CAPÍTOL 2

LITERATURA I ESTADÍSTICA

Índex Capítol

CAPÍTOL 2

LITERATURA I ESTADÍSTICA

L'estadística s'ha revelat com una eina important per a l'estudi d'alguns aspectes del llenguatge, tant parlat com escrit. Per exemple, Kucera (1980) remarca que alguns diccionaris inclouen un llistat de les paraules emprades més freqüentment. De la mateixa manera, els comptatges de les freqüències d'aparició de lletres en els alfabets ha influenciat clarament el desenvolupament dels codis Braille i Morse en els anys 1830, en el disseny del teclat de les màquines d'escriure i els codis per encriptar (i desencriptar) missatges, encara que difícilment podríem considerar que aquests exemples van ser els iniciadors dels estudis estadístics aplicats al llenguatge o a la literatura.

Williams (1970) descriu com els Masoretes, escribes jueus dedicats a la conservació exacta dels texts del Vell Testament entre el 500 i el 1000 dC, varen comptar el nombre de paraules i lletres de cada llibre, així com el nombre de vegades que es repetien determinades paraules d'especial rellevància, habitualment noms. En el segle XIX algunes obres en Anglès van ser estudiades intensament, especialment Shakespeare (Clarke (1845), Fleay (1876), Mendelhall (1887)), comptant les repeticions de paraules i les variacions en la mètrica en poesia. L'objectiu d'aquesta escola d'estudiosos era el de caracteritzar l'estil literari quantitativament, i resoldre problemes de cronologia i autoria usant només mètodes matemàtics elementals.

Entre els pioners dels estudis estadístics aplicats a la literatura trobem Mendenhall (1887, 1901), Zipf (1932), Yule(1944), Guiraud (1959), Morton i McGregor (1964), Mosteller i Wallace (1964, 84), Muller (1967), Brainerd (1970), Hantrais (1976). Aquests autors han emprat mètodes estadístics per datar obres literàries, identificar o discriminar entre autors o gèneres, comparar vocabularis i estructures literàries i fer estudis comparatius de llengües diferents.

En la secció 2.1 es repassen el problemes que aborda l'estilometria, deixant per la secció 2.2 les característiques que han de tenir les unitats d'estadística textual per a quantificar l'estil literari i alguns dels estudis que s'han fet emprant aquestes unitats. A Mosteller i

Wallace (1964, 84), Holmes (1985) i Ginebra i Cabos (1998) es poden trobar extensos resums sobre aplicacions de l'estilometria.

2.1 Estilometria i Autoria

L'anàlisi estadística de l'estil literari és una part de l'estilometria que busca característiques quantificables de l'estil d'un text, i les aprofita per comparar-lo amb l'estil d'altres textos. Quan l'objectiu és determinar l'autoria del text, les característiques escollides han de ser pròpies de l'autor i no del gènere, de l'època en que el text ha sigut escrit o de l'editor que ha ordenat i normativitzat els textos.

L'estilometria sovint estudia aspectes d'estil que l'autor difícilment controla de forma conscient, i per tant aprofita la informació menys evident continguda en els textos.

Quantificar l'estil sense un esforç previ d'un especialista familiaritzat amb els aspectes no quantificables dels textos no porta enlloc i, per tant, les aproximacions a aquests problemes estan necessàriament a cavall de la lingüística, l'anàlisi de textos i del discurs, l'anàlisi de continguts, la informàtica i l'estadística. Amb l'arribada de Internet, dels CD-ROMS i dels scanners, arriba una explosió en l'aplicació de les tècniques estilomètriques que descriurem a continuació que, tal com explica Lebart (1994), són també aplicables a l'anàlisi de preguntes obertes d'enquestes.

Els problemes de determinació de l'autoria d'un text es poden classificar en tres grans famílies:

- En un primer grup de problemes tenim dos (o més) candidats a ser l'autor i es disposa de textos d'aquests dos (o més) autors que són "comparables" amb el text de paternitat discutida. Un exemple que és tractat amb molt detall per Mosteller i Wallace (1984) estudia els *Federalist Papers*, una col·lecció d'articles escrits entre 1787 i 1788 per Hamilton, Jay i Madison amb l'objectiu de recolzar l'aprovació de la constitució americana, i que van ser publicats tots ells sota un sol pseudònim. Dels 77 articles, se'n coneix l'autoria de 65, però n'hi ha dotze que es poden atribuir tant a Madison com a Hamilton. Fent servir altres articles escrits per Madison i Hamilton arriben a la conclusió que els dotze articles disputats s'haurien d'atribuir a Madison.

  Un segon exemple és l'estudi de Kjetsaa (1979) sobre l'autoria real dels quatre volums del *Don de plàcides aigües*, començat a publicar el 1928 com a obra d'en Xolokhov, a partir de la seva comparació amb altres obres d'en Xolokhov i textos d'en Kyriukov, mort de tifus el 1920.

  Aquest primer grup de problemes són els més estructurats i fàcils de tractar; cal assignar el text disputat a un dels dos (o més) autors coneguts fent servir eines estadístiques de classificació i anàlisi discriminant.

- En una segona família de problemes es té un candidat, del que es disposa de textos "comparables" al text en disputa, i una sèrie de candidats alternatius dels que no es disposa de textos reconeguts que puguin servir per fer la comparació. Exemples d'aquest tipus són els estudis de Brinegar (1963) per decidir si les cartes publicades al *New Orleans Daily Crescent* el 1871 sota pseudònim de Quintus Curtius Snodgrass foren escrites per Mark Twain, i les anàlisis començades per Mendelhall

(1887) per investigar si les obres de Shakespeare podrien haver sigut escrites per Bacon. Més recent és l'anàlisi que ha permès destapar que l'autor del llibre *Primary Colours*, sobre la campanya electoral de Bill Clinton, era el periodista de Newsweek, M. Klein. Un exemple de la literatura catalana és el plantejat per en Josep Guia (1995, 96) al voltant del *Tirant lo Blanc* i la hipotètica autoria d'en Joan Roís de Corella.

Aquí el problema estadístic és el de la comparació de mostres de dues o més poblacions, i per tant cal determinar si la variabilitat de les característiques d'estil escollides dins del text disputat i dins dels textos de l'autor conegut és més petita que la variabilitat entre el text disputat i els d'autor conegut, o bé si són comparables.

- Un tercer cas estudia l'homogeneïtat d'estil d'un text, intentant o bé detectar canvis d'autoria o bé modelar l'evolució de l'estil d'un sol autor al llarg del text.

  Dos exemples clàssics són els plantejats per Adams i Rencher (1973) i per Morton (1978) al voltant de la hipotètica existència de més d'un autor del *Llibre d'Isaïes* i de les *Epístoles de Sant Pau*.

  Una caracterització de l'evolució estilística d'un autor al llarg de tota la seva obra pot servir per datar textos, com s'ha fet amb les *Ètiques* d'Aristòtil.

  Aquests tipus de problemes, són els de més difícil aproximació i els menys tractats en la literatura estadística.

  En la literatura catalana, l'exemple més estudiat és el de la possible existència d'una frontera d'estil dins del *Tirant lo Blanc*. Aquest és el problema central que aborda la tesi.

Un primer dilema plantejat en tots els casos és entre basar l'estudi en les formes gràfiques o bé lematitzar prèviament aquestes formes, reduint totes les paraules a vocabulari. Algunes regles de lematització proposades consisteixen en convertir totes les formes verbals a l'infinitiu, tots els substantius al singular, tots els adjectius al masculí singular i les formes elidides a formes sense elidir. Alguns d'aquests criteris tenen més sentit en unes llengües que en d'altres; on en català fem servir *parlo, parles parleu, parlem, parlen*, en anglès només fan servir *speak*, i per llengües amb declinacions, apareixen moltes més formes gràfiques. Mosteller i Wallace (1984) es mostren partidaris de no lematitzar els textos, per raons pragmàtiques ja que per trencar ambigüitats sovint cal fer ús del context. Per textos petits aquesta lematització es pot fer a ma, però fins i tot en aquest cas ens trobem que alguns dels criteris de lematització poden ser discutibles. Lebart (1994) posa per exemple que *defensa de la llibertat* té connotacions diferents de *defensa de les llibertats*. Lebart elabora més sobre aquest dilema, i descriu un estudi fet per Labbe (1990) analitzant els discursos polítics radiotelevisats de Mitterrand, fet tant lematitzant com sense lematitzar. En aquest estudi, s'ha decidit no lematitzar.

Fins ara, pràcticament tota la recerca s'ha fet sobre aplicacions a textos anglesos, alemanys i en menor mesura francesos i clàssics. Adaptar totes aquestes eines a d'altres llengües no és immediat. Un dels objectius d'aquest treball és llistar alguns dels problemes específics pel català, junt amb les solucions que hi hem donat en la nostra l'aproximació a l'estudi del *Tirant lo Blanc*.

En aquesta direcció, l'aparició a través de l'I.E.C. de bancs de dades textuals de la literatura catalana a disposició dels investigadors pot donar una bona empenta a les

aplicacions a la nostra llengua com ho ha fet en el cas del francès la creació del *Tresor General des Langues et Parles Français* que permet accedir en suport informàtic a les obres principals dels segles XIX i XX de la literatura francesa.

2.2 Caracterització de l'estil literari

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

Per emprendre l'estudi estadístic de l'estil literari cal identificar-ne trets que puguin ser quantificables amb precisió i sense ambigüitats. Bailey (1979) llista les propietats generals que haurien de tenir aquestes característiques:

> *"haurien de ser rellevants, estructurals, freqüents, fàcilment quantificables i relativament inmunes al control conscient de l'escriptor."*

Mesurant i controlant aquests trets estilístics s'espera identificar característiques d'un autor o d'un gènere, i distingir diferències genuïnes d'estil de les variacions en l'ús degudes a una tria conscient.

La quantificació de l'estil literari es pot fer en tres nivells, en funció del grau de complexitat a caracteritzar. Un primer nivell compara les freqüències d'ús de unitats lingüístiques fàcils d'identificar i de comptar, que siguin freqüents i difícilment controlables conscientment per l'autor, per exemple la llargada de paraula o frase, o la proporció d'ús d'algunes paraules. Escollir les unitats a estudiar és un problema obert, que convé adaptar a cada autor, gènere i temps. En la tesi s'escolliran a nivell de lèxic, tot i que també es podria haver quantificat a nivell fonètic, morfològic, sintàctic, semàntic o pragmàtic. En els apartats 2.2.1, 2.2.2 i 2.2.3 es fa un repàs d'algunes d'aquestes unitats. En un segon nivell, que obliga a fer un inventari de totes les paraules i a comptabilitzar quantes vegades apareix cadascuna en el text, es pot caracteritzar la riquesa i diversitat del vocabulari. Aquest segon nivell s'estudia en els apartats 2.2.4 i 2.2.5. Un tercer nivell modela l'ordre d'aparició del vocabulari i d'altres unitats lingüístiques, estudiant el llenguatge com a procés estocàstic, i es repassa en l'apartat 2.2.6.

2.2.1 Llargada de paraula i frase

Una característica que es considera pròpia dels autors és l'ús de paraules llargues o curtes. Augustus de Morgan a l'any 1851 ja va considerar el problema de distingir entre diferents autors a través de la llargada de les paraules en llurs texts. La llargada de paraula s'acostuma a mesurar pel seu nombre de lletres.

Mendenhall (1887) va proposar la distribució de les paraules segons la seva llargada com a possible característica distintiva d'autor, i ho va utilitzar per explorar si Bacon o Marlowe podrien haver escrit les obres de Shakespeare. Mendenhall va trobar que Shakespeare usava més paraules de 4 lletres i Bacon més paraules de 3 lletres, tot i que hi havia una semblança gran entre les distribucions de freqüències de les llargades de paraula entre Shakespeare i Marlowe. Williams (1975), repassant l'anàlisi de Mendenhall, va destacar que els texts de Shakespeare i Marlowe eren en vers, mentre que els de Bacon eren en prosa. La diferència entre llargades de paraula pot ser atribuïda a la diferència entre els estils de composició. Les freqüències de llargada de paraula dels treballs de Mill, Dickens i Shakespeare mostren diferències significatives, que poden ser usades per identificar cadascun dels autors.

Brinegar (1963) va recórrer a la longitud de paraula per mostrar que Mark Twain no va escriure "The Quintus Curtius Snodgrass Letters" i Mosteller i Wallace (1984) la utilitzen en l'estudi de l'autoria dels "Federalist Papers" arribant a la conclusió que és poc fiable quan es comparen dos autors perquè les diferències són sovint més grans entre gèneres que entre autors. Ambdós estudis usen el test de Chi-quadrat per a comparar distribucions, una tècnica basada en la hipòtesi –no provada- que els texts d'un autor són mostres aleatòries extretes de la seva pròpia distribució de freqüències fixa de llargades de paraula. El fet que el vocabulari depengui del context implica que, com a molt, això pot ser una aproximació.

Smith (1983) senyala que, quan es comparen texts de gèneres diferents o escrits en èpoques diferents, les diferències en les característiques observades són, de llarg, molt superiors a qualsevol de les que permetin identificar, de forma fiable, autors. És més, comparant texts de diferents autors contemporanis i que pertanyen al mateix gènere literari, les seves distribucions de llargada de paraula són tan semblants que semblen ser escrits pel mateix autor. Arriba a la conclusió que la distribució de les llargades de paraula no hauria de ser utilitzat en estudis d'autoria.

Hilton i Holmes (1993) proven els gràfics *weighted Cusum* amb les paraules curtes (dues i tres lletres), per a comprovar si dos texts poden ser atribuïts al mateix autor, i ho comparen amb els resultats obtinguts en aplicar el test de QSUM. Ho apliquen a tres conjunts de texts: els Federalists Papers, a novel·les escrites per Jane Austen, i a novel·les de James Bond. Tot i que els *weighted Cusum* són marginalment millors que el QSUM, no donen resultats consistentment fiables, i ho atribueixen a que els autors no segueixen patrons de forma suficientment rígida de manera que permeti determinar l'autoria de forma correcta.

La llargada de paraula, doncs, hauria de ser una mesura especialment indicada per a l'estudi de l'homogeneïtat d'estil d'un sol text, o per comparar textos d'un mateix gènere i època.

Fucks (1952) proposa, en una anàlisi sobre autors anglesos i alemanys, mesurar la llargada de paraula pel seu nombre de síl·labes, i calcular el nombre mig de síl·labes per paraula, les freqüències relatives de les paraules de *s* síl·labes i la seva distribució en el text. Conclou que poden ser identificats, de forma quantitativa, trets diferencials entre autors o entre gèneres (poesia o prosa) en base a aquestes característiques. Fucks i Lauter (1965) descobreixen que la distribució de freqüències de síl·labes per paraula discrimina millor entre diferents llengües que entre autors. Bruno (1974) analitza el nombre mig de síl·labes per paraula, a partir de les distribucions de freqüència de les paraules amb *s* síl·labes, en un estudi de l'homogeneïtat d'estil en l'obra èpica de l'edat mitja germànica *Nibelungenlied*. Brainerd (1974) en un estudi de la distribució de

síl·labes per paraula en texts anglesos troba que s'obtenen ajustos millors amb un model basat en la distribució Binomial Negativa traslladada que amb la Poisson traslladada, proposada per Fucks i indica que poden haver-hi grans canvis en la distribució del nombre de síl·labes en canviar d'estil, per exemple del narratiu al diàleg, concloent que els estils d'alguns autors són més homogenis que d'altres pel que fa a les síl·labes. En l'estudi del *Tirant* no s'analitzarà el nombre de síl·labes.

Yule (1938) va suggerir utilitzar la llargada de frase com a mètode per a determinar l'autoria de texts, i la va utilitzar en un estudi sobre l'autoria de *The imitation of Christ*. Va concloure que estadístics basats en la llargada de frase no són fiables en estudis d'aquest tipus, però van aflorar qüestions importants per a l'estilometria relatives a la definició de "frase". Williams (1940) va descobrir que la distribució de freqüències dels logaritmes del nombre de paraules per frase és una aproximació a una llei Normal per a cada autor. El model lognormal proposat per Williams i usat per Wake (1957) ha de ser rebutjat per diversos motius, un dels quals és que la majoria de les distribucions de freqüència de la llargada de frase, després de la transformació logarítmica, son negativament asimètriques. Sichel (1974) proposa una barreja de distribucions de Poisson per a la distribució de llargada de frase. Tant els estudis de Williams com els de Sichel són per l'anglès. Morton (1965) usa la llargada de frase en tests d'autoria per a prosa grega, i Kjetsaa (1979) ho aplica a l'autoria del *Don de plàcides aigües*. Mosteller i Wallace (1984) troben que per als *Federalist Papers* la llargada de frase no és un bon discriminant, doncs dels articles d'autoria coneguda es desprenia que la mitjana de paraules per frase era extraordinàriament semblant per els dos autors en disputa: 34,55 per Hamilton i 34,59 per Madison, i que la variabilitat és massa gran per trobar-hi diferències significatives: la desviació tipus de les llargades mitges és 19,2 per Hamilton i 20,3 per Madison.

La llargada s'acostuma a mesurar en nombre de paraules per frase, i és habitual considerar com a frase tot el que acaba en un punt, un signe d'interrogació o bé un signe d'exclamació. És important notar que aquesta variable depèn de la puntuació i, per tant, queda sota el control conscient de l'autor. Per tant, té sentit emprar-la sobre tot quan els texts preserven la puntuació de l'autor o, en el seu defecte, tots han sigut editats per la mateixa persona. Aquesta és una qüestió molt important en el cas dels texts medievals, atès que la puntuació no s'introdueix fins més tard, i és obra de l'editor.

Ginebra i Cabos (1998) fan el primer anàlisi estilomètric del *Tirant*. Parteixen d'una base de dades limitada, composada per *10* trossos d'aproximadament *4000* paraules consecutives, presos a l'inici i al final de cadascuna de les 5 parts en que es composa l'edició de Riquer (1983), atès que en aquell moment no es disposava d'una versió digitalitzada del llibre. Analitzen l'homogeneïtat d'estil en el *Tirant lo Blanc* a partir de la llargada de paraula, trobant que les paraules més llargues són més abundants a la zona final que a la resta del llibre, denotant una possible doble autoria. Mitjançant l'anàlisi de la distribució de les paraules i de la distribució de les frases segons la seva llargada aprecien diferències entre els *10* trossos i la *Història de Josef*, obra en prosa religiosa de Roís de Corella. En un estudi en el que es disposa d'una versió completa del *Tirant*, Riba i Ginebra (2001) detecten un punt de canvi en la distribució de les llargades de paraula en el capítol 371.

En els capítol 8, 9 i 10 de la tesi s'estudiarà l'homogeneïtat d'estil en el *Tirant* fent servir la llargada de paraula, de frase i de capítol, respectivament.

2.2.2 Vocabulari i paraules eina

Sovint podem reconèixer el vocabulari d'un autor identificant la freqüència amb la que fa servir unes paraules determinades.

L'ús de paraules ofereix grans possibilitats per a la discriminació entre estils. La proporció d'ús d'algunes paraules varia molt en diferents obres de l'autor, mentre que d'altres presenten molta estabilitat en totes les obres del mateix autor, degut sobre al fet que hi ha paraules que depenen fortament del context o de l'argument mentre que altres en són força independents. A l'hora de discriminar entre autors necessitem aquest darrer tipus de paraules, aquelles que la seva presència sigui tan independent com sigui possible del context, per no confondre l'efecte de l'autor amb l'efecte del context. Aquestes paraules s'anomenen en anglès *function words,* en francès *mots outil* mentre que en català ho traduïm per *paraules eina*. Lebart (1994) proposa emprar com a paraules eina les conjuncions, i Mosteller i Wallace (1984) hi afegeixen els articles i de vegades els pronoms. També escullen paraules eina d'entre les més freqüents.

En el problema de l'autoria dels *Federalist Papers,* tractat per Mosteller i Wallace, els dos autors candidats són ben coneguts, així poden triar les paraules eina entre les que discriminen millor entre els estils de Hamilton i Madison, a partir de texts dels que se sap qui és l'autor. Van descobrir, per exemple, que Hamilton tendia a usar *while* mentre que Madison preferia *whilst*, i ho van aprofitar per atribuir l'autoria dels texts disputats. També van notar que Hamilton va emprar la paraula *upon* prop de 18 vegades més freqüentment que Madison. Mosteller i Wallace van seguir quatre mètodes diferents en el seu estudi basats en la freqüència d'ús de paraules. El principal té enfocament Bayesià, el segon es basa en l'anàlisi discriminant clàssic, el tercer és una anàlisi bayesiana robusta mentre que el quart consisteix en un estudi simplificat de les freqüències emprant mètodes clàssics.

Ellegard (1962) proposa, en el seu estudi de l'autoria de *The Junius Letters*, comparar la freqüència d'ús de una paraula en el text disputat amb la freqüència d'ús d'aquesta paraula en un conjunt molt més gran de texts de referència. Va obtenir una llista ordenada de paraules correlacionades positivament o negativament amb el text en estudi, en el sentit de que eren usades més o menys freqüentment que pels seus contemporanis.

Morton (1978) desenvolupa tècniques per a estudiar la proporció d'aparicions d'una paraula eina seguida o precedida per una altra paraula determinada, respecte al nombre d'aparicions de la paraula eina, com per exemple *and the* respecte de *and*, *to be* respecte de *to* o bé *not only* respecte de *not*. Això ho aprofita per a estudiar si els atacs que va patir Sir Walter Scott van afectar el seu estil. De totes maneres, els mètodes proposats per Morton han estat molt criticats, Smith (1985) demostra que no és capaç de distingir entre els treballs d'autors de teatre Elisabetians i Jacobins.

Burrows (1987) fa un estudi detallat i exhaustiu de les sis novel·les publicades de Jane Austen, emprant les 30 paraules més freqüents sense distingir entre paraules eina i paraules lligades al context. Mostra com dintre de cadascuna de les tres divisions formals – narrativa pura, narrativa de personatge, diàleg – les sis novel·les mostren patrons semblants. Aquest patró es veu alterat quan es comparen texts de diferents categories. Burrows també va trobar diferències entre els estils narratius de Jane Austen en diferents etapes de la seva carrera. Burrows i Hassall (1988) usen l'Anàlisi en Components Principals per distingir entre Henry i Sarah Fielding, calculant el rati d'ocurrència de les cinquanta paraules més freqüents en els texts d'autoria disputada i

en treballs per els que no hi ha dubte de qui dels dos és l'autor. Representant les dades en l'espai generat per les dues primers Components Principals apareix molt clarament a qui cal atribuir els texts en disputa. Més endavant Burrows (1992) dóna més exemples d'aquest tipus d'anàlisi en un conjunt divers d'autors i gèneres.

El concepte de paraula eina es pot generalitzar al de *locució eina* que podria incloure locucions prepositives com a *causa de*, *enmig de* o *per tal de*, locucions conjuntives com *tot seguit*, *en tant que* o expressions com *seguretat social, nivell de vida* o *fer esport*. A aquestes unitats Lebart (1994) les anomena *segments repetits*. Miller, Newman i Friedman (1958) donen una llista de 363 paraules eina en anglès, però no coneixem cap compilació semblant pel català.

Ginebra i Cabos (1998), fent servir l'anàlisi de correspondències al *Tirant lo Blanc*, donen una llista de *44* paraules eina en el *Tirant*, formada per conjuncions, preposicions, articles i alguns elements de locucions prepositives o conjuntives. Troben que la freqüència d'aparició d'algunes d'elles és significativament més alta al principi que al final del llibre, mentre que per altres paraules passa el contrari, confirmant la manca d'homogeneïtat en el llibre detectada en l'estudi de la llargada de paraula. També troben diferències en lús d'algunes d'aquestes paraules entre els 10 trossos que analitzen del *Tirant* i la *Història de Josef* de Corella.

En el capítol 11 de la tesi s'estudia l'homogeneïtat d'estil en el *Tirant* mitjançant l'anàlisi de les 25 paraules més freqüents no sensibles al context, i es llisten les paraules, entre les 100 més freqüents, que discriminen entre els dos estils que es troben.

2.2.3 Distribució de les parts de l'oració

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

Antosch (1969) estudia el comportament del rati verb-adjectiu per varis gèneres literaris i mostra com el rati depèn en gran mesura del tema de l'escrit. Per exemple en contes populars el rati pren valors elevats mentre que en treballs científics pren valors baixos.

Brainerd (1974) presenta un estudi meticulós sobre si la freqüència d'ús d'articles pot ser considerada com a indicador d'estil. Conclou que el nombre de vegades que apareix un article no és específic d'un autor, però sí que apareixen diferències en diferents tipus d'escrits. Els articles periodístics, la narrativa i l'exposició formal tendeixen a fer servir, en general, més articles que les novel·les, la correspondència i les autobiografies. També troba relació entre el nombre d'articles i el nombre de pronoms en un text.

En un altre article, Brainerd (1973) utilitza els articles i els pronoms per intentar distingir entre novel·la i el gènere romàntic, pel fet que aquestes parts de l'oració són especialment sensibles a canvis en el nivell de formalitat de l'escrit. Va trobar que les novel·les tendeixen a tenir freqüències més elevades de pronoms i més baixes d'articles, però també va trobar una manca de homogeneïtat entre autors de gènere romàntic.

En la tesi no s'estudia la distribució de les parts de l'oració.

2.2.4 Diversitat i Riquesa de Vocabulari

A l'hora de comparar textos, on hi ha més informació és a nivell de lèxic, perquè és on hi ha més dades. Una de les famílies de mesures més importants a l'hora de quantificar l'estil són les que intenten caracteritzar la riquesa o la diversitat del vocabulari d'un autor.

Es parteix de la hipòtesi que l'autor disposa de un vocabulari, composat per una llista de paraules que pot emprar quan escriu, de les quals tendeix a utilitzar unes més que d'altres. Els textos de l'autor representen mostres d'aquest vocabulari teòric, i per tant la freqüència d'aparició de les paraules en ells reflexa les característiques del vocabulari de l'autor. L'objectiu és el de mesurar la quantitat de paraules de què disposa l'autor a partir de les freqüències d'aparició de totes les paraules en el text per caracteritzar la diversitat, que es podrà emprar a l'hora de fer comparacions entre texts o autors.

El concepte de diversitat s'ha estudiat en altres camps de la ciència. Té l'origen en l'ecologia i l'estudi de la biodiversitat, i ha estat aplicat també, entre d'altres contextos científics, a la genètica, a l'anàlisi de la concentració industrial, a l'economia, a les desigualtats econòmiques i a la diversitat lingüística. La idea comú és la de definir un cert nombre de categories i d'assignar a cadascuna una proporció del total d'una determinada quantitat. Aquesta quantitat pot ser en forma de recursos, inversions, temps, energia, abundància i més (Patil i Taillie, 1982).

Un índex de diversitat és una mesura de "dispersió qualitativa" d'una població de individus que pertanyen a vàries categories qualitativament diferents. De la mateixa manera que estadístics com la variança, la desviació tipus o el rang mesuren variabilitat d'una variable aleatòria discreta o contínua, els índexs de diversitat mesuren variabilitat en variables aleatòries categòriques (Pielou, 1982).

Es diu que un estil literari (o una comunitat biològica) té gran diversitat si empra moltes paraules (hi ha moltes espècies) diferents, i per a un nombre donat de paraules (espècies) diferents trobades, la diversitat serà més gran quant més semblants siguin llurs abundàncies relatives.

Definim N com el nombre total de paraules que apareixen en un text, i que en direm nombre d'ocurrències i V com el nombre de paraules diferents que formen el vocabulari observat en el text, i en direm formes. També definim V_d com el nombre de paraules que apareixen d vegades en el text, de manera que tenim:

$$N = \sum_{d=1}^{\infty} d \cdot V_d ,$$

$$V = \sum_{d=1}^{\infty} V_d .$$

El quocient $\hat{p}_d = \frac{V_d}{V}$ és la proporció de paraules diferents que apareixen d vegades en el text i, en, quan el nombre de textos de llargada N creix, el límit del promig d'aquestes proporcions el denotarem com p_d, tenint que $p_d = E(V_d/V)$. La suma de les p_d per totes les d val *1*. La majoria de les mesures de diversitat del vocabulari es basen en $(p_1, p_2, .., p_d, ..)$.

Donat un text de llargada N, com més gran és V més ric i divers és el llenguatge; mentre que per un text amb N i V donades, com més gran p_d per d petites i més petit p_d per d

grans, més divers és el vocabulari. Qualsevol índex de diversitat ha de respectar aquests dos nivells d'ordenació.

Sichel(1986) i Yule(1944) plantegen el debat sobre si els estudis de riquesa del vocabulari d'un autor o text s'han de fer per paraules de tota mena o bé si és millor restringir-se a paraules d'alguna classe especial com substantius, adverbis, verbs o preposicions i conjuncions. El present treball s'ha realitzat sense desagregar el vocabulari segons les funcions de les paraules.

Riba i Ginebra (2000a,b) analitzen l'evolució dels índexs de diversitat per estudiar l'homogeneïtat d'estil en el *Tirant*, detectant un canvi important en el capítol 383 i un de menys evident en el capítol 119.

En el capítol *13* es fa un repàs als índexs de diversitat emprats en estudis estilomètrics i es llisten referències que les utilitzen. Part d'aquesta literatura està escrita pensant en la caracterització de la diversitat biològica d'un ecosistema, on el nombre d'espècies correspon al nombre de paraules o formes diferents i l'àrea de l'ecosistema a la llargada del text. En el mateix capítol es faran servir aquests índexs per analitzar l'evolució de l'estil en el *Tirant lo Blanc*.

2.2.5 Distribucions de Vocabulari

Per cada paraula *i* de les que composen el vocabulari teòric d'un autor, existeix un valor π_i que és la proporció de vegades que apareixeria la paraula *i* en un text del mateix autor de llargada infinita. Implícitament es fa la hipòtesi que tant la llista de paraules com els valors π_i són constants al llarg de l'obra de l'autor en estudi, suposició que pot ser discutible quan els textos de l'autor comprenen èpoques o gèneres molt diferents. Denotem per n_i el nombre de vegades que apareix la paraula *i* en el text, i $\hat{\pi}_i = n_i/N$ la proporció de vegades que hi apareix; si la hipòtesi d'estacionarietat és vàlida, en augmentar la llargada *N*, $\hat{\pi}_i$ tendeix a π_i. Per a un text, la suma de les n_i val N i a suma de les π_i per a totes les *i* que composen el vocabulari teòric de l'autor val 1:

$$N = \sum_{i=1}^{V} n_i,$$

$$\sum_{i=1}^{V} \pi_i = 1.$$

La seqüència de valors π_i està íntimament relacionada amb la seqüència de valors p_d definits en l'apartat 2.2.4, emprant-se aquesta darrera per a caracteritzar la distribució del vocabulari de l'autor, atès que, en aquest cas, es pot modelar el vocabulari amb menys paràmetres que a través de π_i, perquè el suport de la variable aleatòria n_i és tot el vocabulari, tant l'observat com el no observat, i per tant és desconegut, mentre que el suport de V_d són els números enters. A més, p_d depèn de *N*, mentre π_i no. Els valors de p_d i de π_i s'estimen a partir de les seqüències $\hat{p}_d$ i $\hat{\pi}_i$ que observem en els seus textos. La majoria de les mesures de diversitat del vocabulari es basen en p_d.

Les distribucions de vocabulari s'obtenen de comptar quantes paraules apareixen una vegada en el text, quantes apareixen dues vegades, quantes tres i així en endavant. L'avantatge d'usar-les és que les mesures són independent del context i, per tant, poden ser aplicades a treballs d'arguments completament diferents. La informació que s'obté

revela trets característics sobre l'actitud d'un autor envers l'ús de paraules rares i un vocabulari diversificat. Quan es considera la llista d'unitats lexicals que apareixen en un text de llargada donada, es pot comprovar, per exemple, que el nombre de paraules que han sigut usades una sola vegada (*hapax legomena*) és més gran que la de les paraules emprades dues vegades (*hapax dislegomena*), i que aquestes són més nombroses que les que apareixen tres vegades i així endavant. Hi ha hagut alguns intents per obtenir models matemàtics que expliquin aquest decreixement, per exemple Sichel (1986).

En la tesi no s'ha estudiat la distribució del vocabulari.

2.2.6 Dependència del llenguatge

Podem tractar el llenguatge com una seqüència de paraules i, com a tal, modelar-ne la dependència.

El primer estudi estadístic de la dependència en el llenguatge en el context literari es deu a Markov (1913), on analitza la seqüència de vocals i consonants en el poema "Evgeni Onegin" de Pushkin. Va ser d'aquest estudi que va sortir la definició de la Cadena de Markov simple, en la que una lletra pot prendre dos estats: *vocal* i *consonant*. Una característica bàsica en les cadenes de Markov és la matriu de probabilitats de transició entre els estats:

$$P = \begin{pmatrix} p_{cc} & p_{cv} \\ p_{vc} & p_{vv} \end{pmatrix}$$

on:

$$P_{cc}=Pr\{\text{una consonant sigui seguida per una consonant}\},$$
$$P_{cv}=Pr\{\text{una consonant sigui seguida per una vocal}\},$$
$$P_{vc}=Pr\{\text{una vocal sigui seguida per una consonant }\} \quad \text{i}$$
$$P_{vv}=Pr\{\text{una vocal sigui seguida per una vocal }\}.$$

En qualsevol llenguatge es pot estudiar si alguns fonemes o seqüències de paraules en frases formen Cadenes de Markov (Mandelbrot (1961)). De fet, pot demostrar-se que un llenguatge no pot ser un procés de Markov d'ordre finit, tot i que Good (1969) mostra com es poden obtenir bones aproximacions a cadenes de Markov de tercer ordre per a l'Anglès.

Una tècnica, basada en les cadenes de Markov, per a modelar la dependència en forma de sèrie temporal discreta consisteix en l'ajust de Models de Markov Amagats (HMM en endavant, de la terminologia anglesa *Hidden Markov Models*). En els HMM s'assumeix que cada element de la sèrie prové d'una entre *m* distribucions discretes, de les quals es coneix el nombre de paràmetres que les descriuen, que però són desconeguts, i on la regla per decidir quina distribució és activa en un punt donat és una cadena de Markov no observada. Normalment es trien models amb un nombre d'estats, *m*, petit, i com a distribució discreta la Poisson, la binomial o la multinomial. El mètode, però, pot ser generalitzat de forma molt senzilla per a altres distribucions.

Com a exemple, es pot considerar la sèrie de llargades de paraula $\{S_t: t \in N\}$, on s'assumeix que S_t està generada per una entre *m* diferents distribucions de Poisson truncades, de manera que no es pugui donar el valor *0*. S'assumeix també que, en un

punt *t* donat, es tria la distribució mitjançant una cadena de Markov no observada de *m* estats,

$$C^{(T)} = \{C_t: t=1,....,T\},$$

de manera que, condicionat a $C^{(T)}$, les variables aleatòries $\{S_t: \quad t=1,....,T\}$ són mútuament independents i on:

$$P(S_t = s \mid C_t = i) = \frac{\lambda_i^s e^{-\lambda_i} / s}{1 - e^{-\lambda_i}}$$

és la probabilitat que la i-èssima Poisson truncada doni *s*.

Ajustar aquests models implica estimar els paràmetres de les *m* distribucions i la matriu de transició de la cadena de Markov no observada. Per a més detalls sobre els HMM es pot veure MacDonald i Zucchini (1997).

Riba i Ginebra (1998) modelen, fent servir HMM, les seqüències de llargades de paraula per *10* trossos d'aproximadament *4000* paraules consecutives, presos a l'inici i al final de cadascuna de les 5 parts en que es composa l'edició del *Tirant* de Riquer (1983), per *m=2*. És a dir, assumeixen que la seqüència de llargades de paraula és una mescla de paraules més llargues i més curtes de la mitjana, mesclades segons la matriu de transició estimada. L'anàlisi dels 10 models ajustats mostra una frontera d'estil en la quarta part.

2.2.7 Tècniques Multivariants i de Classificació

Els problemes d'autoria normalment necessiten la selecció de múltiples criteris per tal de distingir trets característics d'un autor. Kjetsaa (1979) va usar una combinació de variables estilístiques, algunes de les quals eren especialment indicades per a texts en rus, en el seu estudi del *Don de plàcides aigües*. Les variables que va seleccionar representaven un gran ventall de categories sintàctiques i de vocabulari. En un article posterior, Kjetsaa (1981), utilitza 15 paràmetres per a detectar els trets característics de l'estil de Dostoyevsky.

En els darrers anys l'ús de tècniques multivariants com l'Anàlisi en Components Principals, l'Anàlisi Factorial, Anàlisi Discriminant o Anàlisi Cluster han sigut utilitzades en l'estudi de l'estil literari. Binongo (1994) aplica l'Anàlisi en Components Principals a l'estudi de l'estil de l'escriptor filipí Nick Joaquin, Miles i Selvin (1966) apliquen Anàlisi Factorial al vocabulari dels poetes del segle XVII i conclouen que és possible determinar trets característics de la influència tant qualitativa com quantitativa de Petrarca, dels clàssics, i de la Bíblia en els poetes d'aquella època. Somers (1966) va aplicar de forma efectiva l'Anàlisi Discriminant als treballs de *Philo Alexandrinus* i a la col·lecció de les epístoles de San Pau per provar la hipòtesi de que *Philo* va exercir alguna influència sobre la *Carta als Hebreus*. Bruno (1974) va aplicar una Anàlisi Discriminant pas a pas per trobar una equació discriminant que permetés distingir entre dinou "high-formulaic" i vint-i-dues "low-formulaic stanzas" escollides del *Nibelungenlied*, i podem trobar exemples d'Anàlisi Cluster aplicats a l'estudi de l'autoria de texts a Bailey (1979) i Boreland i Galloway (1980). Holmes (1992) aplica Anàlisi Cluster jeràrquic i Anàlisi en Components Principals per a mesurar la diversitat del llenguatge com a eina per a l'atribució d'autoria, i ho aplica al corpus de l'escriptura mormona, i Holmes i Forsyth (1995) apliquen les mateixes tècniques als *Federalist*

Papers emprant mesures de diversitat del llenguatge i una llista de paraules eina, atribuint tots els articles en disputa a Madison.

La estilometria és, essencialment, un problema de reconeixement de patrons (*pattern recognition* en llenguatge estadístic). Les xarxes neuronals i els algoritmes genètics tenen l'habilitat de reconèixer l'organització que està a la base de les dades, el que és de vital importància per a qualsevol problema de *pattern recognition* i, per tant, la seva aplicació a l'estilometria hauria de ser útil. El model hauria de ser entrenat prèviament sobre dades provinents de texts d'autoria coneguda, per tal que pugui aprendre a distingir entre, per exemple, dos autors candidats abans d'intentar classificar un text d'autoria desconeguda. Un aspecte important és la tria de variables d'entrada; típicament s'han emprat freqüències d'ús de paraules eina o ratis de combinacions d'ocurrències de paraules.

Matthews i Merriam (1993) van usar una xarxa neuronal per a investigar qüestions relacionades amb treballs de Shakespeare i Fletcher. Van usar dos conjunts de cinc discriminadors cadascun com a variables d'entrada – per exemple el rati *upon*/(*on*+*upon*)- i la xarxa va donar excel·lents resultats en la classificació. En un treball posterior Merriam i Matthews (1994) varen entrenar la xarxa en treballs coneguts de Shakespeare i Marlowe abans de usar-la per a classificar texts d'autoria disputada, obtenint de nou resultats força fiables. Lowe i Matthews (1995) van emprar una arquitectura alternativa en treballs de Shakespeari i Fletcher que va revelar-se com una millora sobre les arquitectures més estàndard. Holmes i Forsyth (1995) apliquen un algoritme genètic al *Federalist Papers*, amb uns resultats que ells consideren encoratjadors.

En els capítols 8, 9, 11 i 12 es fan servir tècniques cluster per agrupar els capítols del *Tirant* en conjunts homogenis, en base a la distribució de les llargades de paraula, de frase, de l'ús de paraules eina i dels índexs de diversitat, respectivament.