UNIVERSITAT DE BARCELONA

Grau de Lingüística

Treball de Fi de Grau

Curs 2022-2023

Elaboració de perfils lingüístics en el marc de l'atribució d'autoria

NOM DE L'ESTUDIANT: Elsa Ubieto Soto

NOM DEL TUTOR: Mireia Farrús

Barcelona, juny 2023

UNIVERSITAT DE BARCELONA

Coordinació d'Estudis
Facultat de Filologia

Gran Via
de les Corts Catalanes, 585
08007 Barcelona

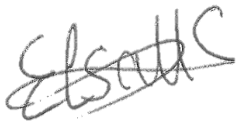
Tel. +34 934 035 594
fil-coord@ub.edu
www.ub.edu

DECLARACIÓ D'AUTORIA

Amb aquest escrit declaro que soc l'autora original d'aquest treball i que no he emprat per a la seva elaboració cap altra font, incloses fonts d'Internet i altres mitjans electrònics, a part de les indicades. En el treball he assenyalat com a tals totes les citacions, literals o de contingut, que procedeixen d'altres obres. Tinc coneixement que d'altra manera, i segons el que s'indica a l'article 18, del capítol 5 de les Normes reguladores de l'avaluació i de la qualificació dels aprenentatges de la UB, l'avaluació comporta la qualificació de "Suspens".

Barcelona, a 18 de juny de 2023

Signatura:



AGRAIMENTS

En primer lloc, vull donar les gràcies a la Mireia Farrús, la meva tutora, per guiar-me en l'elaboració d'aquest treball i ajudar-me a trobar solucions cada vegada que sorgien obstacles pel camí. També agrair als meus amics, companys, i familiars, que m'han donat suport al llarg d'aquest procés. L'interès que hi han mostrat, les paraules d'ànim, i els oferiments desinteressats d'ajudar-me en el que fos possible m'han acompanyat i donat força en tot moment, i per això son una part essencial d'aquest treball.

RESUM

En el següent estudi hem plantejat i documentat la realització d'una tasca d'extracció de característiques lingüístiques de cinc autors d'un mateix període històric, amb l'ajuda d'eines de processament de text. Aquest procés té la finalitat de configurar els perfils lingüístics de cada autor, de cara a aplicacions automatitzades en el marc de la identificació i atribució d'autoria. S'ha avaluat l'eficiència de les dades obtingudes amb diferents classificadors basats en estadística que es fan servir en models d'aprenentatge automàtic. Els resultats obtinguts presenten un gran percentatge d'encert en la identificació automàtica dels autors de les obres.

Paraules clau: Creació de perfils lingüístics, atribució d'autoria, extracció de característiques, classificació automàtica

ABSTRACT

In the following study we have proposed and documented the execution of a task involving the extraction of linguistic features from five different authors of the same historical period, by means of text processing tools. This process aims to configure the linguistic profile of each author for automatic applications in the framework of authorship identification and attribution. The efficiency of the obtained data has been evaluated using different classifiers based on statistic algorithms which are usually employed in machine learning models. The obtained results show a high percentage of success in the automatic identification of the authors and their works.

Keywords: author profiling, authorship attribution, feature extraction, automatic classification

ÍNDEX

1. INTRODUCCIÓ

Els constants avenços tecnològics que presenciem dia a dia van fent-se lloc en la nostra societat amb gran velocitat, amb la influència i incorporació de noves eines en les nostres tasques i labors quotidianes. En relació al camp de la lingüística, una de les àrees més populars i amb més futur és la lingüística computacional, amb branques com el Processament del llenguatge natural, que permet la creació i ús de sistemes de traducció, assistents virtuals, correctors automàtics, etc. Però la difusió d'aquests grans desenvolupaments també comporta grans riscos per als usuaris. Cada cop és més fàcil ser víctima d'usos malintencionats d'aquestes tecnologies, i és en aquests casos on disciplines com l'anàlisi d'autoria hi juguen un paper essencial. L'anàlisi d'autoria col·labora a detectar i combatre situacions com el plagi acadèmic, el frau, o fins i tot accions criminals que es perpetren a través de produccions textuals.

En aquest Treball de Fi de Grau treballarem l'anàlisi i atribució d'autoria mitjançant la creació de perfils lingüístics de diferents candidats, amb l'objectiu de classificar les seves obres automàticament. El treball s'estructura de la següent manera: En la següent secció tractarem el marc teòric de la disciplina. En la secció 3 plantejarem els objectius de la recerca. En la secció 4 descriurem la metodologia emprada. En les seccions 5 i 6 veurem i analitzarem els resultats obtinguts i la seva significació en la tasca proposada. Finalment, en la secció 7 extraurem conclusions de la recerca. En les seccions 8 i 9 trobarem la bibliografia consultada i els annexos.

2. MARC TEÒRIC

L'anàlisi d'autoria és un procés que consisteix en examinar les característiques d'una mostra de text per tal d'extreure conclusions sobre l'autoria del mateix (Zheng, R. et al., 2006). Es considera que deriva de l'estilometria, i està estretament relacionat amb àmbits com el processament del llenguatge natural (PNL) o l'aprenentatge automàtic. Actualment té gran varietat d'aplicacions, sovint en camps com la lingüística forense, seguretat informàtica, o en recerca. Dins l'anàlisi d'autoria es distingeixen camps que s'ocupen de diferents tasques per tal de completar el procés. Les dues principals àrees que tractarem en aquest estudi son la creació automàtica de perfils i l'atribució d'autoria.

La creació automàtica de perfils, també anomenada caracterització de perfils o *author profiling* és una disciplina que busca crear el perfil d'un autor a partir de la suma de

característiques lingüístiques que presenten les seves produccions escrites. Alguns dels trets de l'autor que poden configurar aquests perfils inclouen el sexe, edat, context socioeconòmic, orígens geogràfics, llengües d'ús habitual, etc. Els seus orígens es troben en investigacions d'atribució d'autoria, ja que la elaboració de perfils dels autors permetia discriminar les seves produccions (que en molts casos presentaven una autoria qüestionada) a partir de característiques que distingien els uns dels altres. Es tracta d'una disciplina essencial en la lingüística forense, ja que la creació d'un perfil lingüístic pot permetre la identificació de sospitosos, a més de les evidències en casos de detecció de plagi. A més a més, les aplicacions d'aquests tipus d'eines continuen expandint-se fins àmbits com els de la seguretat o el màrqueting, on hi ha un constant creixement en la demanda de tecnologies que identifiquin els usuaris o potencials clients de determinats serveis i productes.

L'atribució d'autoria, com ja hem introduït, treballa estretament amb la creació automàtica de perfils, i consisteix en determinar la possibilitat d'atribuir una producció textual a un autor concret basant-se en les altres produccions d'aquest mateix autor. Habitualment es dona en casos on hi ha diversos candidats entre els quals s'hi troba l'autor, o però pot ser també que l'autor no es trobi entre els candidats (Stamatatos, E. et al., 2018). Els precedents més destacats d'aquest camp estan datats entre els segles XVIII i XX, el més recent sent el cas de l'autoria dels Federalist Papers de Madison (1964). Ambdues àrees interactuen constantment, ja que complementen les seves respectives tasques (la atribució d'autoria requereix perfils dels autors).

Segons Zheng, R. et al. (2006) el procés a seguir en la identificació o atribució d'autoria consta de quatre fases principals: En primer lloc, configurar el corpus o col·lecció de textos escrits pels autors candidats, amb la finalitat de perfilar els seus respectius estils. En segon lloc, determinar les característiques que n'extraurem, les quals prediem ens ajudaran a discriminar les produccions de cada autor. En tercer lloc, generar els models de classificació a partir de les dades obtingudes. S'esmenta la divisió del corpus entre una secció destinada a l'entrenament del model i una altra a la seva avaluació (*training* i *testing*). Finalment, l'última fase consisteix a provar aquest model amb textos d'autoria desconeguda. En el present estudi només es duran a terme les tres primeres fases, ja que les opcions d'autoria son tancades.

Reprenent l'esmentat anteriorment, en la tasca d'atribució d'autoria cal fer una selecció de les característiques o marcadors en què basarem els perfils. Zhang, C. et al. (2014)

destaquen aquesta tria com una de les qüestions més determinants del procés, ja que en elles s'ha de reflectir l'estil de cada autor. Les descriuen com a marcadors d'estil que han de ser "objectius, quantificables, independents del context i inambigus". També hi reflexionen Grant, T. i Baker, K. (2001), afirmant el risc de la generalització a l'hora de seleccionar aquestes característiques, en tant que cal validar la seva fiabilitat en mostres nombroses.

Pel que fa a l'enfocament més automatitzat d'aquesta disciplina, es fan servir tècniques d'aprenentatge automàtic que permeten extreure els patrons i característiques a tenir en compte a partir dels corpus d'entrenament. La selecció dels mètodes i algorismes de classificació és un altre punt a considerar. Un dels classificadors principals utilitzats en tasques d'identificació d'autoria son les màquines de vectors de suport (SVM) (Zhang, C. et al., 2014). Altres mètodes populars son l'anàlisi discriminant lineal (LDA), els arbres de decisió, les xarxes neuronals o algorismes genètics.

3. OBJECTIUS

L'objectiu principal d'aquest Treball de Final de Grau és dur a terme una tasca d'elaboració de perfils lingüístics de diferents autors. Per a fer-ho, en primer lloc seleccionarem els autors que conformaran el nostre corpus, junt amb les seves respectives obres. Seguidament, analitzarem aquestes obres extraient característiques lingüístiques amb l'ajuda d'eines de processament de text. Els resultats d'aquests trets configuraran els perfils lingüístics dels autors en forma de dades quantitatives i mesurables. A partir d'aquestes dades podrem realitzar càlculs estadístics que ens permetran establir la rellevància d'aquests trets de cara a la distinció de la producció dels diferents autors. Així doncs, intentarem comprovar l'eficiència de les característiques extretes per a la tasca plantejada en el nostre corpus.

4. METODOLOGIA

4.1. Selecció del corpus

El procés de selecció dels autors i dels llibres utilitzats per a aquest corpus ha estat realitzat a partir dels següents criteris: En primer lloc, destacar que, al tractar-se d'un projecte que es focalitza en perfils lingüístics, hem tractat de minimitzar tots els factors

externs que poguessin manifestar-se com a diferències en les seves produccions ja d'entrada. És a dir, s'ha intentat escollir autors que inicialment no personifiquin grans diferències de caire extralingüístic que puguin projectar-se en la seva forma d'escriure (tals com la seva època, edat, temàtica, etc), evitant així un possible biaix en els resultats. És per això que s'ha considerat adient seleccionar un corpus monolingüe on tots els autors (i llibres triats) pertanyessin a una mateixa època, idealment inclús formant part d'una mateixa tendència o moviment literari.

Per tal de dur a terme aquesta selecció hem consultat el repositori web de The Project Guttenberg[1]. Es tracta d'un projecte de biblioteca online d'accés obert que reuneix més de 70.000 llibres en format electrònic, pertanyents a milers d'autors i en nombroses llengües, amb la finalitat de posar-los a disposició de tothom per tal de promocionar la creació i distribució de literatura en aquest format. És en aquest repositori online on hem obtingut els arxius que han configurat el nostre corpus.

En aquest punt del procés de selecció, cal dir que hi ha jugat un paper important la disponibilitat dels títols i autors en les llengües en què ens plantejàvem treballar. En altres paraules, la selecció definitiva del corpus ha estat determinada, en part, per conveniència. S'ha triat un corpus format per llibres en castellà, donat que el nombre d'autors i autores, i els seus respectius títols disponibles eren molt més nombrosos que no pas els que haguéssim pogut trobar en català. També hem establert altres criteris de cara a la tria definitiva dels llibres, de nou**,** guiats per la premissa de facilitar, dins de les possibilitats, la tasca proposada. Pel que fa a l'època, hem cregut convenient no allunyar-nos en excés de la forma que pren la llengua en el present, de forma que no hem volgut considerar obres datades en segles molt anteriors. Pel que fa al gènere literari, també hem trobat adequat descartar obres de poesia i de teatre, ja que no sembla que siguin adients per al tipus de recerca que volem dur a terme. El primer, perquè és un estil que altera la forma d'expressió, o que inclús pot resultar ornamental, i el segon perquè l'estructura del gènere també limita en certa forma l'obra. Veient la direcció que volem seguir en aquest treball, hem cregut adequat focalitzar-nos en la literatura narrativa, ja que creiem que és el gènere on millor és poden reflectir les característiques lingüístiques de cada autor. A més a més, hem volgut assegurar-nos de la presència tant d'autors com d'autores en el corpus. Aquest

[1] *Project Gutenberg*. https://www.gutenberg.org/

factor és el que ha resultat determinant a l'hora de reduir les opcions de cerca perquè, com es podria intuir, la presència d'autores és minúscula en comparació amb la d'autors homes.

En el repositori de The Project Guttenberg en llengua castellana hi havia a penes mitja dotzena d'autores, i d'aquestes, només dues que tinguessin varies obres disponibles. Per fortuna, les dues autores formaven part del mateix moviment literari datat del segle XIX, i entre les seves obres hi havia un mínim de 5 novel·les. Vista la excepcionalitat de la troballa, hem considerat la viabilitat de configurar el corpus al voltant d'aquestes autores i del seu context, de forma que s'ha buscat en el mateix repositori altres autors homes que també fossin exponents del mateix moviment literari, i que hi constessin com a mínim 5 obres seves del gènere narratiu. La cerca ens ha deixat uns altres tres autors que complien els criteris de selecció, de forma que aquests cinc autors (dues dones i tres homes) han sigut els seleccionats definitius per a configurar el corpus del treball.

Els cinc autors son exponents del Naturalisme en la seva vessant espanyola, un moviment artístic i literari que agafa protagonisme a finals del segle XIX i que es caracteritza principalment per la voluntat de mostrar i relatar els successos quotidians des d'una perspectiva realista. Pel que fa als llibres que hem triat per a que conformessin el corpus del treball, han estat seleccionats al atzar d'entre els que hi havia disponibles, i en alguns casos, en funció del format en què es podien descarregar (triant el que facilités en major mesura el posterior processament del text). En la taula mostrada a continuació, estan indicats els autors i els llibres escollits, junt amb altres dades de les obres com l'any de publicació o el codi amb què s'ha desat cada arxiu.

4.2. Selecció de característiques

Un cop seleccionats els llibres que han configurat el nostre corpus, hem procedit a plantejar les característiques que en podríem extreure com a trets definitoris de l'estil dels diferents autors. El que esperem d'aquests trets és que ens ajudin a definir l'escriptura dels autors a través de dades lingüístiques mesurables. Per a fer-ho, hem fet recerca entre la bibliografia consultada amb l'objectiu de seleccionar aquelles que, en estudis anteriors, han resultat eficients en l'elaboració de perfils lingüístics.

	Autor	Novel·la	Any	Arxiu
1	Emilia Pardo Bazán	La Sirena Negra	1908	epb1
		La prueba	1890	epb2
		Dulce Dueño	1911	epb3
		La Tribuna	1883	epb4
		Un Viaje de Novios	1881	epb5
2	José María de Pereda	Los Montálvez	1888	jmp1
		Peñas Arriba	1895	jmp2
		Al primer vuelo	1891	jmp3
		El sabor de la tierruca	1882	jmp4
		La Puchera	1889	jmp5
3	Concha Espina	Agua de Nieve	1911	ce1
		Despertar para Morir	1910	ce2
		La Esfinge Maragata	1914	ce3
		Dulce Nombre	1921	ce4
		La Niña de Luzmela	1909	ce5
4	Benito Pérez Galdós	Bailén	1873	bpg1
		La Desheredada	1881	bpg2
		El amigo Manso	1882	bpg3
		Misericordia	1897	bpg4
		Miau	1888	bpg5
5	Vicente Blasco Ibáñez	Entre Naranjos	1900	vbi1
		Arroz y tartana	1894	vbi2
		La catedral	1903	vbi3
		La bodega	1905	vbi4
		El intruso	1904	vbi5

Figura 1: Autors i novel·les seleccionades per al corpus de l'estudi.

En la gran majoria de casos, les característiques a les que es feia referència es repetien entre les diferents fonts consultades. Tot i així, cal esmentar que no totes aquestes característiques que hem vist han estat incloses en la selecció que hem dut a terme, ja que algunes no hem considerat que fossin viables d'analitzar o extreure en base a l'abast i les limitacions d'aquest treball. Hem seleccionat un total de 26 característiques amb les que configurar els perfils lingüístics dels autors, i les hem dividit per subgrups seguint les propostes de Soler, J., & Wanner, L. (2017) i de Cheng, N. et al. (2011).

En primer lloc, tenim les característiques basades en caràcters, que es recullen i calculen a partir del nombre total de caràcters dels documents. Aquestes inclouen: la proporció de caràcters alfabètics, caràcters especials, caràcters que son espais i també els que son salts de pàgina. A més a més, també n'hem extret el còmput total de caràcters en majúscula i

en minúscula, i els caràcters de puntuació més freqüent, com son el punt, la coma i els guions. El segon subgrup el conformen les característiques basades en paraules. Primer hem extret el nombre total de paraules que suma cadascuna de les obres, i a partir d'aquí hem inclòs els següents trets: La mitjana de lletres per paraula, la riquesa lèxica de cada llibre, el nombre de paraules de més de sis lletres, de paraules d'entre una i tres lletres, i el nombre de paraules gramaticals. El següent subgrup el formen les característiques sintàctiques, i hi estan incloses les freqüències relatives de les categories gramaticals més habituals de les paraules, tals com els noms, pronoms, verbs, adjectius, adverbis, conjuncions, i determinants. Addicionalment, hem calculat la freqüència entre temps verbals en present i en passat, i també hem distingit la presència de verbs auxiliars. Per últim, hem extret característiques estructurals, que treballen sobre la construcció de frases i paràgrafs en forma de recursos estilístics i formals que fan servir els autors. Aquest és el grup amb el menor nombre de característiques, només amb tres: Hi trobem la mitjana de paraules per frase de cada obra, i les mitjanes de frases i de paraules que conformen cada paràgraf.

4.3. Recollida de dades

4.3.1. Preprocessament dels arxius

Com s'ha mencionat anteriorment, hem descarregat els llibres en format digital des del repositori de The Project Guttenberg. Cal esmentar, però, que abans de poder començar a extreure les característiques de cada document, hem fet una breu revisió dels arxius per tal d'assegurar que no presentessin particularitats d'edició o del propi format que dificultessin la tasca. En la majoria de casos, ha calgut corregir algunes d'aquestes particularitats amb l'ajuda dels programes de processament de text que tractarem més endavant .

Les modificacions més destacades que hem dut a terme son les següents: S'ha hagut d'eliminar els caràcters que indicaven els nombres de pàgina del llibre. Han estat esborrades també les capçaleres que dividien alguns llibres en *primera*, *segona*, o *tercera* part en considerar que es tractava d'un afegit d'edició i que no resultava rellevant en la tasca de creació dels perfils dels autors. S'han mantingut, tanmateix, totes les numeracions dels capítols junt amb els seus títols corresponents (en el cas que en tinguessin), ja que considerem que aquests sí son producte dels autors i tenen una

significació en les obres, i també perquè, a diferència del cas comentat anteriorment on es tractava d'una característica marginal en el total dels arxius, els números i noms dels capítols eren una constant present en pràcticament tots els documents. Una altra modificació que hem hagut de fer manualment en un nombre reduït de casos és afegir la primera lletra de cada capítol, ja que en ocasions es presentava en forma d'imatge, fent impossible el seu processament i anàlisis. Finalment, també hi havia uns pocs arxius on el canvi de capítol estava indicat amb una successió de punts de gran longitud. Aquests punts també els hem retirat per tal que no interferissin en el còmput de caràcters i puntuació, en tractar-se d'un recurs indubtablement estilístic. Un cop finalitzada la revisió i efectuades les correccions corresponents, hem pogut procedir amb l'extracció de les dades.

4.3.2. Recursos utilitzats

Per a extreure les característiques que analitzarem hem fet ús de diferents programes de processament de text. El que més destaquem és Python, un llenguatge de programació que ens permet treballar sobre les dades lingüístiques i analitzar-les, a més de processar i manipular grans volums de text, i extreure'n informació significativa. La utilització de Python ens ha servit, entre d'altres coses, per comptabilitzar de forma automàtica les ocurrències que tenen lloc en els arxius. També per segmentar el contingut dels arxius en paraules i en frases, i ens ha permès desar els textos modificats en nous documents per tal de facilitar l'anàlisi de dades, en comptar amb diferents formats de text. Ha estat l'eina amb què hem treballat en major mesura i que ens ha permès extreure la majoria de dades.

També hem fet ús del sistema operatiu Unix, una altra eina per al processament de text i anàlisi i extracció de dades lingüístiques. A diferència del recurs anterior, que hem utilitzat com a processador de capçalera, hem fet ús de Unix en casos específics, on la formulació de les ordres i configuració del llenguatge de programació ens permetia extreure característiques concretes amb més facilitat. Primerament, l'hem fet servir per veure la riquesa lèxica de cada obra. I sobretot, ens ha sigut molt necessari de cara a les característiques sintàctiques, ja que ens ha permès fer l'anotació morfològica de les paraules amb el programa Freeling.

Totes les dades que hem anat obtenint les hem desat en un arxiu Excel, el qual ens ha permès recopilar i organitzar els resultats. A més a més, hem pogut comparar les dades dels diferents autors amb l'aplicació de funcions estadístiques com la prova t, que ens fet possible determinar la significació estadística dels resultats en cada característica extreta. En aquest arxiu Excel hem designat els autors amb les seves inicials, junt amb un nombre del 1 al 5 que denota a quina de les seves obres es refereixen les dades. Per a dinamitzar aquesta memòria, d'ara en endavant ens referirem als autors amb les seves inicials, i complementant-les amb l'ordre que pren cadascun d'ells en la recopilació presentada.

4.4. Experiments de classificació amb Weka

En un últim pas, després d'haver fet l'anàlisi de totes les dades recollides i vist la seva rellevància estadística en la discriminació entre autors, hem volgut comprovar la seva eficiència de cara a una aplicació com a eina en un procés d'identificació automàtic. Per a fer-ho hem fet servir Weka, un software d'anàlisi de dades que posa a la nostra disposició funcions com la classificació, associació, *clustering*, i selecció d'atributs, a més de comptar amb eines per al processament i visualització de dades.

Hem utilitzat quatre classificadors diferents per a veure si, amb les dades introduïdes referents a les característiques de cadascuna de les obres, el software era capaç de categoritzar-les com a pertanyents als autors corresponents. Aquests classificadors son els que s'utilitzen amb més freqüència en aquest tipus de tasques: MultilayerPerceptron, SMO, J48 i RandomForest. Els classificadors estan basats en algoritmes estadístics que es fan servir en aprenentatge automàtic per a la classificació de dades, tant de forma binaria com multiclasse. Concretament, els classificadors J48 i RandomForest treballen amb arbres de decisió.

5. RESULTATS

5.1. Característiques basades en caràcters

Per a fer l'anàlisi d'aquest primer subgrup de característiques, el primer pas ha estat extreure el nombre de caràcters total de cada document. A partir d'aquest nombre hem calculat la freqüència d'ocurrències de les característiques següents. D'aquesta manera, hem pogut observar la utilització del tret en l'escriptura de cada autor, independentment de l'extensió de les obres.

Fent una comparació global, les característiques basades en caràcters han estat el segon subgrup més eficient en la distinció i posterior classificació dels autors. En la classificació de dades que hem fet amb Weka, introduint exclusivament aquestes característiques, el percentatge d'encert en la identificació dels autors oscil·la entre el 92% i el 64%, depenent del classificador utilitzat. Com podem observar en la següent taula, el classificador Multilayerperceptron ha resultat el més encertat, només fallant en la classificació de 2 obres, mentre que J48 ha obtingut els resultats més imprecisos, fallant en 9 obres.

Classificador	Incorrecte	Correcte
Multilayerperceptron	8% (2)	92%
Smo	24% (6)	76%
J48	36% (9)	64%
Randomforest	20% (5)	80%

Figura 2: Percentatges d'error i encert en la classificació automàtica de les obres amb els diferents classificadors utilitzats segons les característiques basades en caràcters.

5.1.1. Caràcters alfabètics

Per a calcular aquesta primera característica hem extret el nombre total de lletres, tant majúscules com minúscules, que hi havia a cada document. Aquest nombre l'hem dividit entre el nombre total de caràcters de les respectives obres, i ens ha quedat com a resultat la proporció de caràcters alfabètics que constitueixen cadascun dels documents. El rang de resultats no és gaire extens, vist que l'obra amb major proporció de lletres té un 0,7993 de lletres per caràcter, i la que en té menys en té 0,7767. Totes les dades recollides i els càlculs que s'han dut a terme en relació a aquesta, i a totes les altres característiques, es poden consultar en l'annex d'aquesta memòria

Pel que fa a la rellevància estadística d'aquestes dades, només en un dels autors ha resultat ser una característica significativa. Es tracta del cinquè autor (VBI) el qual és l'únic que dona valors per sota de 0,05 en relació a tots els altres autors, i per tant és estadísticament rellevant. També podem veure, en el gràfic a continuació, la distribució de les obres de cada autor en relació al rang de valors de les dades obtingudes. Podem observar que en el cas dels quatre autors restants, els resultats de les obres s'estenen, en major o menor mesura, al llarg de l'eix horitzontal. Mentre que les obres de l'autor VBI (presentades en

color rosa), es concentren a la dreta d'aquest, indicant una consistència al voltant del valor més elevat.

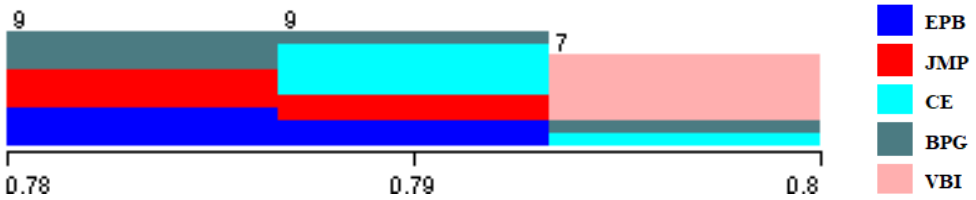


Figura 3: Distribució de les obres de cada autor en el rang de valors per a la característica de caràcters alfabètics. Font: Weka.

5.1.2. Caràcters especials

Per a calcular aquesta característica hem extret el nombre de caràcters especials de cada arxiu, i posteriorment l'hem dividit pel nombre total de caràcters del document. Hem comptat com a caràcters especials tots els que es consideren signes de puntuació: punt (.), coma (,), punt i coma (;), dos punts (:), guions (-), parèntesis (()), claudàtors ([]), cometes simples (‘), dobles (“”) i angulars («»), signes d'interrogació (¿?), d'exclamació (¡!), i d'altres que es fan servir en menys freqüència com ‘&%#@~$*/^{}_’. Els resultats de obtinguts oscil·len entre el 0,0282 i el 0,0461, sent el cinquè autor (VBI) el que concentra resultats amb valors més reduïts i la primera autora (EPB) qui en presenta de més elevats.

No és casualitat llavors que aquests dos autors hagin estat els qui presenten significació estadística en aquesta característica. Com mostren els càlculs realitzats, aquest tret és estadísticament distintiu en la comparació entre el cinquè autor (VBI) i els autors restants, i entre la primera autora (EPB) i tres dels altres autors. Hi ha una única excepció entre la primera autora (EPB) i el quart (BPG), on la comparació de dades no és significativa. Pel que fa als altres tres autors, aquesta característica no representa un tret que faci possible la distinció entre ells, només respecte als dos autors destacats. També ho podem apreciar en el gràfic següent, on veiem les obres de color rosa (VBI) i blau marí (EPB) concentrades en els dos extrems del rang de valors, mentre que les obres dels altres tres autors es dispersen al llarg de l'eix horitzontal.

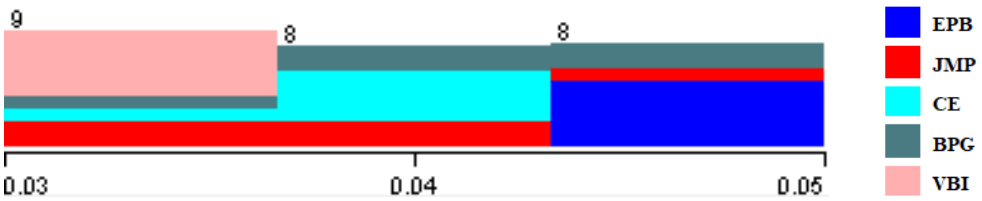


Figura 4: Distribució de les obres de cada autor en el rang de valors per a la característica de caràcters especials. Font: Weka.

5.1.3. Caràcters d'espai

Per al càlcul d'aquesta característica hem comptabilitzat el nombre d'espais en blanc que contenia cada arxiu, i l'hem dividit entre el nombre total de caràcters, obtenint així la freqüència d'aparició dels espais en la totalitat del document. Les dades mostren que els espais en blanc ocupen entre el 0,1601 i el 0,1780 de la totalitat de caràcters que conformen els documents. Parant atenció a les dades podem intuir una certa continuïtat en aquesta característica, ja que els nombres de molts autors no semblen dispersar-se gaire entre les seves respectives obres respecte al rang total.

Pel que fa a la significació estadística d'aquests valors, ens trobem amb una dada molt positiva, ja que en la gran majoria de les parelles comparades resulta ser una característica rellevant. Hi ha dos autors que obtenen valors inferiors al 0,05 en les comparacions amb tots els altres autors, aquests son el segon (JMP) i la tercera (CE). Els tres autors restants no obtenen rellevància estadística en els aparellaments que els relacionen, és a dir, entre EPB i BPG, BPG i VBI, i VBI i EPB. Reprenent el que introduíem abans, en el gràfic podem observar clarament aquesta continuïtat en les dades de tres autors, que concentren les seves obres en la mateixa zona de l'eix horitzontal: El segon (JMP, marcat en vermell, amb els valors més alts), la tercera (CE, marcada en blau clar, amb els més baixos), i el cinquè (VBI, marcat en rosa, al centre de l'eix). Els dos autors restants també tendeixen a la franja central, però amb menys solidesa que els ja esmentats, ja que només hi concentren part de les seves obres.

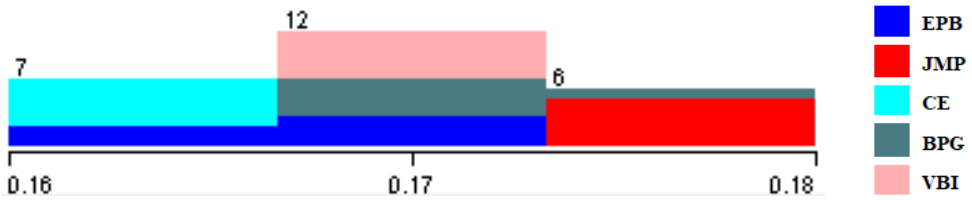


Figura 5: Distribució de les obres de cada autor en el rang de valors per a la característica de caràcters d'espai. Font: Weka.

5.1.4. Caràcters de salt de línia

Aquesta característica la hem extret fent la divisió del nombre de caràcters de salt de línia entre el nombre total de caràcters del document. Veient les dades recollides, observem una uniformitat prou destacada entre quatre dels cinc autors, els resultats dels quals oscil·len entre el 0,0038 i el 0,0074. Pel contrari, despunta la autora número tres (CE), els resultats de la qual van entre el 0,0085 i el 0,0155. Com sembla evident, és aquesta autora (CE) l'única que obté significació estadística en les seves dades en relació a tots els altres autors. Aquest tret és altament rellevant per a caracteritzar la seva producció. Contràriament, els quatre autors restants no obtenen valors estadísticament significatius en ningun dels seus aparellaments.

Podem observar en el gràfic com destaca indiscutiblement la tercera autora (CE, en color blau clar). Les seves obres ocupen els dos terços de l'eix que corresponen als valors més elevats del rang, mentre que els colors corresponents a les obres dels altres quatre autors es concentren en els valors més baixos.

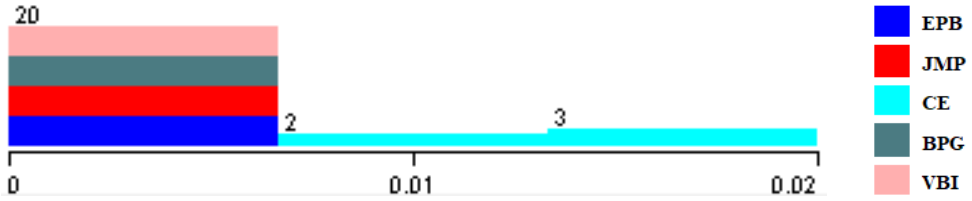


Figura 6: Distribució de les obres de cada autor en el rang de valors per a la característica de caràcters de salt de línia. Font: Weka.

5.1.5. Caràcters en majúscula

Per a l'extracció d'aquesta característica hem comptabilitzat totes les lletres en majúscula que contenia cada document, i hem fet la divisió d'aquest nombre entre el total de les

lletres que hi havia en l'arxiu. Cal destacar que, en aquest cas, hem cregut més adient agafar aquest valor i no el del nombre total de caràcters (com veníem fent en les característiques anteriors), perquè considerem que la rellevància està en la proporció de majúscules en relació a les lletres, i no als caràcters no alfabètics. També mencionar que aquesta característica és complementària al càlcul de la proporció de lletres en minúscula, que no hem calculat apart perquè obtindríem els mateixos resultats a la inversa. En una primera ullada a les dades recopilades veiem que els valors de lletres en majúscula entre les totals van entre el 0,0116 i el 0,0251.

Parant atenció als resultats del càlcul del p valor en els diferents aparellaments entre autors, podem declarar que no és un tret que sigui suficientment significatiu com per categoritzar un autor respecte a tots els altres. Igualment, sí és una característica amb una alta rellevància estadística, ja que en moltes d'aquestes comparacions sí que suposa un tret distintiu: distingeix la primera autora (EPB) del segon (JMP) i el cinquè (VBI); el segon (JMP), de tots els autors excepte del cinquè (VBI); les dades de la tercera (CE) i del quart (BPG) son significatives respecte a tots els autors menys la primera (EPB); i el cinquè (VBI), com ja hem dit, només falla en comparació amb el segon (JMP). En resum, considerem que, tot i no encertar completament en cap autor, és tracta d'una característica molt útil de cara al perfil lingüístic dels autors. Observant el gràfic, podem intuir aquesta irregularitat en la significació veient els solapaments entre les obres dels autors en els diferents nivells de l'eix horitzontal. L'únic autor que es manté en la mateixa franja és el cinquè (VBI, de color rosa). Pel contrari, l'autora número u (EPB, en blau marí) és la que compta amb els valors més dispersos, i la seva obra s'estén per gran part del rang.

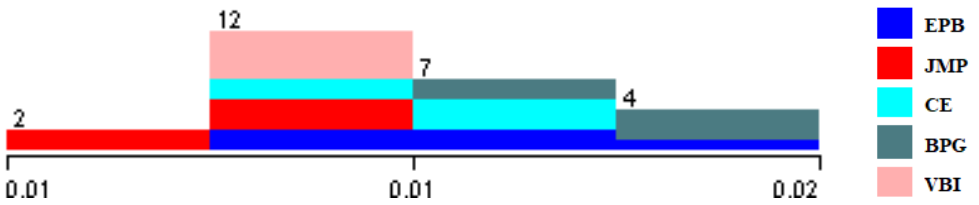


Figura 7: Distribució de les obres de cada autor en el rang de valors per a la característica de caràcters en majúscules. Font: Weka.

5.1.6. Nombre total de punts

Hem extret aquesta característica obtenint el nombre total de punts que hi consten en els arxius i dividint-lo entre la suma dels seus caràcters totals. El rang dels resultats va des del 0,0170 al 0,0070, mostrant així la proporció de punts en relació als caràcters totals que conformen les obres escollides. Inicialment és difícil extreure conclusions sobre la significació d'aquesta característica, ja que observant els valors resultants no sembla que hi destaqui cap dels autors. Tanmateix, parant atenció als aparellaments de dades entre autors per al càlcul de la seva rellevància estadística, sí trobem resultats interessants. La primera autora (EPB), és qui denota més significació de les seves dades, ja que aquest tret serveix per a distingir-la de tres altres autors, a excepció del quart autor (BPG). També notem una rellevància estadística en les comparacions entre el quart autor (BPG) i els autors segon (JMP) i cinquè (VBI). En la resta de casos, no detectem significació entre les dades.

Observant el gràfic, veiem, en efecte, que les dades es mostren més disperses del que hem vist en moltes de les característiques anteriors. Destaquem la primera autora (EPB, en blau marí), que concentra totes les seves obres en el costat de l'eix horitzontal amb valors més alts, i també el cinquè autor (VBI, en rosa), que pràcticament es manté en la seva totalitat en el costat oposat del rang. L'autor número quatre (BPG, en gris), mostra un cas curiós, ja que totes les seves obres estan situades a la franja dels valors més elevats, però una única obra en el costat oposat de l'eix deu alterar notablement les seves dades totals. Els dos autors restants estan repartits al llarg de la totalitat de l'eix.

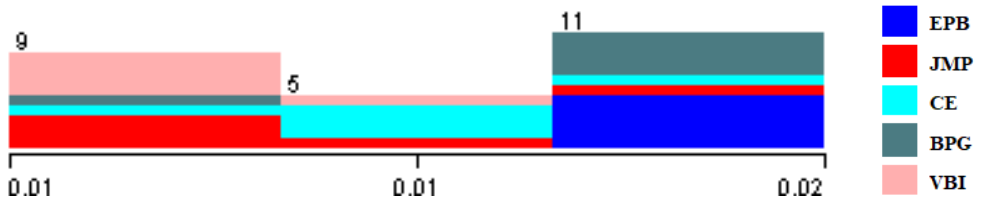


Figura 8: Distribució de les obres de cada autor en el rang de valors per a la característica de nombre total de punts. Font: Weka.

5.1.7. Nombre total de comes

Per al càlcul d'aquesta característica, hem extret el nombre de comes presents en cada arxiu i l'hem dividit pel nombre total dels caràcters. Els resultats obtinguts van entre el

0,0117 i el 0,0176, configurant un rang d'ocurrències no gaire extens. Veient el còmput total de les dades, no sembla que els autors presentin gaires diferències, ja que en la majoria de casos les els resultats que mostren varien entre la major part del rang. Sí hi podríem destacar el cinquè autor (VBI), doncs sembla el que presenta valors més baixos i amb més regularitat que els altres.

En efecte, als càlculs dels p valors, es veu amb claredat que és aquest cinquè autor (VBI) l'únic que presenta una significació estadística en les seves dades respecte dels altres. Aquesta característica només serveix per a distingir-lo a ell, mentre que cap dels altres autors presenta rellevància alguna en els seus resultats. De fet, ni tan sols en el cas d'aquest cinquè autor és una característica concloent, ja que en la comparació amb la tercera autora (CE) les dades no mostren significació estadística. Veient aquests resultats podríem afirmar que és una de les característiques extretes que menys informació aporta de cara a l'elaboració dels perfils lingüístics dels autors. Veient el gràfic, podem detectar aquest cinquè autor (VBI, en color rosa) agrupant les seves obres en els valors més baixos de l'eix, però tot i així, no destaca per damunt dels altres autors, que a excepció del segon autor (JMP, en vermell), concentrat en els valors centrals del rang, estan prou dispersats al llarg del eix horitzontal.

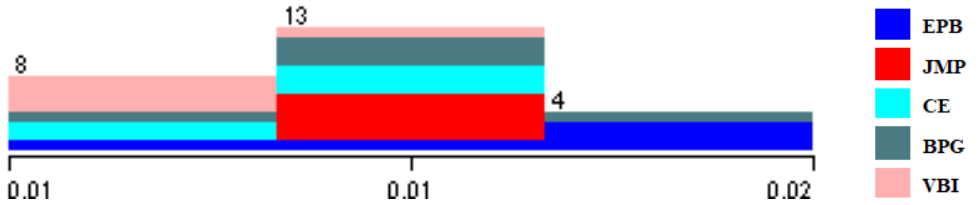


Figura 9: Distribució de les obres de cada autor en el rang de valors per a la característica de nombre total de comes. Font: Weka.

5.1.8. Nombre total de guions

L'última de les característiques basades en caràcters és el nombre de guions, que hem obtingut dividint el nombre de guions que presentaven els documents entre el seu respectiu nombre de caràcters totals. Destacar que en el còmput dels guions hi entraven tant els guions "curts" (-) com els "llargs" o també anomenats ratlles (—). A més a més, creiem que és una característica molt interessant, perquè pot denotar la preferència d'un autor per afegir més o menys contingut en format de diàleg, és a dir, fent ús de l'estil

directe. Els valors obtinguts oscil·len entre el 0,010 i el 0,050, i hi podem observar tendències consistents en l'ús d'aquest caràcter en alguns dels autors, que es mouen en rangs molt més reduïts: La tercera autora (CE) es mou entre els valors més alts, mentre que el cinquè (VBI) ho fa entre els més baixos.

Pel que fa a la rellevància estadística de les dades obtingudes, detectem com destaca notablement el cinquè autor (VBI), els resultats del qual obtenen significació estadística en comparació amb tots els altres autors. Pel que fa a la resta, només hi ha dos casos més on aquesta característica és rellevant: En la distinció entre la primera i quart autor (EPB i BPG) i entre el segon i la tercera (JMP i CE). En els aparellaments restants, no resulta un tret prou significatiu. Aquestes dades es mostren clarament en el gràfic adjuntat, on hi destaca el cinquè autor (VBI, en color rosa) mantenint-se entre els valors més reduïts del rang. Els altres autors es mostren repartits entre les diferents franges del gràfic, a excepció de la primera autora (EPB), que concentra totes les seves obres entre els valors més intermedis.

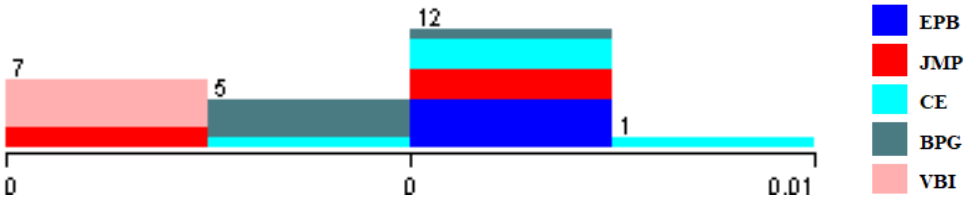


Figura 10: Distribució de les obres de cada autor en el rang de valors per a la característica de nombre total de guions. Font: Weka.

5.2. Característiques basades en paraules

Com ja hem introduït anteriorment, aquest subgrup de característiques gira entorn de la noció de paraula. Per a dur a terme la seva extracció, hem recollit el nombre total de paraules que conformava cada document. A partir d'aquesta dada, hem anat calculant la resta de característiques, de nou, evitant un possible biaix dels resultats a causa de la diversa longitud dels textos.

Aplicant només aquestes característiques en el software Weka, no hem obtingut tan bons resultats com en el subgrup anterior en la classificació automàtica de les obres. Amb un dels classificadors (SMO), el percentatge d'errors és més alt que el d'encerts (56%

d'errors, l'equivalent a 14 de les obres, vers el 44% d'encerts). Els altres tres classificadors, però, sí assoleixen entre un 64% i un 72% d'encert. MultilayerPerceptron i RandomForest son els que més bons resultats obtenen, fallant en la classificació de 7 i 8 obres respectivament de les 25 analitzades en total.

Classificador	Incorrecte	Correcte
Multilayerperceptron	28% (7)	72%
Smo	56% (14)	44%
J48	36% (9)	64%
Randomforest	32% (8)	68%

Figura 11: Percentatges d'error i encert en la classificació automàtica de les obres amb els diferents classificadors utilitzats segons les característiques basades en paraules.

5.2.1. Riquesa lèxica

Per al càlcul d'aquesta característica hem hagut d'extreure el nombre de paraules diferents de cada arxiu (també denominat *types*), i dividir-lo entre el nombre total de paraules del mateix. Observem en els resultats que la xifra més baixa és de 0,1151, mentre que la més elevada arriba als 0,2414. Podem observar com els resultats de cada autor semblen arreplegar-se en rangs més reduïts, denotant així una consistència o regularitat en les seves dades. Per exemple, destaca molt per sobre la primera autora (EPB), amb els valors més elevats, seguida de la tercera autora (CE), i a prou distància del segon autor (JMP), qui presenta els valors més baixos.

Amb els resultats del càlcul dels p valors podem confirmar el que ja intuíem, doncs son justament la primera autora (EPB), el segon (JMP), i la tercera (CE), els únics que presenten significació estadística en els seus resultats. No ho fan en tots els casos, però: Les dades de la primera autora (EPB) i la tercera (CE) no es distingeixen entre elles, mentre que el segon autor (JMP) presenta rellevància estadística amb tots els autors excepte amb el cinquè (VBI). Observant ara el gràfic amb la distribució de les obres en el rang de valors totals obtinguts, veiem clarament la tendència de cada autor en aquesta característica. La primera (EPB, en blau marí) i tercera autora (CE, en blau clar) ocupen les franges amb valors més elevats, mentre que els tres autors restants mantenen, gairebé en la seva totalitat, els resultats corresponents a les seves obres en l'extrem amb valors més baixos de l'eix horitzontal.

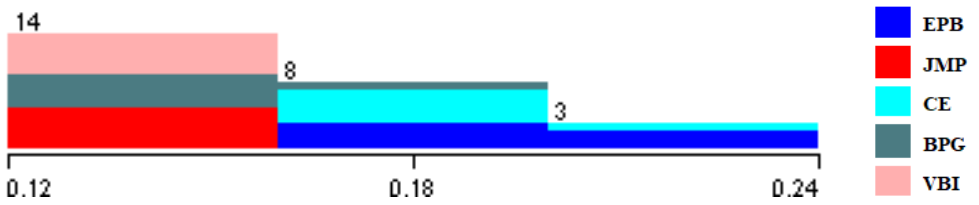


Figura 12: Distribució de les obres de cada autor en el rang de valors per a la característica de riquesa lèxica. Font: Weka.

5.2.2. Mitjana de lletres per paraula

Per a l'extracció d'aquesta característica hem dividit el nombre de paraules de cada document entre el nombre de lletres que contenia (sense incloure puntuació o altres caràcters no alfabètics). Els resultats obtinguts afirmen que la llargada mitjana de les paraules és d'entre 4,2750 i 4,7102 lletres cadascuna. Podem detectar amb facilitat que els valors més alts els concentra el cinquè autor (VBI), mentre que els més baixos els té el segon (JMP). Els càlculs de la rellevància estadística ens mostren com, en efecte, aquesta característica té un alt grau de significació en relació a tres autors: El segon autor (JMP) que obté dades estadísticament rellevants en comparació amb tots els altres autors. El quart (BPG), que només falla en l'aparellament amb la primera autora (EPB), però manté aquest tret com a significatiu en la resta de casos. I el cinquè (VBI), que es distingeix de tots els autors menys de la tercera (CE).

Observant el gràfic de la distribució de les dades, podem constatar aquests resultats en veure com la majoria dels autors presenten una regularitat en els seus valors, que els fa mantenir-se, casi en la seva totalitat, en les mateixes franges de l'eix. La tercera (CE, en blau clar) i el cinquè autor (VBI, en rosa) ocupen els valors més alts. El segon (JMP, en vermell), es situa a l'esquerra de l'eix completament sol. El quart (BPG, en gris) ocupa la franja del mig, incidint lleugerament en els valors alts. La primera autora (EPB, en blau marí) és clarament la que mostra resultats més dispersos, distribuint les seves obres de forma prou uniforme entre les dues franges amb valors més elevats.

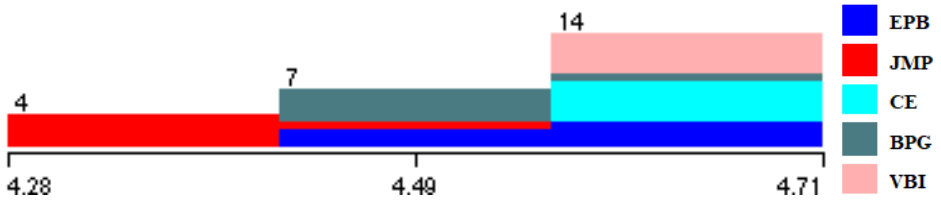


Figura 13: Distribució de les obres de cada autor en el rang de valors per a la característica de mitjana de lletres per paraula. Font: Weka.

5.2.3. Paraules de més de sis lletres

Una altra característica que hem volgut extreure és la proporció de paraules considerades llargues, que son aquelles formades per més de sis lletres. Hem comptabilitzat les paraules a partir de set lletres, i aquest nombre l'hem dividit pel de les paraules totals del document en qüestió. Les dades obtingudes mostren que la proporció de paraules llargues respecte a les paraules totals és d'entre 0,1990 i 0,2720. Podem destacar a primera vista el segon autor (JMP) que compta amb els valors més baixos, mentre que la tercera autora (CE) i el cinquè (VBI) presenten els més elevats.

Pel que fa a la rellevància estadística d'aquesta característica, obtenim resultats optimistes. Seguint amb les suposicions anteriors, aquest tret serveix per a distingir el segon autor (JMP) dels altres quatre. També presenten significació estadística el quart autor (BPG) respecte de tres altres autors, només deixant fora la primera (EPB), i el cinquè autor (VBI), els resultats del qual el distingeixen de tots els autors menys de la tercera (CE).

El gràfic acaba de confirmar les nostres conclusions: El segon autor (JMP, en vermell) ocupa casi en la seva totalitat la franja amb els valors més baixos. La tercera autora (CE, en blau clar) i el cinquè (VBI, en rosa) es troben íntegrament en l'extrem dret de l'eix, on es situen els valors més alts. Les obres dels dos autors restants es distribueixen entre els diferents valors de l'eix. Un detall curiós que creiem interessant destacar és que aquest gràfic és pràcticament idèntic al de la característica anterior (la llargada mitjana de les paraules). Podem afirmar que els dos trets presenten una correlació, doncs sembla clar que els autors que utilitzin amb freqüència paraules més llargues, tindran una mitjana de més lletres per paraula en el còmput total.

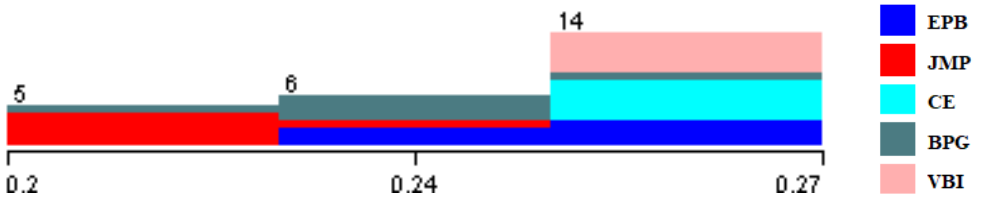


Figura 14: Distribució de les obres de cada autor en el rang de valors per a la característica de paraules de més de sis lletres. Font: Weka.

5.2.4. Paraules d'entre una i tres lletres

En oposició a la característica anterior, aquest cop hem volgut extreure el nombre de paraules curtes dels documents, que son aquelles amb una longitud d'entre una i tres lletres. De nou, hem dividit el nombre de paraules curtes entre aquelles paraules que conformen la totalitat de les obres. Els resultats oscil·len entre el 0,4380 i el 0,4893. Podem observar com els valors més alts els presenta el segon autor (JMP), mentre que la tercera autora (CE) és la que en presenta de més baixos.

La rellevància estadística d'aquesta característica ens mostra que, com es podia intuir, el segon autor (JMP) és distingeix significativament dels quatre restants. Però, en relació als altres autors cap no aconsegueix resultats significatius en la totalitat dels emparellaments. La primera autora (EMP) no presenta rellevància estadística en les seves dades respecte a cap dels altres autors (a excepció, com ja hem dit, del segon). Tanmateix, els autors restants (CE, BPG i VBI) sí es distingeixen estadísticament entre ells en totes les seves comparacions.

Veiem clarament aquest succés en el gràfic, ja que la primera autora (EPB, en blau marí) té les seves obres distribuïdes al llarg de tot l'eix, el qual la fa susceptible a no distingir-se dels altres autors. Pel contrari, el segon (JMP, en vermell) i el quart autor (BPG, en gris) se situen en les franges dreta i mitja de l'eix respectivament. Els dos autors restants reparteixen les seves obres en les franges més baixes del rang de valors, tot i que la tercera autora (CE, en blau clar) sí tendeix als valors més baixos en la majoria de les seves obres.

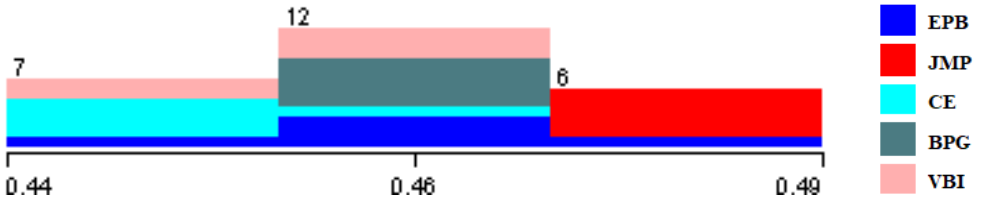


Figura 15: Distribució de les obres de cada autor en el rang de valors per a la característica de paraules d'entre una i tres lletres. Font: Weka.

5.2.5. Paraules gramaticals

L'última característica d'aquest subgrup consisteix a veure la freqüència d'aparició de les denominades paraules gramaticals o *stopwords* respecte al total de paraules de les obres. Es consideren paraules gramaticals aquelles que no tenen de per sí un significat, sinó que només porten informació gramatical, tals com els articles, preposicions, conjuncions, pronoms, etc. Per a extreure aquestes paraules hem fet ús d'una biblioteca de Python, que les processa de forma automàtica. Un cop hem extret el nombre de paraules gramaticals de cada document, les hem dividit entre les paraules totals d'aquest. Els resultats obtinguts oscil·len entre el 1,5258 i el 2,4042. Individualment, destaca sobretot el segon autor (JMP), amb els valors més alts (tots per sobre del 2).

La rellevància estadística d'aquesta característica és principalment útil per a distingir el segon autor (JMP), que com ja intuíem, presenta significació respecte a tots els altres aparellaments realitzats. Els autors restants, però, no obtenen dades tan favorables, doncs només hi ha dos comparacions més en que els resultats d'autors diferents presenten significació estadística: Aquests son la de la tercera i quart autor (CE i BPG), i del quart amb el cinquè (BPG i VBI). En el gràfic veiem clarament aquesta tendència, doncs destaca sobretot el color vermell del segon autor (JMP) situat en els valors més alts de l'eix. També veiem les obres de la tercera (CE, en blau clar) i el cinquè autor (VBI, en rosa) en bloc al costat oposat de l'eix, amb els resultats més baixos. Els dos autors restants distribueixen les seves obres al llarg del rang de valors, tot i que podem destacar que la primera autora (EPB, en blau marí) tendeix als valors menys elevats.

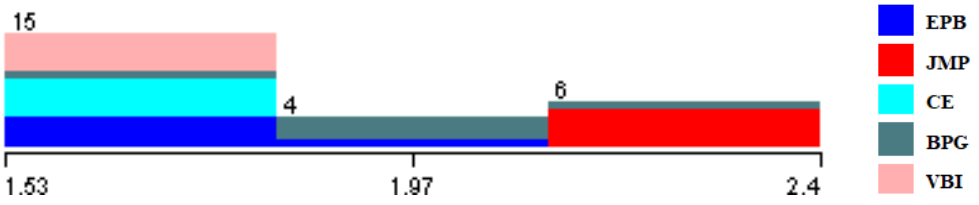


Figura 16: Distribució de les obres de cada autor en el rang de valors per a la característica de paraules gramaticals o *stopwords*. Font: Weka.

5.3. Característiques sintàctiques

Aquest subgrup de característiques que hem extret es basa en mesurar la freqüència relativa de les diferents categories gramaticals de les paraules al llarg dels documents. Per a fer aquest càlcul, com ja hem avançat anteriorment, hem fet servir la biblioteca Freeling,

que ens ha permès dur a terme funcions com l'anàlisi i etiquetatge morfològic de cada paraula de forma automatitzada. Cal fer un incís en la metodologia emprada, perquè Freeling presenta diferents alternatives en el format de presentació de les dades: Depenent de l'ordre utilitzada, pot fer l'etiquetatge morfològic presentant l'opció més probable, en ocasions, afegint el percentatge d'encert, també poden aparèixer totes les etiquetes possibles per a la paraula, ordenades segons la probabilitat o, de nou, amb el percentatge. En aquest estudi hem volgut comptar únicament amb una etiqueta per paraula, de manera que hem programat l'ordre per tal d'obtenir l'opció més probable de categoria gramatical, a risc que algunes paraules (estimem, una proporció molt minoritària) no hagin estat etiquetades correctament.

Aquest grup de característiques és, individualment, el que millors resultats obté de cara a la classificació d'obres duta a terme amb el software Weka. El percentatge d'encerts obtingut comprèn entre el 76% i el 88% de les obres, en funció del classificador utilitzat: MultilayerPerceptron i RandomForest son els que aconsegueixen millors marques, amb un 88% de les obres ben classificades, i només fallant en 3 d'elles. El classificador que més errors obté és J48, amb 6 obres classificades de forma incorrecta.

Classificador	Incorrecte	Correcte
Multilayerperceptron	12% (3)	88%
Smo	20% (5)	80%
J48	24% (6)	76%
Randomforest	12% (3)	88%

Figura 17: Percentatges d'error i encert en la classificació automàtica de les obres amb els diferents classificadors utilitzats segons les característiques sintàctiques.

5.3.1. Freqüència relativa de noms

Per a l'extracció d'aquesta característica hem seleccionat totes les paraules categoritzades com a noms de cada document (independentment de trets com el gènere, nombre, classe, etc). Posteriorment, les hem dividit entre el total de paraules del mateix, obtenint així la proporció de noms per paraula dels arxius. Els resultats mostren que la proporció més baixa és de 0,1953 noms, mentre que la més alta és de 0,2611. A primera vista despunta el segon autor (JMP) com el que presenta resultats més baixos, mentre que la tercera (CE) i el cinquè (VBI) destaquen per obtenir els més elevats.

Aquesta característica ha resultat ser molt significativa a l'hora d'establir distincions entre els autors. En primer lloc, tenim el segon autor (JMP) que, com ja avançàvem, obté rellevància estadística en els seus resultats en comparació a tots els altres. També obtenen bons resultats el quart autor (BPG), que es aconsegueix distingir-se de tots els autors menys de la primera (EPB); i el cinquè (VBI), que només manca significació estadística en relació a la tercera autora (CE). Observant el gràfic, constatem la solidesa del segon autor (JMP, en vermell), íntegrament en la franja de valors més baixos del rang, i del cinquè (VBI, en rosa), ocupant l'extrem dret de l'eix. Els tres autors restants es solapen, entre ells i amb els altres dos, al llarg de l'eix horitzontal. Si bé, podem destacar la tendència de la primera autora (EPB, en blau marí) a mantenir-se en els valors centrals, mentre que els altres dos s'apropen més als extrems del rang.

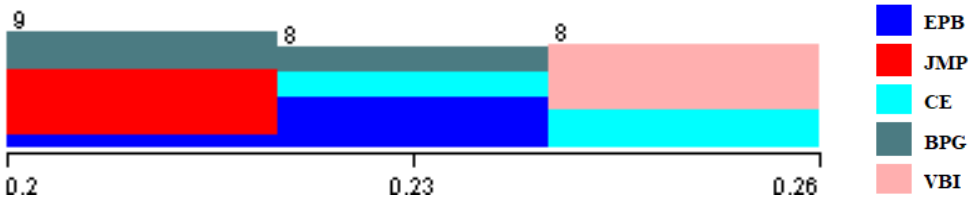


Figura 18: Distribució de les obres de cada autor en el rang de valors per a la característica de freqüència relativa de noms. Font: Weka.

5.3.2. Freqüència relativa de pronoms

Aquesta característica la hem calculat dividint el nombre de pronoms detectats en cada obra entre el nombre total de paraules que contenia aquesta. Els resultats mostren que la proporció més baixa és de 0,0609, en oposició a la més alta que arriba als 0,1217. De nou, s'intueixen tendències en l'ús dels pronoms de cada autor: Veiem els valors més alts concentrats en les obres del segon autor (JMP), però seguit de prop per la primera (EPB) i quart autor (BPG). Pel contrari, hi ha un gran descens en els valors de la tercera autora (CE), que presenta els valors més baixos, a penes sobrepassant el 0,085.

La significació estadística d'aquest tret no atorga distinció absoluta a cap dels subjectes. Sí hi destaca la tercera autora (CE), que obté resultats amb rellevància estadística en oposició a tots els autors menys al cinquè (VBI). De manera paral·lela, el cinquè autor (VBI) també es distingeix dels tres autors que presentaven les dades més altes, però no

ho fa amb la tercera (CE). Els altres tres autors no arriben a presentar cap distinció estadísticament significant entre ells. El gràfic, tanmateix, no acaba de mostrar amb claredat aquesta polarització que intuíem, ja que els autors més distintius (CE i VBI, en blau clar i rosa respectivament), ocupen tant la franja dels valors més baixos com la central, on es barregen amb les obres dels altres autors. Podem destacar, però, el segon autor (JMP, en vermell), que clarament concentra les seves obres en els valors més alts de l'eix.

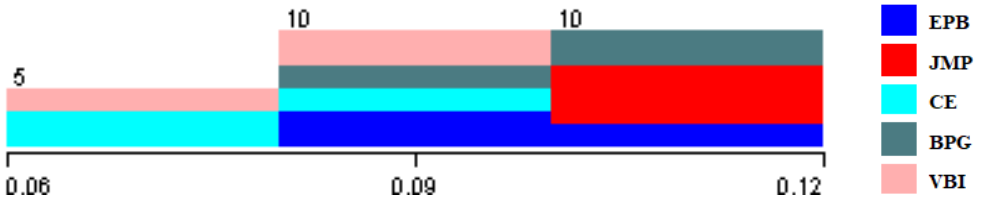


Figura 19: Distribució de les obres de cada autor en el rang de valors per a la característica de freqüència relativa de pronoms. Font: Weka.

5.3.3. Freqüència relativa d'adjectius

L'extracció d'aquesta característica la hem dut a terme fent la divisió entre el nombre d'adjectius detectats en cada obra i el nombre total de paraules de la mateixa. Els resultats obtinguts oscil·len entre els 0,0467 i els 0,0952 adjectius per cada paraula del text. De nou, destaquen els resultats del segon autor (JMP), amb els valors més baixos de la recopilació, sense arribar als valors de 0,06. Pel contrari, la tercera autora (CE) presenta, en les seves cinc obres, els cinc valors més elevats.

La rellevància estadística d'aquesta característica és prou elevada: Com s'intuïa a partir dels resultats, la tercera autora (CE) és distingeix dels quatre restants amb dades estadísticament significatives en tots els aparellaments. També hi estan a prop el cinquè autor (VBI), qui obté resultats distintius en comparació amb tots menys la primera autora (EPB), i el segon (JMP), els resultats del qual només cauen en la irrellevància estadística en relació al quart autor (BPG). El gràfic ens mostra com despunta la tercera autora (CE, en blau clar) amb els valors més elevats, seguida de la primera autora (EPB), que ocupa majoritàriament la franja central. En oposició, els tres autors restants es concentren en la franja dels valors més baixos del rang, si bé, el cinquè autor (VBI, en rosa), també té part de les seves obres en la franja central de l'eix.

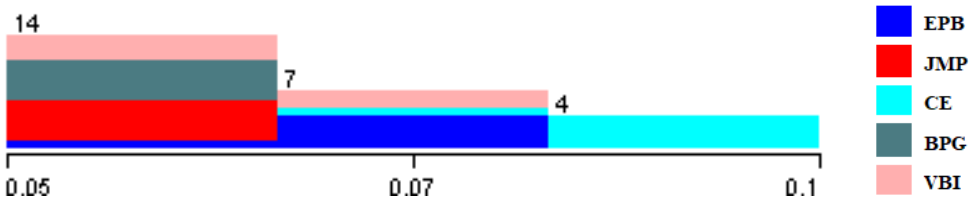


Figura 20: Distribució de les obres de cada autor en el rang de valors per a la característica de freqüència relativa d'adjectius. Font: Weka.

5.3.4. Freqüència relativa d'adverbis

Hem extret aquesta característica comptabilitzant tots els adverbis presents en cada obra per, seguidament, dividir aquest nombre entre el nombre de paraules totals d'aquesta. Els resultats obtinguts mostren que la proporció és d'entre 0,0379 i 0,0714 adverbis per paraula dels documents. Els resultats més baixos de la taula els monopolitzen la tercera (CE) i el cinquè autor (VBI), que no arriben a superar el 0,047 i 0,043 respectivament. Els valors més alts semblen pertànyer a les obres del segon autor (JMP).

Els càlculs de p valors entre les obres dels diferents autors ens proporcionen els resultats més eficients que hem vist fins ara, fent d'aquesta característica la més significativa d'entre les estudiades. Hi ha només un únic aparellament que obté resultats sense rellevància estadística, que és el realitzat entre la primera autora (EPB) i el quart autor (BPG). En la resta de casos, les dades denoten significació estadística, de manera que aquest tret pot distingir les obres dels autors en la gran majoria de les ocasions.

Parant atenció al gràfic podem observar com els resultats de cada autor presenten una forta consistència: Els autors amb resultats més baixos (CE en blau clar, i VBI en rosa) ocupen en la seva totalitat la franja esquerra de l'eix horitzontal. El segon autor (JMP, en vermell) es manté a l'extrem amb els valors més alts, mentre que la primera autora (EPB, en blau marí) ocupa, amb la majoria de les seves obres, la franja mitjana. El quart autor (BPG, en gris), és el que mostra més irregularitat en els resultats de les seves obres, que es reparteixen entre les dues columnes situades a la dreta de l'eix, tot i que podem apreciar la similitud amb la distribució presentada per la primera autora, que els fan estadísticament compatibles.

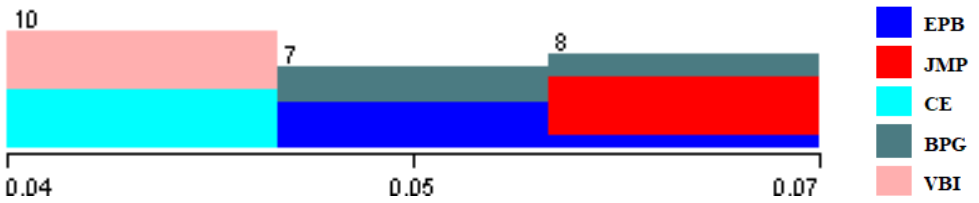


Figura 21: Distribució de les obres de cada autor en el rang de valors per a la característica de freqüència relativa d'adverbis. Font: Weka.

5.3.5. Freqüència relativa de verbs

Hem calculat aquesta característica extraient totes les formes verbals presents en cada document, i posteriorment, hem dividit aquest nombre entre el nombre total de paraules del text. Els resultats indiquen que la proporció de verbs per paraula és d'entre 0,1594 i 0,2109. A primera vista, no s'observen grans diferències entre els resultats dels autors. Les proves de rellevància estadística mostren, en efecte, que les dades de tots els autors son generalment massa similars com per constituir una diferència estadísticament significativa. Tanmateix, hi ha dues excepcions d'aparellaments en els quals aquesta característica sí és distintiva. Aquests son les comparacions del quart autor (BPG), tant amb la tercera autora (CE) com amb el cinquè autor (VBI).

Observant el gràfic de barres, veiem com aquesta característica és clarament poc distintiva, doncs les obres de tots els autors es solapen en les diferents franges sense una tendència clara. Es pot destacar, en qualsevol cas, la consistència de les obres dels autors tercer (CE, en blau clar) i cinquè (VBI, en rosa), que es mantenen, gairebé en la seva totalitat, en els valors més baixos del rang. En comparació a ells, el quart autor (BPG, en gris), només ocupa les dues franges restants de l'eix, que representen els valors més alts.

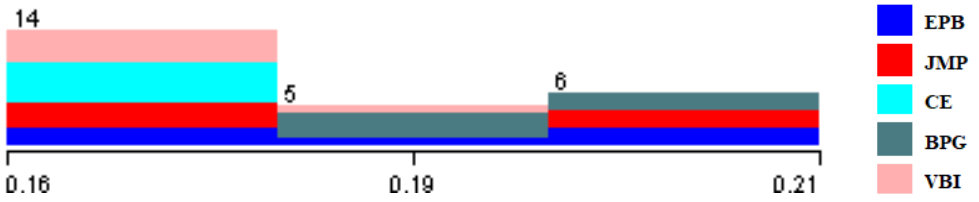


Figura 22: Distribució de les obres de cada autor en el rang de valors per a la característica de freqüència relativa de verbs. Font: Weka.

5.3.6. Freqüència relativa de verbs en passat

En relació a la característica anterior, també volgut extreure la freqüència relativa dels verbs en temps passat. Hem fet el càlcul dels verbs en passat recopilant el nombre d'ocurrències d'aquests i dividint-lo entre el nombre total de verbs de cada document. Ho hem fet així, i no dividint-lo entre el nombre total de paraules, perquè ens interessa veure la proporció de l'ús del temps passat en relació a totes les formes verbals. Creiem que, de fer servir el nombre de paraules total, podríem estar esbiaixant els resultats, en tant que dependrien directament de la producció total de verbs pròpia de cada autor. Els resultats obtinguts oscil·len entre el 0,1754 i el 0,4389. En una primera observació superficial, sembla, més aviat, que l'ús del temps passat sigui una característica pròpia de cada obra, i no dels seus autors, doncs la majoria presenten resultats molt variats. Els únics autors que semblen mantenir certa continuïtat en els seus resultats, és a dir, que es concentren en un rang de valors molt més estret, son el quart (BGP) i el cinquè autor (VBI).

Curiosament, en els resultats de les proves de p valor, l'únic autor que presenta significació estadística en aquesta característica és el cinquè (VBI), i ho fa en les comparacions amb tots els autors, a excepció de la tercera (CE). Cap dels altres autors mostra rellevància estadística en els seus resultats, de forma que és un tret poc distintiu. En el gràfic podem veure com, en efecte, el cinquè autor (VBI, en rosa) destaca per sobre dels altres en mantenir totes les seves obres en l'extrem dret de l'eix horitzontal, amb els valors més alts del rang. El quart (BPG, en gris), també aconsegueix mantenir gairebé totes les seves obres en la franja central. Els altres tres autors mostren tendències molt irregulars respecte al posicionament de les seves obres. Concretament, les autores primera (EPB, en blau marí) i tercera (CE, en blau clar), ocupen l'extensió total del rang.

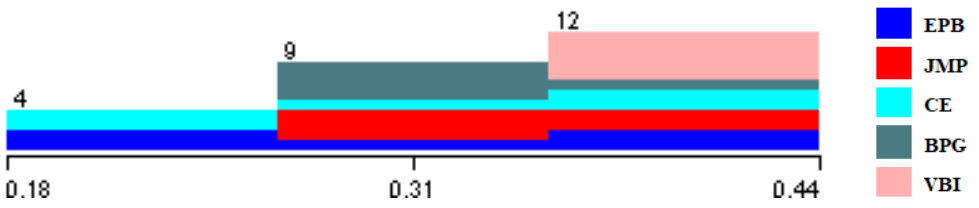


Figura 23: Distribució de les obres de cada autor en el rang de valors per a la característica de freqüència relativa de verbs en passat. Font: Weka.

5.3.7. Freqüència relativa de verbs en present

En oposició a la proporció de verbs en temps passat, hem volgut també extreure la freqüència relativa dels verbs en present. Ho hem fet, de nou, dividint el nombre de verbs en present entre el nombre de total de verbs que apareixen en el document. Es tracta pràcticament de la mateixa operació realitzada en la característica anterior, però a l'inrevés, de manera que podem anticipar uns resultats inversament proporcionals. En aquest cas, el rang de valors obtinguts comprèn entre el 0,1262 i el 0,4174. De nou, observem la mateixa disparitat entre els resultats dels mateixos autors, el qual ens porta a concloure que és un tret més circumstancial que no pas característic de cada autor. Com ja preveiem, els resultats en relació a la significació estadística d'aquest tret son iguals que els vistos en l'anàlisi dels temps en passat: L'únic autor que mostra resultats amb rellevància estadística és el cinquè (VBI), fent-ho en els aparellaments amb la primera (EPB), segon (JMP), i quart autor (BGP), i deixant de ser distintiu en comparació amb la tercera autora (CE).

Veient el gràfic, constatem la irregularitat dels resultats en les obres dels mateixos autors, doncs els veiem repartits al llarg del rang de valors. Els únics casos on es detecta una tendència comuna son els del quart (BPG, en gris) i el cinquè autor (VBI, en rosa), que es concentren en les franges central i esquerra respectivament, de forma íntegra en el cas del quart autor (BPG) i només amb una excepció en el cas del cinquè (VBI). El segon autor (JMP, en vermell), també tendeix a presentar valors més baixos. Comparant aquest gràfic amb el de la característica anterior, veiem clarament com els resultats guarden una relació, si no exactament inversa, molt destacable.

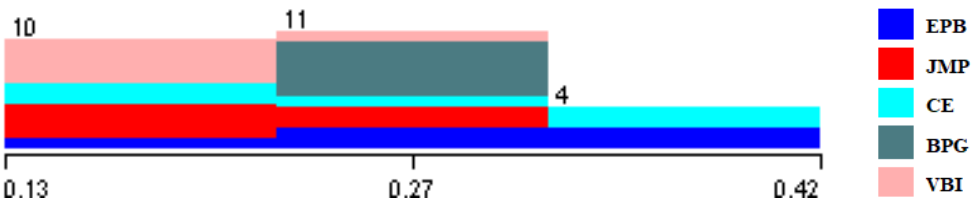


Figura 24: Distribució de les obres de cada autor en el rang de valors per a la característica de freqüència relativa de verbs en present. Font: Weka.

5.3.8. Freqüència relativa de verbs auxiliars

Ja per finalitzar amb les característiques relacionades amb els verbs, hem extret també la freqüència relativa dels verbs auxiliars. Per a fer-ho, hem dividit el nombre de verbs auxiliars comptabilitzats en cada document entre el nombre total de paraules. En aquest cas, hem fet servir el nombre de paraules enlloc del nombre total de verbs, perquè el nostre interès no estava en la proporció de formes verbals auxiliars respecte als verbs totals, sinó a la presencia de formes auxiliars en la totalitat del text. Els resultats obtinguts oscil·len entre el 0,0031 i el 0,0095.

Observant les dades obtingudes en les proves de significança estadística, ens trobem amb un dels resultats més pobres de tota la recerca: Només hi ha dos casos on aquesta característica suposa una distinció estadísticament rellevant entre dos autors. Es tracta de les comparacions entre la primera autora (EPB) i el segon autor (JMP), i entre la primera (EPB) i el quart (BPG). En la resta de casos no hi ha rellevància estadística per aquest tret. En el gràfic podem destacar la tendència dels autors segon (JMP, en vermell), quart (BPG, en gris), i cinquè (VBI, en rosa) a situar-se en la franja amb els valors més alts del rang. Pel contrari, les autores primera (EPB, en blau marí) i tercera (CE, en blau clar) tenen les seves obres repartides al llarg de l'eix horitzontal, per tant insinuant poca uniformitat en els seus valors.

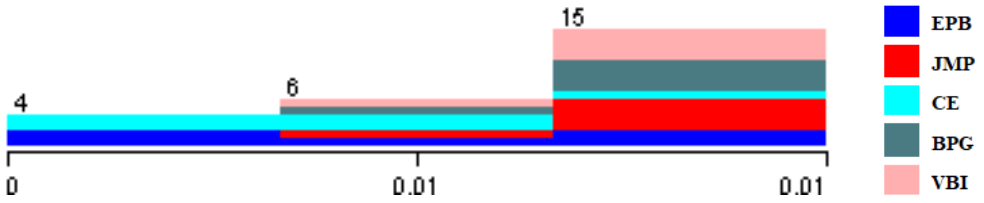


Figura 25: Distribució de les obres de cada autor en el rang de valors per a la característica de freqüència relativa de verbs auxiliars. Font: Weka.

5.3.9. Freqüència relativa de conjuncions

Per a l'extracció d'aquesta característica hem comptabilitat totes les conjuncions presents en cada obra, i les hem dividit pel nombre total de paraules d'aquestes, obtenint així la freqüència relativa d'aparició d'aquestes formes. Els resultats oscil·len entre el 0,578 i el 0,0892. Destaquen individualment el segon autor (JMP) amb els valors més elevats, tots

per damunt del 0,08, i el cinquè (VBI) amb els més baixos. Tot i que, majoritàriament, veiem com els resultats de cada autor semblen moure's en un rang més estret.

Els resultats dels càlculs dels p valors indiquen una forta significació estadística en la gran majoria dels aparellaments entre autors, fent d'aquesta característica una de les més distintives. Només hi ha dos casos on els resultats no tenen suficient rellevància estadística com per considerar-se distintius: Les comparacions de la primera autora (EBP) amb la tercera (CE) i amb el quart (BPG). En tots els altres casos sí suposa un tret distintiu. En el gràfic podem veure-hi, en primer lloc, la concentració de les obres del cinquè autor (VBI, en rosa) en la franja amb valors més baixos, en oposició a les del segon autor (JMP, en vermell), que presenten els valors més alts. També observem la situació de les obres del quart autor (BPG, en gris), gairebé totes en la franja central. Les obres pertanyents a les dues autores restants estan repartides de forma més irregular. Concretament, la primera autora (EPB, en blau marí) ocupa la totalitat de l'eix horitzontal.

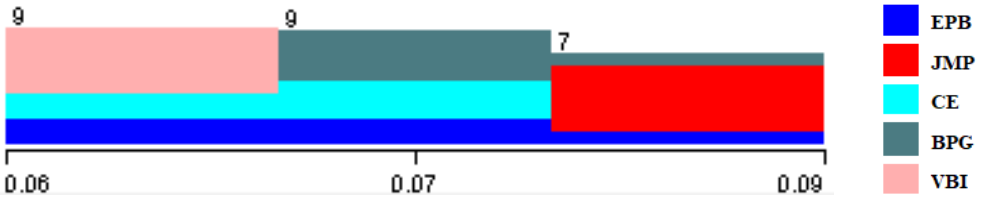


Figura 26: Distribució de les obres de cada autor en el rang de valors per a la característica de freqüència relativa de conjuncions. Font: Weka.

5.3.10. Freqüència relativa de determinants

Hem dut a terme l'extracció d'aquesta característica dividint el nombre total de determinants (incloent-hi articles, demostratius, indefinits, possessius, interrogatius i exclamatius) entre el nombre total de paraules de cada document. Els resultats obtinguts presenten proporcions amb valors d'entre el 0,1406 i el 0,1875. Es pot destacar el cinquè autor (VBI) com el que presenta valors més elevats en els seus resultats, mentre que els valors més baixos son compartits per diferents autors.

Observant els resultats obtinguts en les proves de significació estadística, veiem que aquesta característica esdevé un tret distintiu per a la producció tant de la tercera (CE) com del cinquè autor (VBI), que indiquen rellevància estadística en les comparacions amb

tots els altres autors. No ho és, tanmateix, en cap dels aparellaments dels altres tres autors entre ells. En el gràfic podem constatar aquests resultats, ja que veiem com aquests dos autors que obtenen significació estadística son els únics que ocupen la franja dreta de l'eix, en el cas del cinquè autor (VBI, en rosa) amb totes les seves obres, i en el cas de la tercera (CE, en blau clar), amb la majoria ajustada. Els altres tres autors, com apuntàvem anteriorment, ocupen les dues franges restants, amb els valors més baixos del rang. Podem destacar també el quart autor (BPG, en gris), que concentra totes les seves obres a la franja esquerra de l'eix horitzontal.

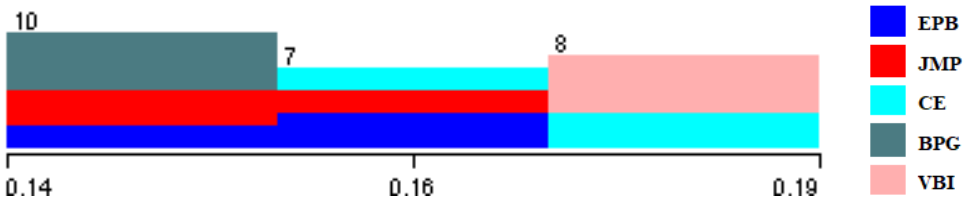


Figura 27: Distribució de les obres de cada autor en el rang de valors per a la característica de freqüència relativa de determinants. Font: Weka.

5.4. Característiques estructurals

Finalment, ens trobem amb l'últim subgrup de característiques que hem extret per tal de dur a terme l'elaboració dels perfils lingüístics de cada autor. Aquest grup el definim com a característiques estructurals, i giren entorn de l'ús dels autors de les paraules, en tant que les distribueixen en frases o paràgrafs. Si bé aquest tipus de trets no sempre té perquè reflectir l'estil de l'autor en tant que pot estar motivat per el contingut que descriu o per les necessitats o fluïdesa del text, hem cregut rellevant analitzar-les. Per a fer-ho, hem hagut d'extreure el nombre de frases i paràgrafs de cada text, i a partir d'aquests valors hem calculat la resta de característiques.

En les proves de classificació automàtica que hem dut a terme amb Weka, aquest subgrup ha estat el que individualment obté pitjors resultats en la identificació d'autors. Tanmateix, no ens resulta sorprenent ni alarmant ja que es tracta, amb diferència, del grup que menys característiques aporta. Els resultats comprenen percentatges d'entre el 44% i el 52% d'encert, només superant la meitat en dos dels classificadors (SMO i RandomForest). Els dos classificadors restants obtenen més errors que encerts, amb 13 i 14 obres classificades de manera incorrecta de les 25 totals. No es tracta d'un grup de

característiques que pugui identificar els autors amb cap garantia de efectivitat. Hem comprovat, però, que la seva aplicació en la realització d'aquesta recerca no perjudica el resultat global de la suma amb les característiques anteriors.

Classificador	Incorrecte	Correcte
Multilayerperceptron	52% (13)	48%
Smo	48% (12)	52%
J48	56% (14)	44%
Randomforest	48% (12)	52%

Figura 28: Percentatges d'error i encert en la classificació automàtica de les obres amb els diferents classificadors utilitzats segons les característiques estructurals.

5.4.1. Mitjana de paraules per frase

La primera característica d'aquest subgrup és la mitjana de paraules per frase, que hem extret dividint el nombre total de paraules de cada document entre el nombre de frases que el constituïen. Els resultats obtinguts indiquen que les obres presenten frases amb un nombre mitjà de paraules d'entre 15,3962 i 35,6793. A primera vista, destaca el segon autor (JMP) com aquell amb les frases més llargues, mentre que els altres quatre mostren valors més semblants.

En efecte, en la taula amb els resultats de les proves de significació estadística veiem que l'únic autor que presenta una distinció rellevant en relació a tots els altres és el segon (JMP). Addicionalment, també presenten dades estadísticament significatives els autors tercer (CE) i cinquè (VBI) en la seva comparació. En tots els altres aparellaments, aquesta característica no suposa un tret distintiu. En el gràfic següent podem observar com el segon autor (JMP, en vermell), ocupa les franges amb els valors més alts, en oposició a la resta d'autors, que concentren el gruix de les seves obres en el costat dret de l'eix horitzontal. Esmentar també la tercera autora (CE, en blau clar) que tendeix a posicionar-se en els valors centrals de l'eix.

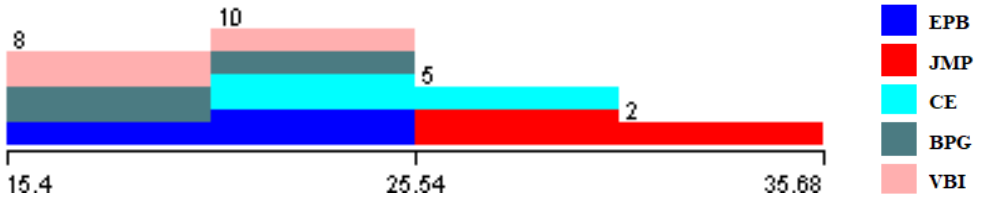


Figura 29: Distribució de les obres de cada autor en el rang de valors per a la característica de mitjana de paraules per frase. Font: Weka.

5.4.2. Mitjana de paraules per paràgraf

Per a l'extracció d'aquesta característica hem comptabilitzat, en primer lloc, el nombre total de paràgrafs de cada obra (entenent per paràgrafs qualsevol frase o cadena de paraules delimitada per salts de línia, de forma que també s'hi inclouen frases aïllades o que formen part d'un diàleg, si és dona el cas). Després, hem dividit el nombre total de paraules per aquest nombre, obtenint així la mitjana de paraules que conformen els paràgrafs del document. Els resultats oscil·len entre les 21,44 i 93,43 paraules per paràgraf, configurant el rang de valors més ample que hem vist al llarg de tot l'anàlisi. Tot i que en alguns casos s'observa prou disparitat entre els resultats d'obres dels mateixos autors, hi destaca notablement la tercera autora (CE), com aquella amb els valors menys elevats. De fet, les seves cinc obres son les que tenen les puntuacions més baixes, cap d'elles superant les 40 paraules de mitjana.

Veient els resultats de les proves de p valor realitzades, confirmem el que s'intuïa a partir dels resultats obtinguts en el càlcul de mitjanes: L'única autora que obté significació estadística en els seus resultats és la tercera (CE), per a qui aquesta característica la distingeix dels quatre autors restants. Els altres autors no obtenen rellevància estadística en els resultats de cap dels seus emparellaments. Parant atenció al gràfic, veiem com la tercera autora (CE, en blau clar) ocupa la franja amb els valors més baixos de l'eix, situant-se a l'extrem esquerre. Els altres quatre autors es concentren sobre la franja central, tots amb quatre de les seves cinc obres.

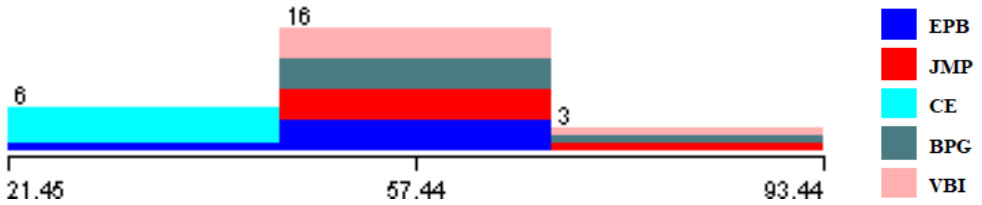


Figura 30: Distribució de les obres de cada autor en el rang de valors per a la característica de mitjana de paraules per paràgraf. Font: Weka.

5.4.3. Mitjana de frases per paràgraf

Per acabar, ens hem fixat de nou en la longitud dels paràgrafs dels autors, en aquest cas agafant la mesura del nombre mitjà de frases que els configuren. Hem fet la divisió del

nombre total de frases detectades en els arxius entre el nombre de paràgrafs en que es reparteixen. Les diferents mitjanes obtingudes presenten un rang de valors que va d'entre 1,02 a 4,23 frases per paràgraf. De nou, l'autora que més destaca és la tercera (CE), que torna a presentar els valors més baixos de tots els recollits; cap de les mitjanes de les seves obres arriba a les dues frases per paràgraf.

La realització de les proves de significació estadística ens mostra que, en efecte, la tercera autora (CE) obté resultats distintius en comparació amb tots els altres autors. També presenta els mateixos resultats el segon autor (JMP), per a qui aquesta característica també esdevé un tret distintiu en oposició als autors restants. Aquests son els dos únics casos, però, ja que els tres autors restants no aconsegueixen resultats estadísticament significatius.

Observant el gràfic, veiem com la tercera autora (CE, en blau clar) destaca un cop més en oposició a tots els altres autors. Segueix sent l'única que se situa íntegrament en la franja de l'eix amb valors més baixos. Tot i la gran similitud al gràfic de la característica anterior, la major diferència que podem observar és que el segon autor (JMP, en vermell), reparteix les seves obres entre les franges esquerra i central, fet que il·lustra la distinció dels seus resultats respecte als autors restants. Si bé 3 de les seves 5 obres segueixen estant a la franja central junt amb el gruix dels altres autors, es desmarca d'ells en tant que s'allunya de la franja dreta, que conté els valors més alts.

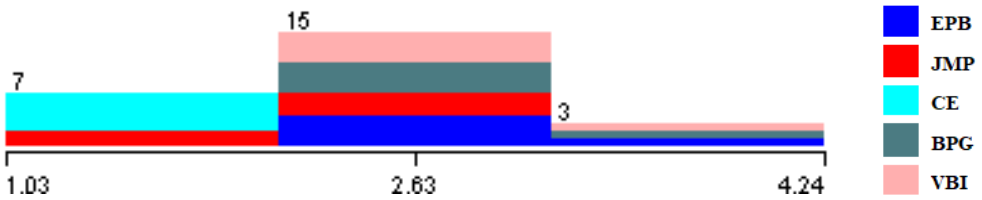


Figura 31: Distribució de les obres de cada autor en el rang de valors per a la característica de mitjana de frases per paràgraf. Font: Weka.

6. DISCUSSIÓ DE RESULTATS

Després d'haver vist els resultats obtinguts amb totes les característiques, tant individualment com en les subcategories en què les hem agrupat, ens trobem finalment amb els resultats de tots els trets en conjunt. Hem comptabilitzat les 26 característiques

en el software Weka per a veure l'eficiència total de les dades extretes en la classificació automàtica de les obres. Els resultats obtinguts es presenten en la taula a continuació.

En els quatre classificadors utilitzats el percentatge d'encerts en la classificació d'obres és d'entre el 80% i el 96%. El menys eficient és J48, que presenta el nombre més alt d'obres classificades erròniament (5 obres, el 20% del total). Pel contrari, SMO només s'equivoca en la classificació d'una única obra, obtenint el percentatge d'encert més elevat que hem vist fins ara i, per tant, sent el classificador més precís. Els dos classificadors restants fallen entre dues i tres obres, i assoleixen percentatges d'encert del 92% i 88% respectivament.

Classificador	Incorrecte	Correcte
Multilayerperceptron	8% (2)	92%
Smo	4% (1)	96%
J48	20% (5)	80%
Randomforest	12% (3)	88%

Figura 32: Percentatges d'error i encert en la classificació automàtica de les obres amb els diferents classificadors utilitzats segons totes les característiques extretes.

Considerem que els resultats obtinguts en els quatre classificadors son molt positius, ja que denoten que les característiques i dades extretes de cadascuna de les obres han resultat (en major o menor mesura) de molta utilitat en la configuració dels perfils lingüístics dels autors i, en conseqüència, en la classificació automàtica de les seves obres. Amb totes les dades introduïdes, el software és capaç de classificar correctament, en el pitjor dels casos, un 80% de les obres, el qual és un percentatge alt que indica l'encert en l'àmplia majoria. Tanmateix, com ja hem anat veient al llarg d'aquest apartat, algunes de les característiques han resultat molt eficients en aquesta distinció dels autors, mentre que d'altres no ho han estat gaire. A continuació repassarem breument quins d'aquests trets han estat els més significatius en la creació dels perfils lingüístics dels autors.

Les característiques que han demostrat ser més rellevants en la distinció dels autors han estat, en primer lloc, la freqüència relativa d'adverbis, obtenint significació estadística en 9 dels 10 possibles aparellaments entre autors. Seguidament, el nombre de caràcters d'espai i la freqüència relativa de les conjuncions, amb 8 de 10 aparellaments estadísticament distintius. I a continuació, trets com el nombre de caràcters en majúscula, la riquesa lèxica, la mitjana de lletres per paraula, les paraules de més de sis lletres i les

d'entre una i tres lletres, i les freqüències relatives de noms i adjectius, que obtenen rellevància estadística en 7 dels 10 aparellaments possibles.

Cal destacar també algunes característiques que han resultat estadísticament distintives en autors concrets, és a dir, que les dades que les conformaven aconseguien distingir un dels autors dels altres, tot i no obtenir resultats significatius en els autors restants. Exemples d'aquestes característiques son les paraules gramaticals, la mitjana de paraules per frase, i la de frases per paràgraf en el cas del segon autor (JMP). En el cas de la tercera autora (CE), el nombre de caràcters de salt de línia, la freqüència relativa de determinants, i les mitjanes de paraules i frases per paràgraf. I finalment, en el cas del cinquè autor (VBI), la proporció de caràcters alfabètics, de caràcters especials, el nombre total de guions, i la freqüència relativa de determinants. Com anàvem dient, l'eficiència d'aquestes característiques ha estat lligada a autors concrets, de forma que, tot i la seva evident utilitat en aquest estudi, no podem afirmar amb certesa que siguin objectivament eficaces en altres tasques o comparacions similars amb altres autors.

En el costat oposat, veiem també les característiques que han resultat menys eficients en l'elaboració dels perfils. Aquestes son la freqüència relativa de verbs i de verbs auxiliars, que només han indicat significació estadística en 2 dels possibles aparellaments realitzats, seguides pel nombre total de comes i les freqüències relatives de verbs en passat i present, amb distinció en 3 dels 10 aparellaments.

7. CONCLUSIONS

Aquesta recerca es plantejava com una tasca en la que configurar els perfils lingüístics de cinc autors diferents, a través d'obres de la seva producció. Per a fer-ho, hem vist com seleccionar les característiques a analitzar, com hem dut a terme l'extracció de les dades, i com visualitzar i interpretar els resultats obtinguts segons la seva significació estadística. Havent realitzat aquesta tasca i comentat amb detall els resultats, tant en conjunt com individualment en cadascuna de les característiques, considerem que s'han acomplert els objectius establerts. A més, reprenent el comentat en l'apartat anterior, trobem que els resultats assolits han estat molt favorables, i per consegüent enalteixen la feina realitzada.

Hem vist com les característiques seleccionades han aconseguit classificar les obres corresponents a cada autor en una gran majoria dels casos, amb percentatges d'entre el

80% i el 96% d'encert. Hem pogut veure també quins dels subgrups en què hem agrupat aquestes característiques resultaven més eficients en aquests experiments de classificació, i quins, pel contrari, obtenien menys èxit. Hem observat cadascuna de les característiques analitzades, parant atenció a la seva efectivitat en cada autor, i en tots els aparellaments possibles entre aquests. En alguns casos, sent aquesta gairebé nul·la, mentre que en d'altres esdevenien trets altament distintius. Considerem que aquesta recerca ha resultat en una matriu de dades, pertanyents a cadascun dels cinc autors, que aconsegueix configurar els seus perfils lingüístics de forma majoritàriament exitosa, tot i que lluny de ser infal·lible.

Convé ressaltar que durant la realització d'aquest treball han sorgit limitacions i dificultats que han condicionat la tasca, i creiem necessari fer-ne menció. El factor més destacat és la manca de recursos accessibles que ens hem trobat principalment en la fase de selecció del corpus: Inicialment es pretenia fer aquesta recerca a partir d'autors i llibres en llengua catalana, però l'escassetat d'obres disponibles en català ho ha fet impossible, així que s'ha treballat amb un corpus en castellà. A més a més, tot i comptar amb moltes més possibilitats en aquesta llengua, cal recordar que el corpus s'ha hagut de configurar a partir de les dues autores triades, que eren les úniques de tota la biblioteca consultada amb més de cinc obres disponibles. A més, els tres autors restants eren, de nou, els únics autors contemporanis a elles amb aquest mateix nombre d'obres. En definitiva, podem concloure que aquesta escassetat en literatura digital d'accés obert ha condicionat indiscutiblement la selecció del corpus, que podria haver inclòs un major nombre d'autors que, a la vegada, hagués enriquit la recerca.

Enllaçant amb això, cal ser conscients de les limitacions del treball, i acceptar que la tasca realitzada conforma un anàlisis en format reduït. Al cap i a la fi, s'ha analitzat un total de 25 obres, corresponents a 5 autors diferents. Els resultats i conclusions obtingudes difícilment es poden generalitzar fins al punt de poder confirmar o desmentir, per exemple, l'eficàcia de les característiques extretes. Tanmateix, sí podem concloure que, en la nostra recerca, aquests trets lingüístics seleccionats han estat de gran utilitat en la tasca. Addicionalment, la bibliografia consultada proporcionava un gran nombre de característiques lingüístiques que no hem extret en aquesta recerca (parcialment a causa de l'abast del treball i del desconeixement de l'autora). Sens dubte, l'extracció i anàlisis d'aquestes altres característiques col·laboraran en millorar la precisió i eficiència de la creació de perfils i classificació automàtica en altres tasques de major extensió.

8. BIBLIOGRAFIA

Benzebouchi, N. E., Azizi, N., Hammami, N. E., Schwab, D., Khelaifia, M. C. E., & Aldwairi, M. (2019). Authors' writing styles based authorship identification system using the text representation vector. *2019 16th International Multi-Conference on Systems, Signals & Devices (SSD)*, 371-376.

Chaski, C. E. (2001). Empirical evaluations of language-based author identification techniques. *International Journal of Speech Language and the Law, 8*(1), 1-65. https://doi.org/10.1558/sll.2001.8.1.1

Cheng, N., Chandramouli, R., & Subbalakshmi, K. P. (2011). Author gender identification from text. *Digital Investigation*, *8*(1), 78-88. https://doi.org/10.1016/j.diin.2011.04.002

Grant, T., & Baker, K. (2001). Identifying reliable, valid markers of authorship: a response to Chaski. *International Journal of Speech Language and the Law*, *8*(1), 66-79. https://doi.org/10.1558/sll.2001.8.1.66

Houvardas, J., & Stamatatos, E. (2006). N-gram feature selection for authorship identification. En *Artificial Intelligence: Methodology, Systems, and Applications* (pp. 77-86). Springer Berlin Heidelberg.

Iyer, R. R., & Rose, C. P. (2019). A machine learning framework for authorship identification from texts. En *arXiv [cs.CL]*. http://arxiv.org/abs/1912.10204

Project Gutenberg. (s. f.). Project Gutenberg. Recuperat 14 de juny de 2023, de https://www.gutenberg.org/

Queralt, S., & Giménez García, R. (2019). La imitación como contraargumento en peritajes de atribución de autoría: estudio de un caso. *Estudios de lingüística aplicada, 68,* 131. https://doi.org/10.22201/enallt.01852647p.2018.68.746

Rangel, F., Rosso, 12 Paolo, Koppel, M., Stamatatos, E., & Inches, G. (s. f.). *Overview of the author profiling task at PAN 2013*. Upv.es. Recuperat 14 de juny de 2023, de

https://riunet.upv.es/bitstream/handle/10251/46636/CLEF2013-AuthorProfiling.pdf?sequence=2&isAllowed=y

ShaukatTamboli, M., & S. Prasad, R. (2013). Authorship analysis and identification techniques: A review. *International Journal of Computer Applications*, *77*(16), 11-15. https://doi.org/10.5120/13566-1375

Soler, J., & Wanner, L. (2017). On the relevance of syntactic and discourse features for author profiling and identification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*.

Stamatatos, E., Rangel, F., Tschuggnall, M., Stein, B., Kestemont, M., Rosso, P., & Potthast, M. (2018). Overview of PAN 2018: Author identification, author profiling, and author obfuscation. En *Lecture Notes in Computer Science* (pp. 267-285). Springer International Publishing.

Weren, E., Kauer, A., Lucas, Viviane, Wives, L. P., & Moreira De Oliveira, J. (2014). Examining Multiple Features for Author Profiling. *Journal of Information and Data Management*, *5*, 266-279.

Zhang, C., Wu, X., Niu, Z., & Ding, W. (2014). Authorship identification from unstructured texts. *Knowledge-Based Systems*, *66*, 99-111. https://doi.org/10.1016/j.knosys.2014.04.025

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, *57*(3), 378-393. https://doi.org/10.1002/asi.20316

9. ANNEX

Resultats de l'extracció de les característiques

ARXIU	num caràct	caràcters alfabètics		caràcters especials		espais [ ]		salt de línia [\n]	
epb1	234839	184005	0,7835368	10126	0,0431189	39406	0,1678001	1288	0,0054846
epb2	361822	283034	0,7822465	15080	0,041678	61642	0,1703655	2056	0,0056824
epb3	391116	305618	0,7813999	18065	0,0461883	64501	0,1649153	2912	0,0074454
epb4	374993	294679	0,7858253	15439	0,0411714	62501	0,1666724	2363	0,0063015
epb5	395081	310445	0,7857756	16631	0,0420952	65571	0,1659685	2434	0,0061608
jmp1	599675	472692	0,788247	19549	0,0325993	104326	0,1739709	3097	0,0051645
jmp2	749058	591050	0,7890577	22394	0,0298962	132767	0,1772453	2847	0,0038008
jmp3	564453	438433	0,7767396	23742	0,042062	98855	0,1751342	3409	0,0060395
jmp4	454356	353859	0,7788144	17558	0,0386437	79543	0,1750676	3380	0,0074391
jmp5	689757	537866	0,7797906	25085	0,0363679	122819	0,1780613	3979	0,0057687
ce1	455300	361088	0,7930771	16039	0,0352273	74276	0,1631364	3897	0,0085592
ce2	336725	265386	0,7881387	13229	0,0392873	54451	0,1617076	3659	0,0108664
ce3	524038	413209	0,7885096	20073	0,0383045	83923	0,1601468	6823	0,01302
ce4	212894	167746	0,787932	7418	0,0348436	34993	0,1643682	2737	0,0128562
ce5	210618	165230	0,7845008	6999	0,0332308	34337	0,1630298	3277	0,015559
bpg1	368925	292381	0,7925215	12285	0,0332995	62079	0,16827	1939	0,0052558
bpg2	815583	638805	0,7832495	34101	0,0418118	136901	0,1678566	5554	0,0068099
bpg3	519684	407718	0,7845498	20096	0,0386697	88217	0,1697512	3601	0,0069292
bpg4	476384	371186	0,7791739	20049	0,0420858	82052	0,1722392	3033	0,0063667
bpg5	555389	435498	0,7841315	21882	0,0393994	95333	0,1716509	2622	0,004721
vbi1	480358	383743	0,7988688	13811	0,0287515	79887	0,1663072	2909	0,0060559
vbi2	536757	428589	0,7984786	15834	0,0294994	89361	0,1664832	2957	0,005509
vbi3	572025	457234	0,7993252	16159	0,0282488	96336	0,1684122	2292	0,0040068
vbi4	586015	466227	0,7955889	18188	0,0310367	97704	0,1667261	3886	0,0066312
vbi5	567644	453620	0,7991276	16415	0,0289178	94596	0,1666467	3009	0,0053009

ARXIU	num caràct	majúscules		punts [.]		comes [,]		guió [—]	
epb1	234839	3224	0,0175213	3515	0,0149677	4058	0,0172799	751	0,0031979
epb2	361822	5389	0,0190401	5838	0,016135	4850	0,0134044	1225	0,0033856
epb3	391116	7298	0,0238795	6681	0,0170819	6897	0,0176342	1361	0,0034798
epb4	374993	5264	0,0178635	5376	0,0143363	5733	0,0152883	1273	0,0033947
epb5	395081	5744	0,0185025	6257	0,0158373	6271	0,0158727	1498	0,0037916
jmp1	599675	5639	0,0119295	4875	0,0081294	8738	0,0145712	1064	0,0017743
jmp2	749058	6862	0,0116098	5316	0,0070969	10675	0,0142512	1175	0,0015686
jmp3	564453	7651	0,0174508	7907	0,0140083	8606	0,0152466	1875	0,0033218
jmp4	454356	6117	0,0172865	4772	0,0105028	6994	0,0153932	1554	0,0034202
jmp5	689757	8865	0,0164818	6479	0,0093932	10130	0,0146863	2148	0,0031141
ce1	455300	6130	0,0169765	5081	0,0111597	6381	0,0140149	1437	0,0031562
ce2	336725	4881	0,0183921	5286	0,0156983	4855	0,0144183	1011	0,0030025
ce3	524038	7537	0,0182402	5547	0,0105851	7101	0,0135505	2632	0,0050225
ce4	212894	3310	0,0197322	2286	0,0107377	2653	0,0124616	836	0,0039268
ce5	210618	3253	0,0196877	1493	0,0070887	3010	0,0142913	644	0,0030577
bpg1	368925	6025	0,0206067	3438	0,009319	5431	0,0147211	982	0,0026618
bpg2	815583	16056	0,0251344	12731	0,0156097	11897	0,0145871	2217	0,0027183
bpg3	519684	8047	0,0197367	8517	0,0163888	7120	0,0137006	1375	0,0026458
bpg4	476384	8163	0,0219917	7248	0,0152146	8004	0,0168016	1449	0,0030417
bpg5	555389	9521	0,0218623	8001	0,0144061	8232	0,014822	1292	0,0023263
vbi1	480358	5902	0,0153801	4473	0,0093118	5638	0,0117371	750	0,0015613
vbi2	536757	6508	0,0151847	4876	0,0090842	7400	0,0137865	559	0,0010414
vbi3	572025	8205	0,0179449	5410	0,0094576	7119	0,0124453	868	0,0015174
vbi4	586015	8155	0,0174915	6120	0,0104434	7665	0,0130799	1061	0,0018105
vbi5	567644	7621	0,0168004	5571	0,0098142	6928	0,0122048	1016	0,0017899

ARXIU	num parau	types	riquesa lèx	mitjana lle	tres/parau	paraules +6		paraules 1-3		paraules gramaticals	
epb1	40410	9755	0,2414006	184005	4,5534521	9990	0,247216	19017	0,4706013	17682	1,76997
epb2	62673	11865	0,189316	283034	4,5160436	15052	0,2401672	29879	0,4767444	29021	1,9280494
epb3	66450	13850	0,2084274	305618	4,5992175	17244	0,2595034	30895	0,464936	27641	1,6029344
epb4	64077	12755	0,1990574	294679	4,5988264	16247	0,2535543	29474	0,4599778	28991	1,784391
epb5	67284	12926	0,1921111	310445	4,6139498	16971	0,2522294	29976	0,4455145	29544	1,740852
jmp1	106154	13660	0,128681	472692	4,4528892	24637	0,2320873	51138	0,4817341	54780	2,223485
jmp2	134547	16466	0,122381	591050	4,3928887	29405	0,2185482	65564	0,4872944	69566	2,3657881
jmp3	100817	12528	0,1242648	438433	4,3488003	20993	0,2082288	48397	0,480048	50375	2,3996094
jmp4	81779	12182	0,1489624	353859	4,3270155	17054	0,2085376	39453	0,4824344	38384	2,250733
jmp5	125815	14487	0,1151453	537866	4,2750546	25041	0,1990303	61570	0,4893693	60205	2,404257
ce1	76871	12717	0,165433	361088	4,6973241	20916	0,2720922	33779	0,4394245	31914	1,5258176
ce2	56609	10125	0,1788585	265386	4,6880531	15343	0,2710346	24855	0,4390645	23678	1,5432445
ce3	88066	14280	0,1621511	413209	4,6920378	23722	0,2693662	38574	0,4380124	38067	1,6047129
ce4	36608	7362	0,2011036	167746	4,5822225	9196	0,2512019	16829	0,4597083	16655	1,8111135
ce5	36092	7104	0,1968303	165230	4,5780228	9055	0,2508866	16033	0,4442259	15966	1,7632247
bpg1	63363	10403	0,164181	292381	4,6143806	16430	0,2592996	28901	0,4561179	29326	1,7849057
bpg2	140653	18033	0,1282091	638805	4,541709	34132	0,2426681	64063	0,4554684	63010	1,8460682
bpg3	90402	13718	0,1517444	407718	4,5100551	21388	0,2365877	42010	0,4647021	39645	1,8536095
bpg4	83873	12819	0,1528382	371186	4,425572	18517	0,2207743	39532	0,4713317	39354	2,1252903
bpg5	97195	14272	0,1468388	435498	4,4806626	22832	0,2349092	45485	0,4679767	42622	1,8667659
vbi1	81710	11969	0,1464815	383743	4,6964019	21621	0,2646065	36622	0,4481948	37074	1,7147218
vbi2	90990	13637	0,1498736	428589	4,7102868	24727	0,2717551	40827	0,4486977	41653	1,6845149
vbi3	97772	13630	0,139406	457234	4,6765332	26428	0,2703023	45056	0,4608272	45541	1,7232102
vbi4	100142	13780	0,1376046	466227	4,655659	25895	0,2585828	46261	0,461954	46601	1,7996138
vbi5	96637	13144	0,1360142	453620	4,6940613	25766	0,2666267	44093	0,4562745	42757	1,6594349

ARXIU	num parau	noms		pronoms		adjectius		adverbis		verbs	
epb1	40410	9093	0,2250186	4262	0,1054689	2814	0,0696362	2215	0,0548132	8168	0,2021282
epb2	62673	13317	0,2124838	7069	0,1127918	3517	0,0561167	3941	0,0628819	11819	0,188582
epb3	66450	15474	0,2328668	6712	0,1010083	4470	0,0672686	3694	0,0555907	13336	0,2006922
epb4	64077	15144	0,2363407	5845	0,0912184	4099	0,0639699	3321	0,0518283	11133	0,1737441
epb5	67284	16001	0,2378129	6152	0,0914333	4719	0,0701355	3759	0,0558677	11677	0,1735479
jmp1	106154	20741	0,1953859	11764	0,1108201	6334	0,059668	6990	0,0658477	17741	0,1671251
jmp2	134547	27220	0,2023085	14782	0,109865	7478	0,0555791	8403	0,062454	21532	0,1600333
jmp3	100817	19800	0,1963954	12270	0,1217057	5031	0,0499023	7025	0,0696807	17615	0,1747225
jmp4	81779	17530	0,2143582	8864	0,1083897	4157	0,0508321	5159	0,0630847	15859	0,1939251
jmp5	125815	24983	0,1985693	14944	0,1187776	5888	0,0467989	8986	0,0714223	24770	0,1968764
ce1	76871	20074	0,2611388	4688	0,0609853	7160	0,0931431	3205	0,0416932	12257	0,1594489
ce2	56609	14229	0,2513558	3859	0,0681694	5394	0,0952852	2552	0,0450812	9539	0,1685068
ce3	88066	22308	0,25331	5651	0,0641678	7178	0,0815071	3900	0,044285	14075	0,1598233
ce4	36608	8393	0,2292668	2986	0,0815669	2841	0,077606	1652	0,0451267	6205	0,1694985
ce5	36092	8365	0,2317688	3078	0,0852821	2927	0,0810983	1700	0,0471019	6295	0,1744154
bpg1	63363	13961	0,2203336	6220	0,0981645	3327	0,052507	3181	0,0502028	11403	0,1799631
bpg2	140653	32024	0,2276809	13984	0,099422	8280	0,0588683	7930	0,0563799	26143	0,1858688
bpg3	90402	19500	0,2157032	10495	0,1160926	5436	0,0601314	5740	0,0634942	18394	0,2034689
bpg4	83873	18111	0,2159336	9610	0,114578	4561	0,0543798	4796	0,0571817	15414	0,1837779
bpg5	97195	21075	0,2168321	11267	0,1159216	5091	0,0523792	6216	0,0639539	20501	0,2109265
vbi1	81710	19610	0,2399951	6882	0,0842247	5423	0,0663689	3341	0,0408885	14043	0,1718639
vbi2	90990	22284	0,244906	6891	0,0757336	5994	0,0658754	3772	0,0414551	14998	0,1648313
vbi3	97772	23880	0,2442417	7911	0,0809127	6113	0,062523	3857	0,0394489	16171	0,165395
vbi4	100142	24438	0,2440335	8323	0,083112	5611	0,0560304	3801	0,0379561	17107	0,1708274
vbi5	96637	23486	0,2430332	7970	0,0824736	5806	0,0600805	4251	0,0439894	17828	0,1844842

ARXIU	num parau	verbs passat		verbs present		verbs auxiliars		conjuncions		determinants	
epb1	40410	1433	0,1754407	3170	0,3880999	313	0,0077456	2656	0,0657263	6337	0,1568176
epb2	62673	3641	0,3080633	3556	0,3008715	439	0,0070046	5078	0,0810237	9236	0,1473681
epb3	66450	2593	0,1944361	5036	0,3776245	528	0,0079458	4307	0,0648157	10434	0,1570203
epb4	64077	4458	0,4004312	2280	0,2047966	202	0,0031525	4732	0,0738487	10276	0,1603696
epb5	67284	4388	0,3757815	2650	0,2269419	235	0,0034927	4828	0,0717555	10302	0,1531122
jmp1	106154	6804	0,3835184	3536	0,1993123	881	0,0082993	9118	0,0858941	16648	0,1568288
jmp2	134547	8739	0,405861	3522	0,1635705	1073	0,0079749	11399	0,0847213	21296	0,1582793
jmp3	100817	4979	0,2826568	5355	0,3040023	960	0,0095222	8847	0,0877531	14200	0,1408493
jmp4	81779	4766	0,3005234	4205	0,2651491	546	0,0066765	7296	0,0892161	12451	0,1522518
jmp5	125815	8253	0,3331853	5071	0,2047235	985	0,007829	10930	0,0868736	18714	0,1487422
ce1	76871	3047	0,2485926	3891	0,3174513	328	0,0042669	5316	0,0691548	13381	0,1740708
ce2	56609	3601	0,3775029	1585	0,16616	318	0,0056175	3908	0,069035	9461	0,1671289
ce3	88066	4033	0,2865364	4582	0,3255417	411	0,004667	6011	0,0682556	15598	0,1771172
ce4	36608	1324	0,2133763	2590	0,4174053	250	0,0068291	2459	0,0671711	6558	0,1791412
ce5	36092	2677	0,4252581	1112	0,1766481	290	0,008035	2585	0,0716225	5782	0,1602017
bpg1	63363	4097	0,3592914	2728	0,2392353	596	0,0094061	4723	0,0745388	9864	0,1556744
bpg2	140653	8480	0,3243698	7437	0,2844739	1158	0,008233	9881	0,0702509	21422	0,1523039
bpg3	90402	5824	0,316625	4479	0,2435033	841	0,0093029	6955	0,0769341	13363	0,1478175
bpg4	83873	4897	0,3176982	4072	0,2641754	645	0,0076902	6909	0,0823745	11797	0,1406531
bpg5	97195	5811	0,2834496	5220	0,2546217	699	0,0071917	7606	0,0782551	14309	0,1472195
vbi1	81710	5745	0,4091006	1942	0,1382895	652	0,0079794	4783	0,0585363	14986	0,1834047
vbi2	90990	6474	0,4316576	1894	0,1262835	552	0,0060666	6203	0,0681723	16126	0,1772283
vbi3	97772	5686	0,3516171	4093	0,2531074	757	0,0077425	6046	0,0618377	18304	0,1872111
vbi4	100142	7509	0,4389431	2467	0,14421	742	0,0074095	6219	0,0621018	18691	0,186645
vbi5	96637	6929	0,3886583	2575	0,1444357	731	0,0075644	5595	0,0578971	18123	0,1875369

ARXIU	num parau	num frases	num parag	paraules per frase	paraules per paràgraf	frases per paràgraf
epb1	40410	1866	664	21,655949	60,858434	2,810241
epb2	62673	3438	1031	18,229494	60,788555	3,3346266
epb3	66450	4316	1483	15,3962	44,807822	2,9103169
epb4	64077	2849	1229	22,491049	52,13751	2,3181448
epb5	67284	2979	1226	22,586103	54,880914	2,4298532
jmp1	106154	3499	1566	30,338382	67,786718	2,234355
jmp2	134547	3771	1440	35,679395	93,435417	2,61875
jmp3	100817	3787	1729	26,621864	58,309427	2,1902834
jmp4	81779	3128	1705	26,144182	47,964223	1,8346041
jmp5	125815	4083	2005	30,814352	62,750623	2,036409
ce1	76871	2804	1986	27,414765	38,706445	1,4118832
ce2	56609	2012	1868	28,135686	30,304604	1,0770878
ce3	88066	3978	3437	22,13826	25,622927	1,1574047
ce4	36608	1691	1399	21,648729	26,167262	1,2087205
ce5	36092	1726	1683	20,910776	21,445039	1,0255496
bpg1	63363	2762	934	22,940985	67,840471	2,9571734
bpg2	140653	8637	2804	16,284937	50,161555	3,0802425
bpg3	90402	4521	1826	19,996019	49,508215	2,4759036
bpg4	83873	3935	1537	21,314612	54,569291	2,5601822
bpg5	97195	4981	1336	19,51315	72,750749	3,7282934
vbi1	81710	3904	1462	20,929816	55,889193	2,6703146
vbi2	90990	4079	1493	22,306938	60,944407	2,7320831
vbi3	97772	4894	1155	19,977932	84,651082	4,2372294
vbi4	100142	5137	1943	19,494257	51,539887	2,6438497
vbi5	96637	4744	1509	20,370363	64,040424	3,1438038

Resultats proves de significació estadística

Caràcters alfabètics

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,693595529	X			
CE	0,070776357	0,052758083	X		
BPG	0,72719998	0,398042248	0,075060249	X	
VBI	0,000487189	0,003342443	0,002817714	0,001749719	X

Caràcters especials

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,018389968	X			
CE	0,004318476	0,923176092	X		
BPG	0,146205215	0,278759142	0,170674206	X	
VBI	0,000472382	0,041365461	0,006399489	0,00192504	X

Caràcters d'espai

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,001289775	X			
CE	0,013523998	0,000247342	X		
BPG	0,161671262	0,00485045	0,000779898	X	
VBI	0,859986892	0,000576354	0,012718276	0,029522961	X

Caràcters de salt de línia

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,333610885	X			
CE	0,004452449	0,003443071	X		
BPG	0,658739704	0,646337879	0,009438427	X	
VBI	0,379676142	0,839587808	0,008264136	0,505150886	X

Caràcters en majúscula

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,029006964	X			
CE	0,596190367	0,024523969	X		
BPG	0,224091156	0,023589268	0,025340269	X	
VBI	0,043485035	0,107830363	0,016811062	0,015701883	X

Nombre total de punts

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,005618906	X			
CE	0,030654377	0,599008111	X		
BPG	0,250848973	0,025798979	0,148618619	X	
VBI	0,000714801	0,866870001	0,39722166	0,02062517	X

Nombre total de comes

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,196582112	X			
CE	0,07281494	0,121714655	X		
BPG	0,408759914	0,84796967	0,213249912	X	
VBI	0,040107312	0,007723059	0,105412013	0,013909647	X

Nombre total de guions

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,085053972	X			
CE	0,669931781	0,039720993	X		
BPG	0,015529033	0,928450889	0,061380662	X	
VBI	0,000159024	0,024113965	0,005515325	0,003371165	X

Riquesa lèxica

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,0016507	X			
CE	0,182033085	0,00288581	X		
BPG	0,000599781	0,035203249	0,041390913	X	
VBI	0,00212291	0,103721995	0,015309806	0,395107537	X

Mitjana de lletres per paraula

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,008826762	X			
CE	0,164264561	0,000084146	X		
BPG	0,243837048	0,000822969	0,002017049	X	
VBI	0,01234137	0,000300738	0,18556806	0,002592346	X

Paraules de més de sis lletres

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,009165983	X			
CE	0,140241567	0,000195574	X		
BPG	0,230674955	0,002674017	0,004006171	X	
VBI	0,023472218	0,000909474	0,422247601	0,008647436	X

Paraules d'entre una i tres lletres

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,028989263	X			
CE	0,067549758	0,00086517	X		
BPG	0,959438152	0,004546652	0,002201119	X	
VBI	0,317612324	0,001679609	0,029469693	0,003716942	X

Paraules gramaticals

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,001387032	X			
CE	0,242037038	0,00057243	X		
BPG	0,164521355	0,005453605	0,002847501	X	
VBI	0,459639074	0,000401968	0,284638302	0,014005468	X

Freqüència relativa de noms

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,006394616	X			
CE	0,172056894	0,008205852	X		
BPG	0,234229153	0,014409742	0,010022322	X	
VBI	0,040812018	0,000187067	0,76629179	0,000491033	X

Freqüència relativa de pronoms

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,067925737	X			
CE	0,028794546	0,001599578	X		
BPG	0,344899365	0,199298969	0,000738347	X	
VBI	0,022061708	0,000375283	0,125628139	0,002176407	X

Freqüència relativa d'adjectius

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,027846783	X			
CE	0,01714972	0,000092281	X		
BPG	0,059803726	0,338343589	0,001266802	X	
VBI	0,40108284	0,003818153	0,000268408	0,041038452	X

Freqüència relativa d'adverbis

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,021596958	X			
CE	0,003776096	0,000221638	X		
BPG	0,552014951	0,011518759	0,002208906	X	
VBI	0,000707515	0,000072084	0,020105161	0,002211974	X

Freqüència relativa de verbs

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,509857204	X			
CE	0,076396046	0,110019221	X		
BPG	0,628819726	0,106155105	0,01005698	X	
VBI	0,136801634	0,256993975	0,152239677	0,015363132	X

Freqüència relativa de verbs en passat

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,409460313	X			
CE	0,728020098	0,471922535	X		
BPG	0,619325664	0,380444568	0,845282118	X	
VBI	0,047804035	0,038514757	0,107832599	0,008319446	X

Freqüència relativa de verbs en present

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,172011423	X			
CE	0,766681131	0,2002814	X		
BPG	0,371677198	0,374937199	0,674348375	X	
VBI	0,015185967	0,01003216	0,06237728	0,026999379	X

Freqüència relativa de verbs auxiliars

EPB	JMP	CE	BPG	VBI

EPB	X				
JMP	0,041324854	X			
CE	0,993507486	0,105107814	X		
BPG	0,021252963	0,421030035	0,092444191	X	
VBI	0,252307762	0,230270577	0,144841647	0,093265655	X

Freqüència relativa de conjuncions

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,009325075	X			
CE	0,474842495	0,000155532	X		
BPG	0,281926656	0,001291029	0,031880962	X	
VBI	0,008796519	0,000340026	0,030202172	0,011610683	X

Freqüència relativa de determinants

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,473411765	X			
CE	0,002393775	0,016649743	X		
BPG	0,207606077	0,434160788	0,009210584	X	
VBI	0,000037424	0,002265243	0,024221961	0,000876759	X

Mitjana de paraules per frase

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,012064563	X			
CE	0,15164836	0,009499892	X		
BPG	0,966109338	0,018825517	0,122442498	X	
VBI	0,749406551	0,001818066	0,041011748	0,699064965	X

Mitjana de paraules per paràgraf

	EPB	JMP	CE	BPG	VBI
EPB	X				
JMP	0,133531733	X			
CE	0,000550894	0,006146026	X		
BPG	0,402925587	0,501520854	0,005026384	X	
VBI	0,343048085	0,798669179	0,008691333	0,632495113	X

Mitjana de frases per paràgraf

	EPB	JMP	CE	BPG	VBI
EPB	X				

JMP	0,000838936	X			
CE	0,001290644	0,002736028	X		
BPG	0,544131714	0,033027511	0,002364076	X	
VBI	0,38591083	0,00424625	0,00424625	0,782076166	X