquen bé un 61,8% dels casos.

```
Summary of Classification with Cross-validation

                 True Group
Put into Group       0      1
0                  175     91
1                   23    147
Total N            198    238
N correct          175    147
Proportion       0,884  0,618

N = 436           N Correct = 322           Proportion Correct = 0,739
```

Taula 6.4 Resultat anàlisi discriminant quadràtic amb *cross-validation*

Observant la Taula 6.5, els majors desencerts es troben a *R1* i *R2,* amb només una probabilitat d'encert del 50%. La meitat dels seus capítols s'assembla més a *HP*, i l'altre a *noHP.* En alguns gràfics de l'apartat 6.2, aquests dos llibres representen l'inici del canvi respecte *HP,* amb una progressió que es desmarca ja a partir del *R3.* Té sentit que sigui la causa d'aquests errors en la predicció. Un fenomen semblant ocorre per *H7,* però aquest en menor manera (25% de desencert). També es pot comprovar en el gràfics de la Figura 6.3 com els punts que més s'allunyen del centre, i per tant dels capítols *HP*, pertanyen molt més a *R3* i a *CV* que no pas a *R1* o *R2.* Probablement per això les seves prediccions són millors que en els dos primers thrillers, que es troben més a prop o coincidint amb el núvol de punts de *HP.*

S'ha de mencionar que hi ha forces casos de 4 o menys capítols desencertats seguits, amb tandes de capítols gairebé consecutius que estan mal predits. En aquest estudi, es classifiquen correctament tots els capítols de *H2,* i més d'un 94% dels capítols de *H1* i *H5*. Els capítols de *HP* més allunyats cap a la dreta del núvol de punts de *HP* en els gràfics d'eixos principals (Figura 6.3), coincideixen amb els errors en l'anàlisi discriminant.

H1	5,88%	**H5**	5,26%	**R2**	50%
H2	0%	**H6**	13,33%	**R3**	27,41%
H3	13,63%	**H7**	25%	**CV**	33,75%
H4	10,81%	**R1**	47,82%		

Taula 6.5 Desencerts en la predicció per novel·la

Un cop acabat l'estudi, si es compara amb el realitzat mesurant paraules per frase (no s'inclou en la memòria) enlloc de per caràcters, s'observa com per paraules l'anàlisi discriminant obté millors resultats a l'hora de classificar els capítols.

7. ÚS PARAULES

7.1. Presentació de les dades

Tota persona que escriu ho fa amb una sèrie de paraules. La majoria de les paraules més utilitzades per a cada persona solen ser més o menys les mateixes que per a tothom, però sí que varia la *freqüència* d'ús entre una persona i una altra. Semblava doncs interessant estudiar l'evolució de l'ús de paraules de J.K. Rowling al llarg de les seves 11 novel·les. Per fer-ho, primer de tot es van recollir totes les paraules usades amb la funció *Paraules()*, detallada en el capítol 4, a partir d'un document que contenia els 11 llibres junts. Se'n van seleccionar les 50 primeres i es va crear una llista en el programa definitiu que contenia aquestes 50 paraules més usades. En el procés de selecció es van eliminar els noms propis o paraules més relacionades amb la temàtica de les novel·les (com podria ser el nom *Harry* o *wand*, que significa vareta en anglès). Dins la mateixa funció, que també calculava la diversitat, es va extreure per cada capítol un document *txt* amb les vegades que s'utilitzava cada paraula. En la Taula 7.1 es mostra una porció de la taula realitzada.

	H1C1	H1C2	H1C3	H1C4	H1C5	...	Total
the	203	180	217	123	273	...	81240
to	94	65	91	53	122	...	42983
and	102	85	112	75	158	...	41788
of	75	45	72	54	113	...	34680
a	112	93	74	111	177	...	34091
he	137	90	77	103	119	...	31849
was	90	70	65	56	85	...	25411
his	69	63	68	65	56	...	21553
in	60	52	54	37	85	...	21325
had	40	65	34	24	49	...	20343

Taula 7.1 Les 10 paraules més usades

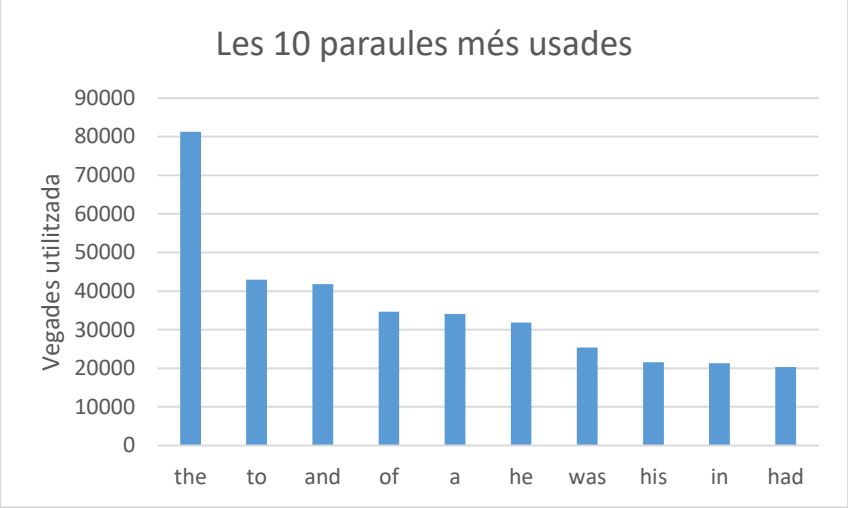

Figura 7.1 Gràfic amb les deu paraules més usades

Tal com es veu a la Figura 7.1, la paraula més usada en diferència és *the,* seguida per *to* i *and,* que apareixen gairebé la meitat de vegades que la primera. La desena paraula més utilitzada, *had,* no arriba a una quarta part de l'ús del *the.* Com els capítols no tenen la mateixa longitud de paraules, s'ha dividit cada valor pel nombre total de paraules del capítol al que pertany, obtenint així els perfils de probabilitat d'ús dins de cada text. Llavors s'ha graficat l'evolució dels perfils de cada paraula al llarg dels capítols.

Entrant a estudiar els gràfics de la Figura 7.2, pel que fa a la paraula *the,* no es distingeixen a simple vista diferències clares entre els capítols de *HP* i *noHP.* La variància, com sempre, és major pels capítols de les novel·les que no pertanyen a la saga de mags.

L'ús de la paraula *and* és més curiós. Els tres primers llibres, el cinquè i el sisè de *HP* tenen usos força semblants entre els seus propis capítols i entre ells. El quart sembla tenir dos tipus de capítols diferenciats: els que usen un 0,03% de vegades *and,* i els que l'utilitzen menys d'un 0,025%. Entre mig hi ha una separació no menyspreable. El setè llibre té cinc capítols que se'n van de la norma, amb una certa tendència a augmentar en els capítols finals. El seu ús total és major que als altres *HP.* Els *noHP* són més variables, tenint les mitjanes d'ús fortament diferenciades. Els thrillers van decreixent en la utilització de la paraula (destacant el tercer per ser el més compacte), però en el *CV* la mitjana remunta considerablement, tenint un ús força semblant que en la setena novel·la (la que el precedeix temporalment).

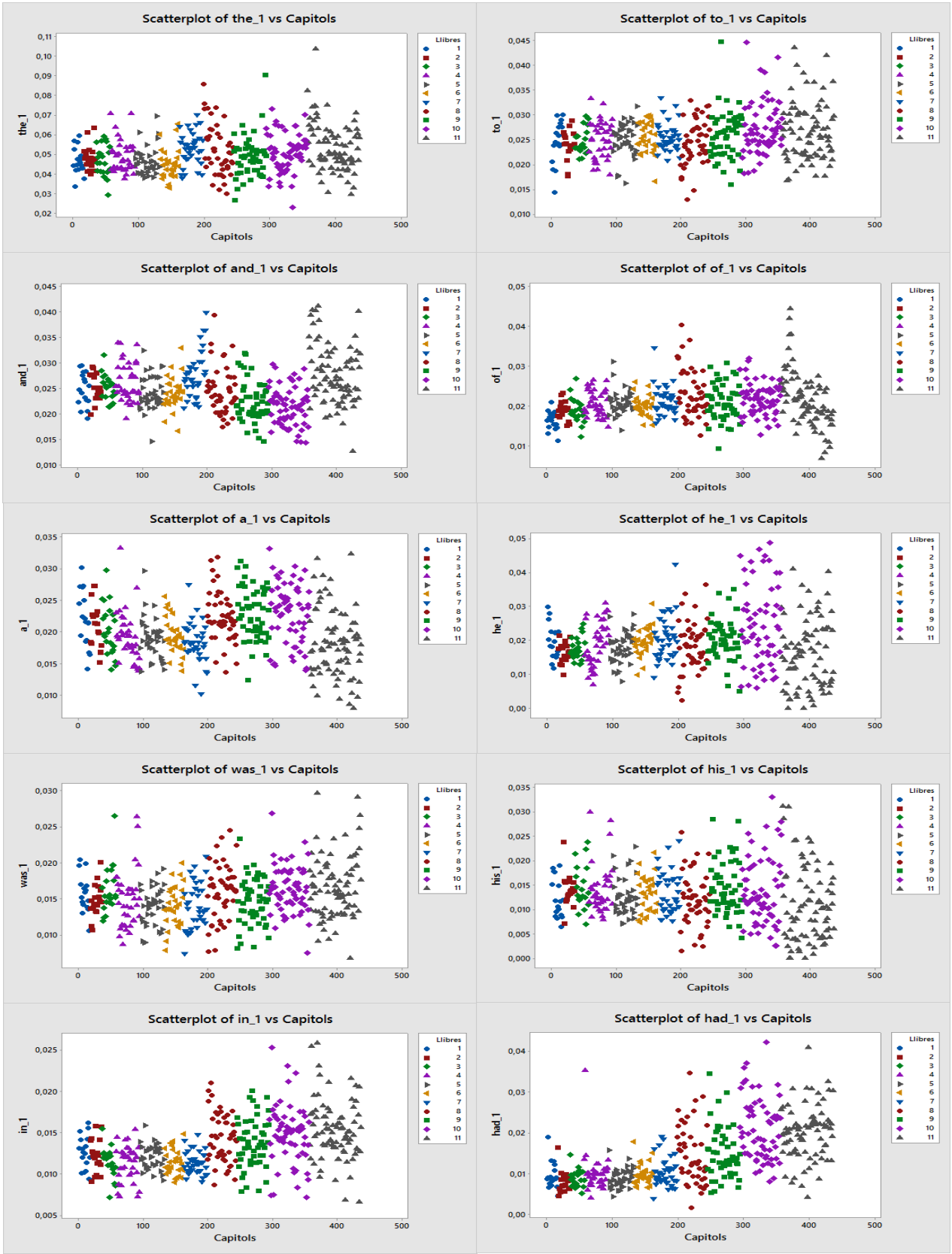

Figura 7.2 Gràfics perfils de probabilitats per cada paraula al llarg dels llibres

La preposició *a* presenta desviacions tipus força elevades i no constants. S'intueix una disminució en l'ús de la paraula en els capítols de les novel·les *HP. CV* té un ús semblant a l'últim episodi de la saga de mags, però molt més variable. Sembla que en aquest cas la progressió de Rowling és més temporal que en altres casos (recordar que *CV* es va escriure després de *H7*). Aquí la saga de misteri es desmarca de la resta, utilitzant molt més *a,* amb un lleuger increment al llarg de les 3 novel·les.

Pel que fa a *he*, els capítols del segon i tercer llibres són els més compactes. La quarta novel·la de *HP* torna a semblar estar diferenciada en dos tipus de capítols: la majoria de la primera part usa *he* menys d'un 0,02%, i la major part de la segona part l'utilitza més d'un 0,02%. A partir del sisè llibre comença a augmentar l'ús de la paraula, però decau força amb *CV* (altre cop el més variable).

Was segueix un ús més o menys constant al llarg dels llibres de Rowling. Excepte als seus dos darrers del gràfic, on apareix un creixement. Com la majoria de les vegades, els capítols de *noHP* presenten més variabilitat. Un altre aspecte rellevant és que el setè llibre de la saga de mags sembla tenir, majoritàriament, dos tipus de percentatge, però no segueixen una progressió temporal. Es podria afirmar que en total els *noHP* l'utilitzen més, però per poc.

En el gràfic per *his,* el setè llibre de *HP* també sembla seguir dos tipus diferenciats de capítols. El més compacte és el segon llibre. A més, els *HP* no semblen seguir una progressió: el primer presenta usos força baixos; el segon, el tercer i el sisè, de més elevats. Els capítols finals del quart llibre tendeixen a utilitzar més aquest pronom possessiu. A l'igual que el que passava amb *he*, però a partir dels últims capítols i no des de la meitat. Una possible explicació seria que en el final d'aquest quarta novel·la, els personatges que apareixen són només masculins. La narració llavors es centra en descriure tot el procés, sense saltar-se cap detall, de com els personatges manipulen diferents objectes: preparen una olla, es relata com un perd la mà però li'n posen una de nova, com lluiten amb les *seves* varetes, o el protagonista es retroba amb els *seus* familiars. Tot això podria ser l'explicació. Pel que fa als *noHP,* el primer thriller té un ús de la paraula *his* baix, però a partir del segon, que augmenta considerablement l'ús de la paraula respecte el seu anterior, s'inicia un descens. *CV* sempre el més variable.

El gràfic de *in* és molt més agradable d'estudiar. S'observen clares diferències entre els *HP* i *noHP,* tant a nivell de mitjana com de variabilitat. Pel que fa a la saga de mags, sembla seguir un subtil descens més o menys progressiu. Destaquen aquest cop les tres últimes entregues per la seva reduïda variabilitat. Pel que fa als *noHP,*

segueixen una progressió més o menys constant, utilitzant més *in* que *HP.*

Per últim, la desena paraula més usada, *had*, presenta el gràfic més evident a simple vista. Els *HP* segueixen una progressió constant i compacta en l'ús de la paraula, però els *noHP* l'utilitzen molt més, diferenciant-se els dos primers dels dos darrers. *R3* i *CV* encara utilitzen més *had,* però alhora presenten major variabilitat.

7.2. Diferències entre *HP* i *noHP.*

Com es tenen 10 dimensions (10 paraules més usades), i és impossible fer gràfics de deu dimensions, es decideix realitzar un anàlisi de correspondències a partir de la taula de contingència (Taula 7.1).

Analysis of Contingency Table

Axis	Inertia	Proportion	Cumulative	Histogram
1	0,0123	0,2712	0,2712	******************************
2	0,0116	0,2564	0,5276	*****************************
3	0,0060	0,1319	0,6595	**************
4	0,0044	0,0979	0,7574	**********
5	0,0030	0,0653	0,8226	*******
6	0,0028	0,0613	0,8839	******
7	0,0023	0,0511	0,9350	*****
8	0,0017	0,0370	0,9720	****
9	0,0013	0,0280	1,0000	***
Total	0,0453			

Taula 7.2 Resultats Minitab anàlisi de correspondències

Les dues primeres components (Taula 7.2) expliquen un percentatge de la informació molt semblant, sumant un 52,75% entre les dues.

En el gràfic de les columnes de la Figura 7.3, s'observa com la primera component (la horitzontal) separa un conjunt de paraules a la dreta del gràfic, i un altre a l'esquerra. Mentre que moltes es troben relativament a prop, situades en el centre, les paraules que es troben més distanciades són *his* i *he,* a l'esquerra, i *had* a la dreta de tot. La segona component (eix vertical), distingeix *his, he* i *had* de la resta. El resultat és un conjunt de categories en el centre, *his* i *he* a la part superior esquerra, i *had* a la zona superior de la dreta.

En el gràfic de baix de tot de la Figura 7.3, s'observen clares diferències en el en els centres dels capítols de *HP* i dels *noHP*, tendint aquests últims més a la dreta i ocupant

molt més espai, de dalt a baix del gràfic, mentre que *HP* és un cúmul de punts situat més a l'esquerra. Forces punts es superposen, però molts altres queden ben separats. D'aquests capítols de *noHP* que es troben més distanciats, la majoria formen part de *R3* i de *CV*. Sembla indicar que serà més fàcil distingir aquests que els dos primers thrillers dels no*HP.* Això ja s'intuïa en la majoria dels gràfics de l'apartat 7.1, on en *R1* i *R2* s'iniciava la diferenciació amb els *HP,* consolidada ja amb *R3* i *CV.*

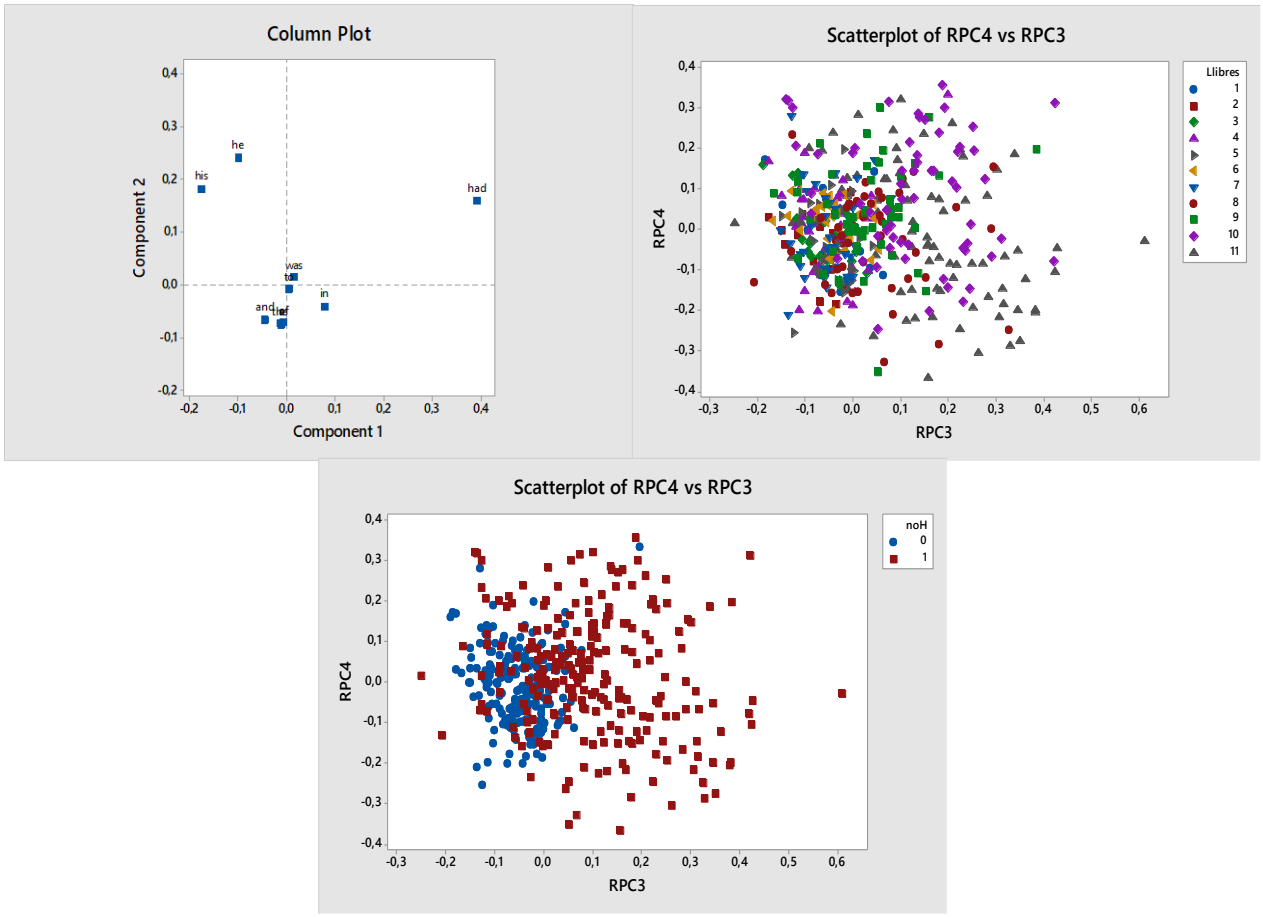

Figura 7.3 Gràfics anàlisi de correspondències: per categories i per capítols (segons llibre o *HP*)

La paraula *had* té sentit que es trobi més separada de la resta, en la zona on predomina *noHP,* i en especial *R3* i *CV,* ja que són els llibres que més usen la paraula. Pel que fa a *he,* en l'apartat 7.1 no s'han extret conclusions clares, però sí que per *his* s'ha observat com els capítols de *HP* l'utilitzen molt més. Per això és la categoria que es troba més a l'esquerra.

Les paraules *in, was* i *to,* situades en el centre però delimitades per la primera component a la part de la dreta, es caracteritzen per a ser paraules més utilitzades pels *noHP* (s'observa en els gràfics de la Figura 7.2).

Pel que fa a *and* i *his* situada a l'esquerra per la primera component, destaca per a utilitzar-se més per les novel·les de *HP.* Per tant, sembla ser que la primer component separa les paraules segons si són més usades pels *HP* o els *noHP*.

7.3. Es pot predir si és *HP* o *noHP*?

```
Summary of Classification with Cross-validation

                  True Group
Put into Group       0      1
0                  179     44
1                   19    194
Total N            198    238
N correct          179    194
Proportion       0,904  0,815

N = 436           N Correct = 373           Proportion Correct = 0,856
```

Taula 7.3 Resultats Minitab anàlisi discriminant quadràtic amb *cross-validation*

Tal com s'intuïa, *HP* i *noHP* estan força diferenciats. L'anàlisi discriminant permet classificar correctament el 85,6% dels 436 capítols. S'ha usat el mètode quadràtic i el cross-validation. Com sempre, es prediuen molt millor els *HP* (90,4%), que els *noHP* (81,5%).

H1	23,52%	**H5**	5,26%	**R2**	34%
H2	5,55%	**H6**	3,33%	**R3**	12,9%
H3	13,63%	**H7**	8,33%	**CV**	10%
H4	13,51	**R1**	23,91%		

Taula 6.4 Taula d'errors de predicció

Les novel·les que presenten més errors de predicció són els *R1* i *R2* (Taula 7.4). Tal com es preveia. En els gràfics de l'apartat 7.1 ja s'apreciava que els dos primers thrillers eren més semblants als *HP* que els dos últims llibres. En el núvol de punts de la Figura 7.3, *CV* i *R3* disposaven de més punts allunyats del cúmul de capítols de HP, sent més fàcils d'identificar.

Sorprèn el mal percentatge de *H1*, i destaca el *CV* per a només mostrar un 10% d'error en la previsió, i els *H2*, *H5*, *H6* i *H7* per la seva bona predicció. En definitiva, hi ha diferències clares entre *HP* i *noHP*.

8. DIVERSITAT I RIQUESA

8.1. Presentació de les dades

Hi ha moltes maneres diferents de mesurar la diversitat d'un text. Les eines que s'utilitzaran en el treball són:

- *Índex de Simpson, D*: eina típica de mesura de la diversitat. Més endavant s'explicarà amb més detall, però és important saber que és la mesura que menys depèn de *N*, el nombre total de paraules del text.
- *V*: indica el nombre de paraules diferents que apareixen en un capítol. Aquesta mesura presenta una distribució que depèn molt de *N*, fet que si els capítols no tenen llargades similars, alhora de comparar diversitats s'ha de tenir en compte la *N.* La *V* és una mesura de la diversitat molt concreta, anomenada usualment com a *riquesa*.
- V_1 i V_2: eines de diversitat que indiquen el nombre de paraules que apareixen 1 o 2 vegades, respectivament. Tal com passa amb *V*, depenen molt del nombre total de paraules del text.

El programa informàtic calcula les eines de mesura mencionades (*D*, *N, V*, V_1 i V_2) i es situen de manera que cada capítol disposi de cinc columnes, una per a cada eina de mesura de diversitat. En la Taula 8.1 es mostra un petit fragment de la taula resultant.

	D	N	V	V1	V2
H1C1	0.00756485	4581	1155	648	193
H1C2	0.00899508	3428	962	574	139
H1C3	0.0088615	3816	1100	684	156
H1C4	0.00669062	3678	1028	592	171
H1C5	0.00661482	6545	1597	871	270
H1C6	0.00751944	6261	1371	719	207
H1C7	0.00807036	4435	1237	707	204
H1C8	0.00761191	3043	1011	634	153
…	…	…	…	…	…
CC78	0.01051333	1171	503	352	69
CC79	0.01014263	789	339	230	54
CC80	0.00988678	7541	1910	1183	293

Taula 8.1 Diversitat per capítols

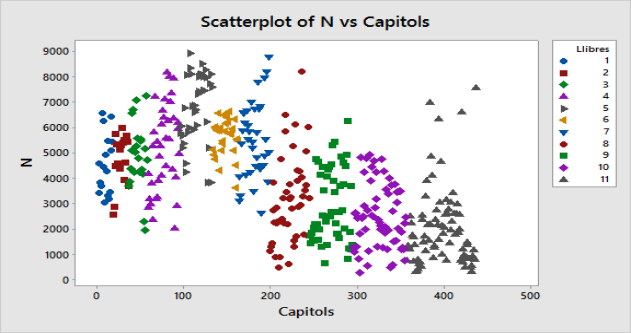

Figura 8.1 *N* al llarg dels capítols

És important prestar atenció a la Figura 8.1. S'observa com la *N* va augmentant fins *H6*, on pateix un fort decreixement. La majoria de capítols de *HP* oscil·len entre les 3000 i 6000 paraules. A partir de llavors, la *N* no fa més que disminuir al llarg de les novel·les restants, arribant alguns capítols a contenir menys de 1000 paraules. Per tant, al tractar-se de *N*'s tant diferents, no es podran comparar les eines així com així.

8.2. Índex de Simpson (diversitat)

L'índex de Simpson depèn menys de *N* (nombre de paraules d'un text) que les altres mesures mencionades. Aquest coeficient indica la probabilitat que al agafar dos elements, és a dir, dues paraules qualsevols d'un text, aquestes dues siguin iguals. Per tant, com major sigui el coeficient, implicarà que hi ha menor diversitat en el text. Com menor sigui la probabilitat de que al agafar dues paraules siguin iguals, menor serà *D* i major diversitat tindrà el text (un text divers és aquell que conté més vocabulari diferent). Per a entendre la fórmula a continuació, n_i és el nombre total de vegades que apareix la paraula *i* en un text, i *N* el total de paraules del text (capítol).

$$D = \frac{\sum_i n_i \cdot (n_i - 1)}{N \cdot (N - 1)}$$

En la Figura 8.2, el gràfic de la *D* al llarg dels capítols i en el de la *D* mitjana de cada novel·la, s'observa com la *D* augmenta a partir del setè llibre de la saga de mags. Per tant, això indicaria que els capítols de *HP* tenen menor diversitat. El llibre amb major coeficient de Simpson és *CV*. La progressió de *H7,* passant pels thrillers i arribant a *CV* seria una progressió força lineal, si no fos per *R2,* que sofreix una disminució considerable de la *D,* i per tant un augment en la diversitat. És interessant observar com *R1* té uns primers capítols amb la *D* molt elevada, i la meitat final amb una de més baixa, aproximant-se a *R2.*

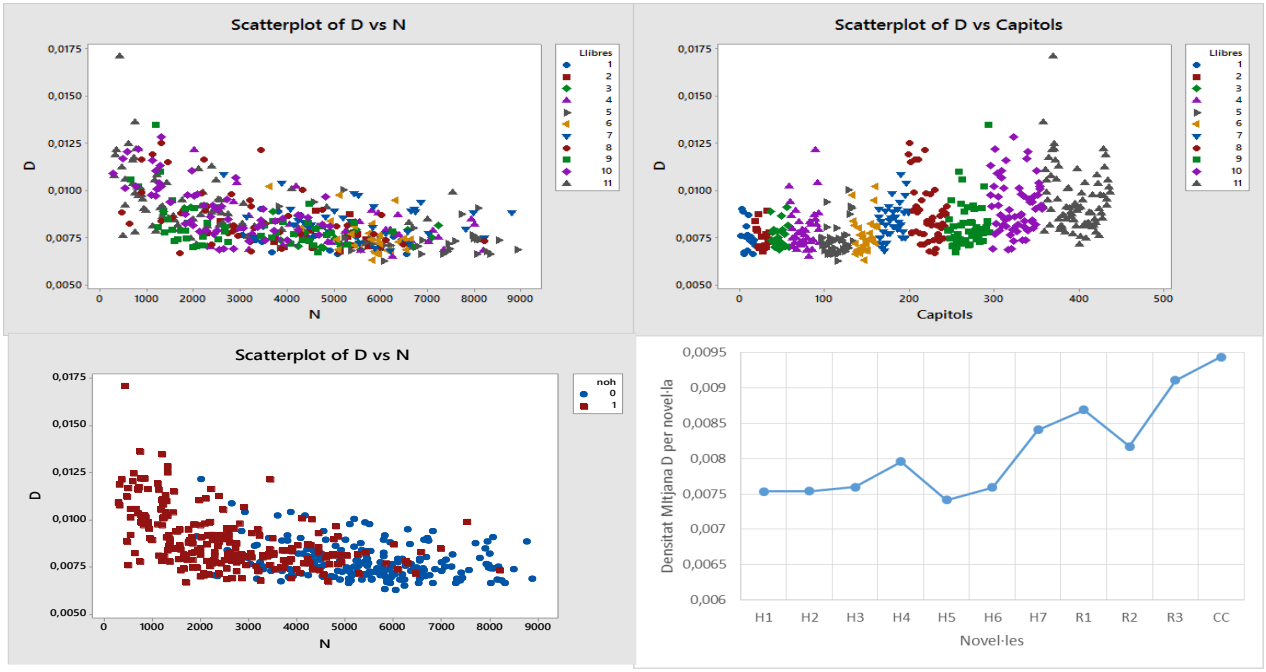

Figura 8.2 Gràfics *D*

Per a sortir de dubtes, i per evitar dependències amb la *N,* es va decidir realitzar una regressió lineal, amb la *D* com a resposta, i la variable binària *noh* com a explicativa (0 si és *HP,* 1 si és *noHP*). Aquest és un mètode de comparació de mitjanes. La constant és la mitjana, i el coeficient de la variable categòrica és la diferència de mitjanes. Si aquest coeficient surt significatiu, és que les mitjanes entre *noHP* i *HP* són suficientment diferents. Per a que el model compleixi les hipòtesis de linealitat i variància constant, s'ha treballat amb el logaritme de *D.*

```
Ln(D)  vs  noh

Regression Equation

lnD = -4,8649 + 0,1349 noh_1

Model Summary

        S     R-sq  R-sq(adj)     PRESS  R-sq(pred)
0,140817   18,61%      18,42%   8,68317      17,88%


Coefficients

Term          Coef  SE Coef         95% CI         T-Value  P-Value
Constant   -4,8649   0,0100  (-4,8846; -4,8452)   -486,13    0,000
noh
  1         0,1349   0,0135  ( 0,1083;  0,1615)     9,96    0,000
```

Taula 8.1 Regressió lineal *ln(D) vs noh*

Tal com es veu a la Taula 8.1, hi ha diferències significatives entre *HP* i *noHP.* La T frega el valor 10 (comença a ser significativa a 2, ja que s'agafa un interval de confiança del 95%). Per tant, *HP* és més divers que *noHP.*

Com a comentari, per a sortir de dubtes es van eliminar els capítols amb *N* baixes. Els resultats van ser els mateixos. *HP* és més divers pel que fa a l'eina de mesura *D.*

8.3. *V* (riquesa)

V mesura el nombre de paraules diferents que conté cada capítol. A diferència de *D*, es comprova que *V* depèn molt de *N.*

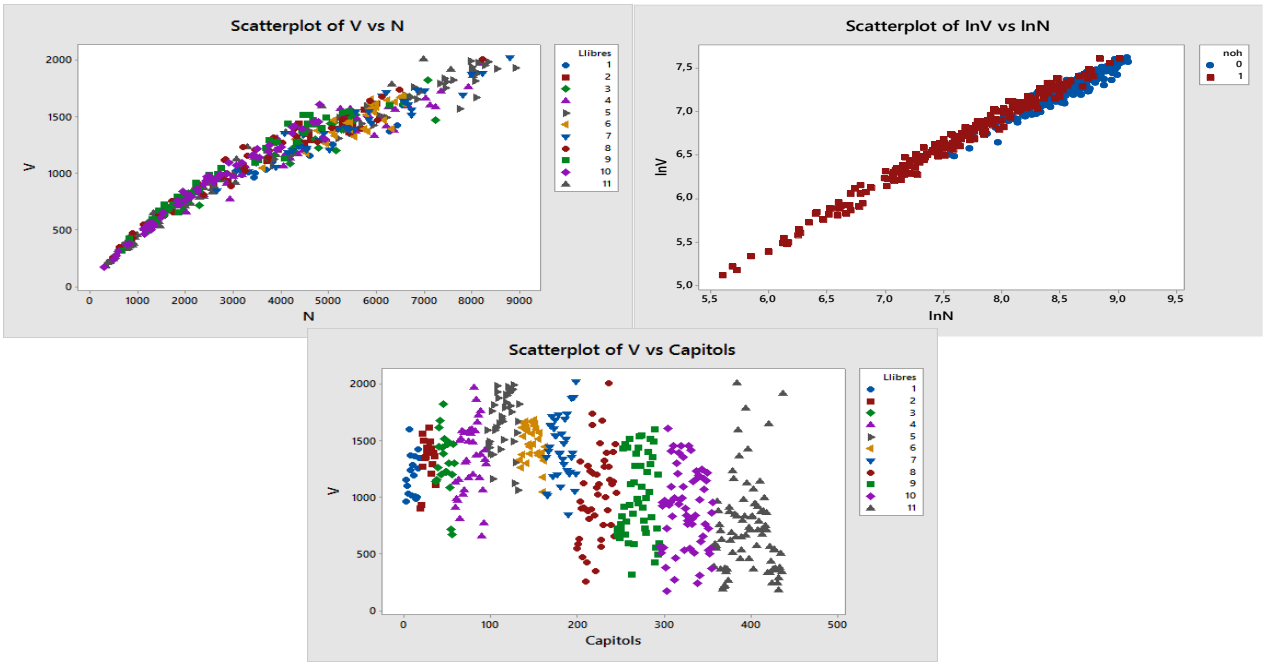

Figura 8.3 Gràfics de riquesa

El gràfic de *V* respecte els capítols (Figura 8.3) mostra com disminueix al llarg de les novel·les a partir de *H6*. Això no és conclusiu, ja que la *V* depèn de *N*, i els *noHP* tenen *N*'s molt petites (hi ha capítols que contenen menys de 1000 paraules). Es pot observar que la relació entre *V* i *N* no és lineal en el primer gràfic de la Figura 8.3. Per a sortir de dubtes, només queda fer una regressió lineal, amb la resposta *ln(V),* explicada per *ln(N)* i *noh*. Com *noh* és una variable categòrica, s'ha de crear una nova variable que és el producte de *ln(N)* per *noh*, i així poder distingir els pendents.

```
Regression Equation

noh
0    lnV = 1,346 + 0,6876 lnN

1    lnV = 1,0117 + 0,7411 lnN

Model Summary

         S    R-sq  R-sq(adj)  R-sq(pred)
0,0611072  98,47%     98,46%      98,44%


Coefficients

Term         Coef  SE Coef  T-Value  P-Value
Constant    1,346    0,125    10,78    0,000
lnN        0,6876   0,0146    47,23    0,000
lnN*noh    0,0535   0,0157     3,41    0,001
noh
  1       -0,334    0,133    -2,52    0,012
```

Taula 8.2 Restultats regressió LnV amb lnN i noh

Tal com es veu en el gràfic de la Figura 8.4, en els capítols amb menys de 600 paraules, les diferències són negligibles, però a partir de les 600, els *noHP*, sorprenentment, tenen una significativa millor riquesa. Cal mencionar que els pendents de les dues rectes són diferents, sent més fort el de *noHP*. Els residus surten bé, i s'explica més d'un 98% de la variabilitat de la resposta (R ajustada).Tant *lnN,* com *noh* i el producte dels dos tenen T's majors a 2. Tot i això, *noh* no té una T gaire elevada. Els residus són bons (no s'han inclòs en els resultats).

Que els capítols de les novel·les de *noHP* siguin més rics sembla contradir els resultats obtinguts en l'apartat anterior amb la D, i també el darrer gràfic de la Figura 8.3. Però, cal recordar que *V* depèn molt de *N,* cosa que fa que els gràfics de la Figura 8.3 puguin resultar enganyosos, i que diversitat i riquesa són dues mesures diferents.

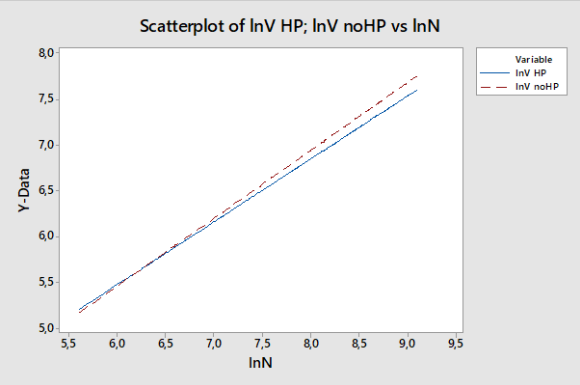

Figura 8.4 Rectes de regressió de ln(*V*) segons *noh*

8.4. V_1 i V_2 (diversitat)

S'estudiarà la V_1 (número de paraules diferents que apareixen només una vegada en el text) i V_2 (nombre de paraules diferents que apareixen només dues vegades en el text). Com més grans siguin, més divers serà el text. Però altre cop hi ha el problema de les *N* diferents i de la dependència de les eines de mesura amb *N*.

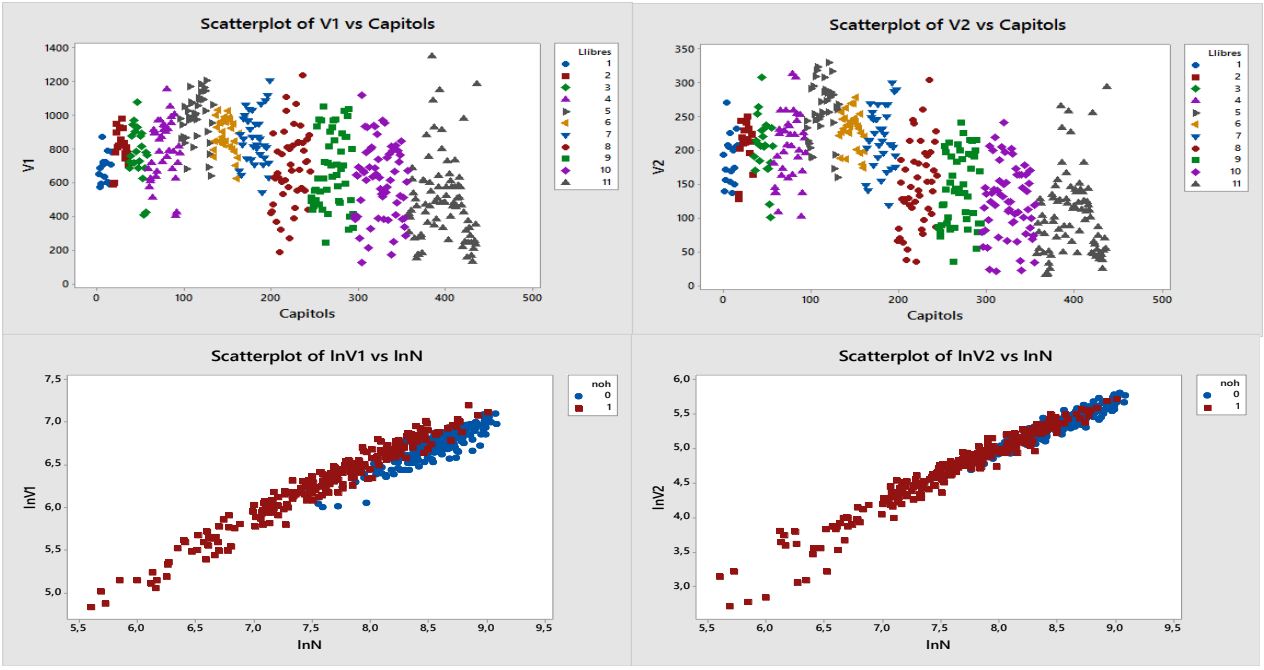

Figura 8.5 Gràfics de V_1 i V_2 respecte els capítols, i relació lineal entre logarítmes de les *V*'s amb *lnN*

Els gràfics de la Figura 8.5 són molt semblants als de la *V* (Figura 8.3). Es decideix realitzar un model lineal idèntic al de l'apartat anterior però intercanviant *lnV* per lnV_1. Per aquesta eina els logaritmes també redueixen la dependència amb *N*, tal com es veu en els gràfics de la part inferior de la Figura 8.5.

```
Regression Equation

noh
    0    lnV1 = 1,390 + 0,6221 lnN
    1    lnV1 = 1,1409 + 0,6751 lnN

Coefficients

Term         Coef  SE Coef  T-Value  P-Value
Constant    1,390    0,194     7,18    0,000
lnN        0,6221   0,0226    27,54    0,000
lnN*noh    0,0530   0,0243     2,18    0,030
noh
  1        -0,250    0,206    -1,21    0,225
```

Taula 8.4 Restultats regressió lnV_1 amb ln*N* i noh

La Taula 8.4 mostra els resultats del model per lnV_1. Tot i que *noh* apareix com a no significatiu, no es pot remoure del model, degut a que el seu producte amb *lnN* sí que és significatiu (T>2). *LnN* és molt significatiu. Per tant, els resultats afirmen que existeixen diferències entre *HP* i *noHP* pel que fa a la mesura de la diversitat per mitjà de V_1. Si s'observa la Figura 8.6, es comprova com els capítols de *noHP* són més diversos per qualsevol *N.*

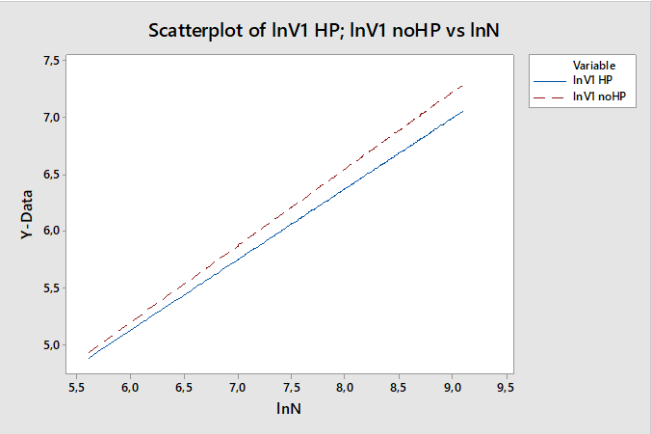

Figura 8.6 Gràfic de lnV_1 respecte *lnN,* diferençiant-se per *Hp* i *noHP*

Pel que fa a V_2, es procedeix d'igual manera. El model resultant (Taula 8.5) indica que la categoria *noh* és significativa. És a dir, hi ha diferències entre *HP* i *noHP* pel que fa a la diversitat mesurada per V_2. Però en aquest cas els resultats són més diferents. Si s'observa atentament la Figura 8.7, es veu com per *N*<2400 paraules, aproximadament, els capítols de *HP* són més diversos, i *noHP* menys. Cal mencionar, que *HP* gairebé no té capítols amb menys de 2400 paraules, però els *noHP* sí. A partir de *N*>2400, els capítols de *noHP* es tornen més diversos respecte *HP* (en aquest interval de *N, HP* conté la gran majoria de capítols). Per tant, com a conclusió, s'extreu que els capítols de les novel·les *noHP* majors de 2400 paraules són més diversos que els capítols de *HP.*

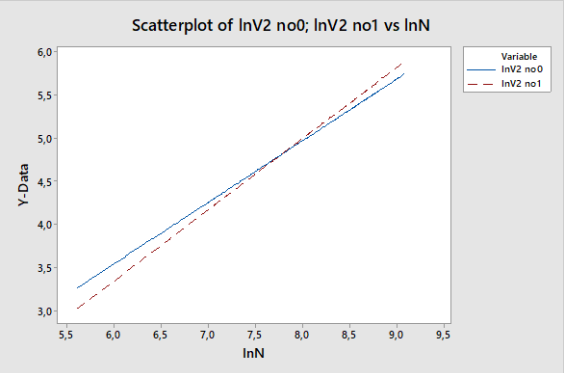

Figura 8.7 Gràfic de lnV_2 respecte *lnN,* distingint *noh*

```
Regression Equation

noh
0    lnV2 = -0,746 + 0,7138 lnN

1    lnV2 = -1,6002 + 0,8234 lnN

Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0,111316  96,17%     96,15%      96,06%

Coefficients

Term        Coef  SE Coef  T-Value  P-Value
Constant  -0,746    0,227    -3,28    0,001
lnN*noh   0,1096   0,0286     3,84    0,000
lnN       0,7138   0,0265    26,92    0,000
noh
  1       -0,854    0,241    -3,54    0,000
```

Taula 8.5 Resultats regressió lnV2

És interessant observar com diferents eines de mesura de diversitat poden donar resultats diferents. Mentre que *D* mesura major diversitat per *HP*, *V, V_1* i *V_2* mesuren major diversitat per *noHP*. De fet, observant els gràfics d'aquestes tres últimes eines al llarg dels capítols (Figures 8.3 i 8.5), s'aprecia com són gairebé idèntics.

Un cop finalitzat l'estudi de diversitat i riquesa, el lector es preguntarà per què no s'ha realitzat un anàlisi discriminant com en els capítols anteriors. La resposta és que, al dependre les eines de mesura de *N*, no es podien relacionar de cap manera en l'anàlisi discriminant. Per això s'ha procedit a la realització de models lineals, per a comparar mitjanes entre els capítols de *HP* i *noHP.*

També cal mencionar que una eina molt utilitzada en la mesura de la diversitat, que és l'Entropia, no s'ha usat degut a la seva elevada dependència amb *N.*

9. ALTRES ESTUDIS

Un cop acabat l'estudi principal, s'ha decidit afegir un novè capítol on ampliar-lo i donar idees sobre altres possibles respostes, donant la possibilitat a que l'autor o algú continuï amb l'estudi. Per tant, els següents apartats estaran menys detallats.

9.1. Anàlisi discriminant amb totes les variables

Cap al final del treball va sorgir una idea: fer un anàlisi discriminant amb totes les variables. S'intentaria esbrinar si totes elles juntes expliquen si és *HP* o no. Per a fer-ho, s'ha de tenir present les variables redundants, com per exemple en *lletres per paraula,* que només s'han d'agafar 9 variables de les 10 existents, ja que sumen 1 en total (una d'elles és prescindible). Si s'afegís, generaria redundàncies i el programa començaria a trobar correlacions massa elevades.

S'han utilitzat com a variables explicatives la *D,* les lletres per paraula, els caràcters per paraula i les 10 paraules més usades. La *V* no s'ha agafat degut a l'alta relació que té amb la llargada dels capítols. L'anàlisi discriminant s'ha fet quadràtic amb *cross-validation.* Els resultats no poden ser millors (Taula 9.1). Es classifiquen correctament un 93,8% dels capítols. Un 92,9% d'encerts en *HP* i un 94,5% en *noHP* (curiós, normalment *noHP* era el pitjor explicat). Cal mencionar que la *D* no afecta en l'estudi. El resultat és idèntic si la *D* s'utilitza en el model o no.

La conclusió és la mateixa que s'extreia dels darrers apartats: hi ha diferències significatives entre *HP* i *noHP.* L'estil de Rowling varia entre els dos grups.

Summary of Classification with Cross-validation

	True Group	
Put into Group	0	1
0	184	13
1	14	225
Total N	198	238
N correct	184	225
Proportion	0,929	0,945

N = 436 N Correct = 409 Proportion Correct = 0,938

Taula 9.1 Anàlisi discriminant *HP* o *noHP*

9.2. Anàlisi discriminant de les 11 novel·les

A posteriori es va intentar realitzar el mateix estudi, però enlloc de diferenciar només entre *HP* i *noHP*, fer-ho envers les 11 novel·les. És a dir, es poden classificar correctament els capítols segons la novel·la a la que pertanyen? El *Minitab* falla per anàlisi quadràtic. S'ha de fer un anàlisi discriminant lineal amb cross-validation. Els resultats no estan gens malament. Com la taula del programa contenia masses números i era difícil de llegir, s'ha fet un resum en la Taula 9.2.

Es classifiquen bé un 47% dels capítols. Si algú es dediqués a escollir nombres aleatoris de l'1 a l'11, i els agafés com a prediccions, es classificarien bé un de cada 11 capítols, un 9,09%. Aquí hi ha un 47% d'encerts. No és una predicció ideal, ja que més de la meitat de les vegades la predicció seria errònia, però indica que hi ha diferències suficients com a distingir forces vegades unes novel·les d'altres. La novel·la pitjor predita és *H4* amb un 29% d'encerts. La millor és *H3*, acostant-se al 60%, i fins a 4 novel·les superen la meitat de prediccions ben classificades.

Total	**47%**
H1	47,1%
H2	33,3%
H3	59,1%
H4	29,7%
H5	55,3%
H6	46,7%
H7	44,4%
R1	41,3%
R2	42%
R3	51,6%
CV	55%

Taula 9.2 Anàlisi discriminant per els 11 llibres, en percentatge segons encerts

9.3. El setè llibre de *Harry Potter*

Durant el transcurs de l'estudi, s'ha pogut anar observant com *H7*, segons el gràfic, començava a tendir cap a *R1.* Aquesta sensació ha pres força en l'estudi de riquesa i diversitat. Sobretot pel que fa a *V.*

Així que es va decidir efectuar un petit canvi. Enlloc de separar *HP* de *noHP*, s'afegiria

l'últim llibre de *Harry Potter* com a *noHP.* És a dir, la creació d'una variable binària que fos 0 per les novel·les de *l'H1* a *H6*, i 1 a partir de *H7* fins *CV*. La variable es va anomenar *7-11.*

Primer de tot es va realitzar un anàlisi discriminant per a llargada paraula, després per llargada frase, i finalment per ús de paraules. El total dels capítols classificats correctament era sempre força elevat, però més baix que a l'estudi amb *noh.* Si es miraven amb atenció els errors, s'observava que els nous errors apareguts al distingir per la variable *7-11,* els mateixos que disminuïen el percentatge d'encerts, eren els capítols de *H7.* És a dir, el setè llibre de *Harry Potter* té més en comú amb els altres volums de la saga que no amb *noHP.* Es podria provar, però, alguna combinació de novel·les o d'altres estudis per a veure més relacions entre *H7* i *R1,* i també *H7* i *CV,* la novel·la que es va escriure després.

9.4. Els 3 blocs

Una altra idea era, enlloc de separar les novel·les en 2 grups (*HP* i *noHP*), per què no separar per sagues (*HP, R* i *CV*).

Summary of Classification with Cross-validation

	True Group		
Put into Group	0	1	2
0	180	10	0
1	9	133	34
2	9	15	46
Total N	198	158	80
N correct	180	133	46
Proportion	0,909	0,842	0,575

N = 436 N Correct = 359 Proportion Correct = 0,823

Taula 9.3 Anàlisi discriminant 3 grups

Tal com es veu a la Taula 9.3, hi ha un 82,3% d'encerts. No està gens malament, però és pitjor que per *noh.* El grup més mal predit és el 2, que equival a *The Casual Vacancy*. Si s'observen bé tots els seus errors de predicció, aquests són perquè el programa els confon amb *R*, però mai amb *HP.*

Com a conclusió, la majoria de capítols de *HP* estan ben predits, i els de *CV,* tot i ser suficientment diferents als demés grups com per a tenir més de la meitat d'encerts, es confonen només amb els thrillers, tenint per tant moltes més semblances amb ells que no amb *HP.*

10. PRESSUPOST

És de menester realitzar un pressupost del treball realitzat. Degut a que aquest projecte no requeria més que un ordinador amb alguns programes i una sola persona treballant, serà breu.

Per a comptabilitzar el preu pel treballador (el propi autor), s'ha suposat que aquest cobra 10€ l'hora i que en treballa 300, ja que el treball equival a 12 crèdits, i 1 crèdit comporta 25h d'estudi (segons estipula l'escola).

També s'ha considerat el preu de l'ordinador. Suposant que té una durada de vida de 3 anys, que val 700€, i que només s'ha utilitzat durant 4 mesos pel treball (4 mesos entre 3 anys = una novena part del cost). El *Minitab* s'ha considerat gratuït gràcies a la llicència de l'escola, però el Microsoft Office sí que s'ha comptat, assumint que la llicència és per universitaris (la més barata), i que s'ha utilitzat 4 dels 12 mesos oferts.

PRESSUPOST TREBALL	
Treballador: 300 hores * 10€/h	3.000€
Ordinador: 1/9 * 700€	77,77€
Minitab (gratuït a l'utilitzar la llicència de l'escola)	0€
Microsoft Office llicència per universitaris: 1/3 * 80€	26,66€
Cost total del treball	3.104,43€

11. CONCLUSIONS

Un cop finalitzat l'estudi, arriba el moment de resumir les principals conclusions del treball.

General

- Diferències clares, en tots els estudis (particulars i generals), entre els *Harry Potter* i els altres 4 llibres que no formen part de la saga de mags. En l'estudi general, l'anàlisi discriminant ha encertat un 93,8% dels 436 capítols.

- En un anàlisi general diferenciant els 11 llibres, el discriminant obté suficients bons resultats. No per a obtenir una predicció molt segura, però sí per a demostrar que les diferències entre novel·les són notables.

- *H7* presenta una lleugera inèrcia a tendir cap a *noHP*, però sense acabar assemblant-s'hi massa.

- Les novel·les *R1* i *R2* solen estar en un entremig entre els *HP* i els *noHP*, segons l'estudi. El cas més extrem és el de la longitud de frases, on els errors en la classificació de l'anàlisi discriminat d'aquests dos llibres arriba al 50%.

- Es confirma que ha valgut la pena moure *The Casual Vacancy* de lloc en els gràfics. Les dades mostren una millor progressió si es mou *CV* de la vuitena posició (la que li pertoca temporalment) a la onzena.

- L'anàlisi clúster no funciona amb les dades del treball. L'estudi no supervisat no troba diferències suficients.

- Com millor es separa el gràfic de correspondències, millor prediu l'anàlisi discriminant.

- Mesurar les frases per paraules aporta millors resultats en l'anàlisi discriminant. Tot i això, mesurar per caràcters ofereix bons resultats. Si s'hagués fet l'anàlisi general mesurant les frases per paraules i no per caràcters, possiblement hauria retornat millors classificacions.

- Els *noHP* presenten major variància en els gràfics, mentre que *HP* són més compactes. *CV* és el més dispers, i també el que conté més capítols, amb llargades ben diferents (disposa de capítols de 600 paraules i d'altres amb més de 6000).

- Pel que fa al nombre de capítols per novel·les i a les seves longituds, mentre els *HP* en tenen menys, amb longituds entre 3000 i 6000 paraules, els *noHP* contenen més capítols (progressivament), amb una bona part que no supera les 1000 paraules.

Longitud Paraules:

- Les variàncies dels capítols són tal com s'ha mencionat anteriorment. Majors pels capítols de *noHP* i menors pels *Harry Potter.*

- La mitjana de lletres per paraula oscil·la majoritàriament entre 4,3 i 4,6. *CV* presenta forces excepcions, i *H1* té menor mitjana que la resta. Pel que fa a les mitjanes de llargades de paraules de les novel·les, *H6* és el que té major mitjana, i *H7* la menor. Els *noHP* casi no presenten canvis en la mitjana, mentre que els llibres *HP* varien força entre ells.

- *noHP* utilitza més les paraules de 3, 6 i 7 lletres. Els *HP* les de 2, 4 i 5 caràcters. Això es pot apreciar en els gràfics bivariants, i també en l'anàlisi de correspondències, on es distingeixen els dos grups en núvols amb centres diferenciats, cada un més proper als caràcters que més utilitza. La primera component ajuda a diferenciar les categories de paraules amb major ús de les que no. Les menys utilitzades són les paraules d'1, 8, 9 i 10 o més lletres.

- L'anàlisi discriminant classifica correctament un 87,2% dels capítols. Les diferències entre *HP* i *noHP* són considerables. *H7* presenta més errors de predicció, probablement perquè en els gràfics es mostra un inici de tendència a semblar-se a *R1*.

Longitud de frases segons el nombre de caràcters:

- Les mitjanes de les llargades de les frases per les novel·les segueix majoritàriament un creixement. Mentre que *H1* i *H3* disposen de menor mitjana, *CV* és la novel·la amb la mitjana de llargada de frases més elevada.

- Les frases de 1 a 11 caràcters (primera categoria) s'utilitzen més en *HP*.

- Les novel·les amb més caràcters per frase són *R3* i *CV.*

- L'anàlisi de correspondències presenta dos núvols de punts per als capítols amb centres poc diferenciats. Els capítols dels llibres de *noHP,* com sempre, els més disperos.

- La primera component ordena les 10 categories de frases de menor longitud (esquerra) a major (dreta).

- L'anàlisi discriminant prediu bé un 73,9% dels capítols. Això sí, *R1* i *R2* només encerten el 50% de les prediccions. És a dir, la meitat dels seus capítols són classificats com a *HP*, i l'altre meitat com a *noHP*. *H7* és el llibre de *HP* que presenta més errors de predicció (25%), degut a que s'assembla força als thrillers en comparació amb els altres llibres de la saga de mags.

- Els capítols més allunyats de les novel·les de *HP* en el gràfic de l'anàlisi de correspondències, apareixen com a errors en la predicció de l'anàlisi discriminant. *CV* i *R3* tenen forces capítols allunyats del centre del núvol de punts en els gràfics de correspondències. Són els més ben predits de *noHP* en l'anàlisi discriminant, mentre que *R1* i *R2* es troben superposats o més propers al núvol de punts de *HP*.

Ús de les paraules

- Segons cada paraula, els perfils de probabilitats de cada llibre varien bastant.

- *H4* presenta alguns trets especials. Per la paraula *and*, hi ha dos tipus de capítol. Uns tenen una mitjana força alta i els altres una de més baixa. L'espai entre els dos grups de capítols és força important (com una línia horitzontal que separa el grup de dalt amb el de baix). Pel que fa a la paraula *he*, també hi ha dos grups diferenciats, però a més ho són a nivell temporal. Es distingeix una primera meitat amb uns capítols amb poc ús de la paraula, i en els darrers va augmentant. Amb el pronom *his*, els últims capítols de la quarta novel·la l'utilitzen molt més. Una possible resposta és l'elevat nombre de personatges masculins que apareixen en la part final i la quantitat d'accions concretes que realitzen, utilitzant més del compte el pronom possessiu *his*.

- Les paraules situades a la dreta per la primera component, s'utilitzen més en *noHP*. Són *had, in, was* i *of*. *had* presenta una progressió perfecte. Des de *H1* creix fins a arribar a un *CV* amb una densitat altíssima. A l'esquerra hi ha les que s'utilitzen més en *HP*.

- L'anàlisi de correspondències presenta pels capítols dos núvols de punts força diferenciats.

- L'anàlisi discriminant classifica correctament un 85,6% dels capítols.

Diversitat i riquesa

- Es comprova com diversitat i riquesa no són ben bé el mateix, i com diferents eines de mesura de diversitat poden donar resultats diferents. Mentre que per l'índex de Simpson (*D*), *HP* és més divers, per *V, V_1* i V_2, els resultats són que *noHP* és més divers. De fet, per V_2 els capítols de *noHP* són només més diversos a partir de *N*>2400 paraules.

- *N* està altament relacionada amb les eines de mesura de la diversitat i riquesa. Per a linealitzar la seva relació amb les altres variables s'han d'aplicar logaritmes.

BIBLIOGRAFIA CONSULTADA

- Diversity of vocabulary and Homogeneity of Literary Style (A. Riba & J. Ginebra)

- Change-point Estimation in a Multinomial Sequence and Homogeneity of Literary Style (A. Riba & J. Ginebra)

- Bayesian Analysis of a Multinomial Sequence and Homogeneity of Literary Style (J. Girón, J. Ginebra & A. Riba)

- Authorship Attribution (P. Juola): http://www.mathcs.duq.edu/~juola/papers.d/fnt-aa.pdf

- La práctica del análisis de correspondencias (M. Greenacre)

- https://en.wikipedia.org/wiki/Stylometry (Febrer 2017)

- http://www.philocomp.net/humanities/signature.htm (Febrer 2017)

- https://en.wikipedia.org/wiki/Linguistics_and_the_Book_of_Mormon#Stylometry_.28wordprint_studies.29 (Febrer 2017)

- https://support.minitab.com/en-us/minitab/18/ (Febrer – Juny 2017)

ANNEX

S'ha afegit el codi del programa (per l'estudi de la primera novel·la de la saga de *Harry Potter*) en forma d'imatges per a una millor lectura. El IDLE no permet importar el text, i fent el copiar-enganxar tradicional es perden els colors, que ajuden molt en la comprensió.

```
import re


def ParaulesFrase(f):
    Numerossols=[]
    Numeros=[]
    www=1
    while www<201:
        Numerossols.append(www)
        Numeros.append([www,0])
        www=www+1
    a=re.split('[.;!?\n\t…]',f) #Ja tinc llista amb frases
    l=Numeros

    for e in a: #frase
        x=e.split() #frase dividida en paraules
        n=len(x)
        k=False
        if n!=0:
            for  i in x:
                if (i=='' or i==' ' or i==',' or i=='.' or i=='`' or i=='´'
                    or i=="'" or i==';' or i=='(' or i==')' or i=='"' or
                    i== '-' or i=='’'or i=='”' or i==',’'):
                    n=n-1

        if n in Numerossols:
            lloc=Numerossols.index(n)
            l[lloc][1]=l[lloc][1]+1

    longitud=0 #frases totals
    q=0 #nombre total de paraules
    for c in l:
        if c[1]!=0:
            q=q+c[0]*c[1]
            longitud=longitud+c[1]
    mitjana= round (q*1.0/longitud,4)

    mitjana=str(mitjana)
    q=str(q)
    longitud=str(longitud)
    rr='HlPaFra'+str(titol)+'.txt'
```

```
    outfile=open(rr,'w')
    outfile.write(NomS)
    outfile.write('\n')
    outfile.write(q)
    outfile.write('\n')
    outfile.write(longitud)
    outfile.write('\n')
    outfile.write(mitjana)
    outfile.write('\n')
    for t in l:

        outfile.write(str(t[l]))
        outfile.write('\n')

    outfile.close()



def CaractersFrase(f):
    llista=(['q','w','e','r','t','y','u','i','o','p','a','s','d','f','g','h',
             'j','k','l','ñ','z','x','c','v','b','n','m'])

    Numerossols=[]
    Numeros=[]
    www=1
    while www<201:
        Numerossols.append(www)
        Numeros.append([www,0])
        www=www+1
    a=re.split('[.;!?\n\t…]',f) #Ja tinc llista amb frases
    l=Numeros
    longitud=0

    ll=llista+['—']+[',']
    for e in a:
        #per a cada frase
        i=0
        e=e.lower() #tot minúscula
         for c in e:
             if c in ll:
                 i=i+1 #Sumo els bons caracters


             #Sé la longitud de la frase e

         if i in Numerossols:

             lloc=Numerossols.index(i)
             l[lloc][1]=l[lloc][1]+1
         i=0
             #Ja he sumat a llista resultats
             #S'acaben les frases
```

```
    #Ara toca guardar resultats en bloc notes
    q=0
    for c in l:
        if c[1]!=0:
            q=q+c[0]*c[1]
            longitud=longitud+c[1]

    mitjana= round (q*1.0/longitud,4)
    mitjana=str(mitjana)
    q=str(q)
    rr='HlCaFra'+str(titol)+'.txt'
    outfile=open(rr,'w')

    longitud=str(longitud)
    outfile.write(NomS)
    outfile.write('\n')
    outfile.write(q)
    outfile.write('\n')
    outfile.write(longitud)
    outfile.write('\n')
    outfile.write(mitjana)
    outfile.write('\n')
    for t in l:
        outfile.write(str(t[1]))
        outfile.write('\n')

    outfile.close()

def CaractersParaules(f):
    llista=(['q','w','e','r','t','y','u','i','o','p','a','s','d','f','g',
             'h','j','k','l','ñ','z','x','c','v','b','n','m'])

    Numerossols=[]
    Numeros=[]
    www=1
    while www<40:
        Numerossols.append(www)
        Numeros.append([www,0])
        www=www+1
    a=f.split()
```

```
    l=Numeros
    longitud=0
    for e in a:
        #paraules
        i=0
        e=e.lower()
        for c in e:
            if c in llista:
                i=i+1
            #Sé caracters per paraula
        if i in Numerossols:
            lloc=Numerossols.index(i)
            l[lloc][1]=l[lloc][1]+1
        i=0
```

```
    q=0

    for c in l:
        if c[1]!=0:
            q=q+c[0]*c[1]
            longitud=longitud+c[1]
    mitjana= round (q*1.0/longitud,4)
    mitjana=str(mitjana)
    q=str(q)
    rr='HlCaPa'+str(titol)+'.txt'
    outfile=open(rr,'w')
    longitud=str(longitud)
    outfile.write(NomS)
    outfile.write('\n')
    outfile.write(q)
    outfile.write('\n')
    outfile.write(longitud)
    outfile.write('\n')
    outfile.write(mitjana)
    outfile.write('\n')
    for t in l:
        outfile.write(str(t[1]))
        outfile.write('\n')
    outfile.close()

def Lletres(f):
    llista=(['q','w','e','r','t','y','u','i','o','p','a','s','d','f','g',
             'h','j','k','l','ñ','z','x','c','v','b','n','m'])

    a=f.split()
    lletrescapitol=0
    l=([['q',0],['w',0],['e',0],['r',0],['t',0],['y',0],['u',0],['i',0],
        ['o',0],['p',0],['a',0],['s',0],['d',0],['f',0],['g',0],['h',0],
        ['j',0],['k',0],['l',0],['ñ',0],['z',0],['x',0],['c',0],['v',0],
        ['b',0],['n',0],['m',0]])
    for e in a: #per cada paraula
        e=e.lower()
        for c in e:
            if c in llista:
                lletrescapitol=lletrescapitol+1
                lloc=llista.index(c)
                l[lloc][1]=l[lloc][1]+1
    q=0
    for p in l:
        q=q+p[1]
    q=str(q)
    rr='HlLletr'+str(titol)+'.txt'
    outfile=open(rr,'w')
```

```
    outfile.write(NomS)
    outfile.write('\n')
    outfile.write(q)
    outfile.write('\n')
    for tt in l:
        outfile.write(str(tt[1]))
        outfile.write('\n')


    outfile.close()

def Paraules(f):
    llista=(['q','w','e','r','t','y','u','i','o','p','a','s','d','f','g','h',
             'j','k','l','ñ','z','x','c','v','b','n','m'])
    llparaules=([['the',0],['to',0],['and',0],['of',0],['a',0],['he',0],
                 ['was',0],['his',0],['in',0],['had',0],['said',0],['it',0],
                 ['you',0],['that',0],['i',0],['her',0],['she',0],['at',0],
                 ['on',0],['with',0],['as',0],['him',0],['for',0],['for',0],
                 ['but',0],['not',0],['they',0],['be',0],['out',0],['were',0],
                 ['up',0],['what',0],['all',0],['have',0],['from',0],['been',0],
                 ['them',0],['there',0],['into',0],['back',0],['this',0],
                 ['so',0],['who',0],['an',0],['could',0],['me',0],['no',0],
                 ['about',0],['when',0],['is',0]])


    a=f.split()
    l=[]
    ll=llista+['-']
    for e in a:

        e=e.lower() #tot a minuscula
        q=''
        for r in e:
            #mirem caràcters

            if r in ll:
                q=q+r
            # El I've el converteix en ive!!!!!!
            #Enlloc de "e" ara faig servir "q"


         if q!='' and q!=' ':
             k=False
             for x in l:
                 if q==x[0]:
                     x[1]=x[1]+1
                     k=True
                     break
                 else:
                     continue
             if k==False:
                 l.append([q,1])
                 l.sort(key=lambda x:x[1])
                 l.reverse()
             for xx in llparaules:
                 if xx[0]==q:
                     xx[1]=xx[1]+1
```

```
N=0
V1=0
V2=0
for p in l:
    N=N+p[1]

numerador=0
denominador=N*(N-1)
for ert in l: #nombre de vegades que apareix la paraula ert
    numerador= numerador + ert[1]*(ert[1]-1)
D=numerador/denominador

for reret in l:
    if reret[1]==1:
        V1=V1+1
    if reret[1]==2:
        V2=V2+1


V=len(l)
rrr=str(titol)+'Pa'+'.txt'
outfile12=open(rrr,'w')
outfile12.write(NomS)
outfile12.write('\n')
outfile12.write(str(D))
outfile12.write('\n')
outfile12.write(str(N))
outfile12.write('\n')
outfile12.write(str(V))
outfile12.write('\n')
outfile12.write(str(V1))
outfile12.write('\n')
outfile12.write(str(V2))
outfile12.write('\n')

for e in llparaules:
    outfile12.write(str(e[1]))
    outfile12.write('\n')
outfile12.close()
```

```
#Cos del programa
capitols= 80
ppp=1

while ppp<=capitols:
    NomS='HlC'+str(ppp)
    titol='HlC'+str(ppp)+'.txt'
    file=open(titol)
    f=file.read()
    ParaulesFrase(f)
    CaractersFrase(f)
    CaractersParaules(f)
    Lletres(f)
    Paraules(f)
    file.close()
    ppp=ppp+1


lll=([['the',0],['to',0],['and',0],['of',0],
      ['a',0],['he',0],['was',0],
      ['his',0],['in',0],['had',0],['said',0],
      ['it',0],['you',0],['that',0],
      ['i',0],['her',0],['she',0],['at',0],['on',0],
      ['with',0],['as',0],['him',0],['for',0],
      ['for',0],['but',0],['not',0],['they',0],
      ['be',0],['out',0],['were',0],['up',0],['what',0],
      ['all',0],['have',0],['from',0],['been',0],
      ['them',0],['there',0],['into',0],['back',0],
      ['this',0],['so',0],['who',0],['an',0],['could',0],
      ['me',0],['no',0],['about',0],['when',0],['is',0]])


outfile44=open('columnaparaules.txt','w')
for e in lll:
    outfile44.write(str(e[0]))
    outfile44.write('\n')
outfile44.close()
```