L'ÚS DE LES COMBINACIONS DE SEQÜÈNCIES DE CATEGORIES GRAMATICALS COM A NOVA TÈCNICA DE COMPARACIÓ FORENSE DE TEXTOS ESCRITS

Sheila Queralt, Maria Spassova i Maria Teresa Turell
ForensicLab (Laboratori de Lingüística Forense)
Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra

1. INTRODUCCIÓ

En el marc de la lingüística forense, l'àrea de *determinació* (de l'autoria d'un o diversos textos dubitats entre diferents possibles autors) o *atribució* (d'un o diversos textos a un únic autor que pot ser sospitós d'haver comès un delicte) —distinció que proposem explícitament en aquest article— té una llarga tradició d'estudi. Altres lingüistes forenses han proposat altres termes com a identificació d'autoria (Tiersma i Solan 2004), reconeixement d'autoria (Hänlein 1999) o anàlisi d'autoria (Grant 2007 i 2008). En qualsevol cas, la comparació forense de textos escrits que pot conduir a una determinació o atribució d'autoria més fiable i rigorosa s'ha nodrit del concepte *estil idiolectal* proposat per Turell (2010), segons el qual aquest estil no tindria tant a veure amb tot allò que és sistemàtic en la llengua o dialecte sinó amb la manera com aquest sistema lingüístic, que aquesta persona comparteix amb més usuaris, és emprat distintivament, de forma única i irrepetible (Coulthard 2004), i a partir de la proposta de Halliday (1989) d'opcions i seleccions d'aquestes opcions.

Les tècniques de comparació de textos escrits emprades han estat diverses. Hi ha mesures estilomètriques de l'estil d'un individu a partir de diversos mètodes (Burrows et al. 1987, Holmes 1994, Baayen i van Halteren et al. 1996, Foster 2001, Holmes, Robertson i Paez 2001, Love 2002, Feiguina i Hirst 2007, Spassova i Turell 2007, Grant 2007): s'ha portat a terme la identificació d'estils idiolectal (Chaski 2001, Grant i Baker 2001), s'han usat mètodes estilístics (McMenamin 2001, 2002 i 2004) i, finalment, s'han aplicat tècniques desenvolupades a partir de l'anàlisi de vocabulari (Woolls i Coulthard 1998, Coulthard 2004, Turell 2004, Hoey 2005) considerant vocabulari nuclear, hàpax legòmena i dislegòmena, densitat i riquesa lèxica, i altres mesures, que es combinen amb tècniques derivades de la lingüística de corpus per tal d'establir la distribució de freqüència d'ús de determinades paraules en textos forenses dubtosos, textos del món real i textos de corpus de referència.

Alguns estudis forenses de determinació o atribució d'autoria (Spassova i Turell 2007, Grant 2007, Spassova i Grant 2008, Spassova 2009) i també no forenses (Love 2002, Feiguina i Hirst 2007) han demostrat que sembla haver-hi menys variació intraparlant o intraescriptor que interparlant o interescriptor. Al mateix temps, altres estudis més recents sobre el català i l'espanyol han confirmat, d'una banda, que l'estil idiolectal oral o escrit d'un individu no varia gaire per gènere textual, excepte en el cas del vocabulari i les frases fetes, i de l'altra, que aquest estil es manté bastant

© Queralt, S.; Spassova, M.S.; Turell, M.T.(2011)

estable al llarg del temps (Cicres 2007, Spassova i Turell 2007, Spassova 2009).

2. Objectius

Aquest article presenta una nova tècnica aplicada a la comparació forense de textos escrits basada en l'ús de les combinacions de seqüències de categories gramaticals emprades per un individu en la seva producció escrita com a marques d'autoria. Un altre objectiu metodològic que es pretén assolir és demostrar la necessitat de relacionar la lingüística forense amb diverses disciplines, com són la lingüística de corpus, la lingüística computacional i l'estadística. Aquesta interrelació de disciplines ha permès crear una metodologia efectiva per analitzar els textos escrits amb uns resultats més fiables i concloents, tot i que els textos de casos forenses reals presenten certes limitacions a l'hora d'emprendre l'anàlisi, com ara l'extensió (que acostuma a ser curta) i la variabilitat d'ús segons els gèneres textuals.

En els estudis de comparació forense de textos escrits es treballa amb dos tipus de corpus lingüístics: els textos dubtosos (o dubitats), l'autor dels quals és desconegut, i els textos no dubtosos (o indubitats), és a dir, mostres d'un autor o de diferents possibles autors. És imprescindible que els corpus de textos indubitats es composin d'un nombre de mostres suficient per reflectir l'estil idiolectal i la intravariabilitat de cada escriptor. No obstant això, abans d'assolir resultats concloents, la comparació forense de textos escrits s'ha de nodrir de la necessària experimentació en aquest camp amb textos del món real i textos provinents de corpus de referència, per tal d'obtenir freqüències de distribució poblacional de determinades variables escrites amb les quals es puguin comparar les freqüències d'ús que s'observen en els textos de casos forenses reals.

Així doncs, en aquest article es presenten resultats de diversos experiments portats a terme amb corpus escrits del món real, tot i que posteriorment la nova tècnica proposada s'ha aplicat també a textos de casos forenses reals amb bastant bons resultats, emprant aquests textos com a corpus de control.

3. Variables

A la dècada dels anys 80 es van començar a utilitzar mètodes estadístics d'anàlisi multivariant gràcies al desenvolupament de programaris estadístics com l'SPSS, SAS i VARBRUL, que faciliten el tractament d'un gran nombre de dades amb molta facilitat de càlcul. Simplificant molt, aquestes dades analitzades es divideixen en tres tipus de variables: la variable dependent, que constitueix l'objecte d'anàlisi; les variables independents (lingüístiques, socials, estilístiques), que són les hipòtesis formulades i afecten el comportament de la variable dependent, i les variables de control, que, com el seu nom indica, controlen l'investigador.

En la tècnica que s'ha desenvolupat en l'estudi que es presenta s'utilitza l'anàlisi lineal discriminant de variables dependents que es corresponen amb les seqüències de categories gramaticals, anomenades també *n-grames*, les quals representen una unitat de caràcter morfosintàctic. Per a l'anàlisi s'utilitzen dues classes de n-grames: [1] els bigrames, formats per seqüències de dues categories gramaticals, i els trigrames, formats per tres categories gramaticals. A continuació es presenta un exemple de bigrames en la taula 1 i un altre de trigrames

[1] S'adopten els bigrames i els trigrames com a variables d'estudi perquè diversos estudis preliminars han demostrat que són els més discriminants i significants.

en la taula 2:

Taula 1. Exemple de bigrames

1	2	3	4	5	Bigrama
La	Núria				AS.NS
	Núria	és			NS.V3S
		és	una		V3S.JS
			una	heroïna	JS.NS

Taula 2. Exemple de trigrames

1	2	3	4	5	Trigrama
La	Núria	és			AS.NS.V3S
	Núria	és	una		NS.V3S.JS
		és	una	heroïna	V3S.JS.NS

Com a variables independents s'han emprat el gènere biològic de l'autor, el temps en què s'han produït els textos i el gènere textual, i com a variables de control, la varietat lingüística, que en aquest cas correspon a l'espanyol peninsular.

4. Hipòtesis generals i específiques

Les hipòtesis que la lingüística forense, i més específicament l'àrea de determinació i atribució d'autoria, està en curs d'avaluar són les següents:

1) Que la variació intraautor (dins del mateix escriptor) és menor que la variació interautor (entre escriptors).

2) Que el temps transcorregut entre una producció i una altra no afecta de manera significativa l'estil idiolectal d'un autor.

3) Que l'estil idiolectal d'un autor resta relativament constant en els diferents gèneres i registres de les seves produccions.

5. Metodologia

5.1. Criteris de selecció

L'estudi que es presenta en aquest article té a veure amb textos del món real. S'ha de recordar que perquè el corpus reflecteixi l'estil idiolectal i la intravariabilitat de cada escriptor, aquest corpus ha d'estar composat d'un nombre de mostres suficientment alt. El corpus s'ha confegit tenint en compte les hipòtesis que es volen avaluar, a partir de les variables independents següents: el temps i el gènere textual. Per tal de validar la primera hipòtesi, d'una banda, s'han creat dos grups d'autors: el grup femení compost per tres autores i el grup masculí compost per tres autors. D'altra banda, s'ha controlat la varietat lingüística per tal de crear un grup homogeni i s'han escollit només autors de l'espanyol peninsular.

També s'ha tingut en compte la variable independent temps i s'han obtingut mostres de cada autor en dos moments de recollida de dades per tal de poder validar la segona hipòtesi, que planteja que el temps no influeix de manera significativa en l'estil d'un autor. El temps transcorregut entre una producció i l'altra és d'uns deu anys aproximadament en el cas de la novel·la (N), i de sis a vuit anys en el cas dels articles d'opinió (AO).

Per tal de validar la tercera hipòtesi s'han triat dos tipus de gèneres: el literari i el periodístic. Les mostres del gènere literari es composen de cinc fragments de sis-centes paraules d'una novel·la de cada autor i les mostres del gènere periodístic consten de tres a cinc articles d'opinió d'unes tres-centes paraules cada un.

Taula 3. Corpus informants

Autor	Gènere biològic	Varietat lingüística
EL	D	espanyol peninsular
LE	D	espanyol peninsular
RM	D	espanyol peninsular
XM	H	espanyol peninsular
AP	H	espanyol peninsular
JM	H	espanyol peninsular

Taula 4. Nombre de mostres i nombre de paraules per autor en N

Autor	Nombre de mostres	Nombre de paraules
EL	5	~3000
LE	5	~3000
RM	5	~3000
XM	5	~3000
AP	5	~3000
JM	5	~3000

Taula 5. Número de mostres i nombre de paraules per autor en AO

Autor	Nombre de mostres	Nombre de paraules
EL	3	~900
LE	3	~900
RM	2	~600
XM	3	~900
AP	4	~1200
JM	3	~900

5.2. Processos de tractament computacional del corpus

L'anàlisi de seqüències de categories gramaticals requereix un tractament computacional del corpus mitjançant eines de processament de textos. En primer lloc, el corpus ha de ser anotat, és a dir, s'ha d'assignar una etiqueta —categoria gramatical— a cada unitat. Aquest procés és automàtic i dóna com a resultat un text anotat amb totes les categories gramaticals, com el que es mostra a la taula 6.

Taula 6. Exemple d'anotació d'etiquetes

Mot	Etiqueta
La	AS
Núria	NS
és	V3S
una	JS
heroïna	NS

Malgrat que els sistemes d'anotació automàtics disponibles actualment són molt competitius i ofereixen molt bons resultats, les etiquetes resultants han de ser revisades manualment ja que són la base de la nostra anàlisi i constitueixen la variable dependent.

5.3. Extracció de dades

El segon procés que defineix aquesta tècnica és l'extracció de dades. En aquesta etapa de l'anàlisi s'extreuen les seqüències de categories gramaticals (bigrames i trigrames) del corpus etiquetat. Aquesta extracció es fa amb un programari informàtic específic, Legolas 2.0, dissenyat pel ForensicLab per assolir aquesta tasca. Posteriorment es procedeix a estudiar quines són les seqüències de categories gramaticals que resulten més discriminants en la comparació forense de textos mitjançant l'anàlisi estadística.

6. ANÀLISI I EXPERIMENTS APLICATS A CASOS CONCRETS

La lingüística forense parteix de la premissa que l'estil idiolectal d'un individu es únic i irrepetible, però que al mateix temps, exhibeix una certa variabilitat. Aquesta variació s'explica perquè hi ha una variació inherent a totes les llengües que es reflecteix en l'ús lingüístic. No obstant això, tal com s'ha exposat en l'apartat sobre hipòtesis, es parteix de la base que la variació intraautor sembla que és menor que la variació interautor. En aquest article es presenten diversos estudis, l'objectiu dels quals és estudiar els dos tipus de variació esmentats, mitjançant les seqüències de categories gramaticals. Les variables independents que s'han tingut en compte són el temps i el gènere textual.

6.1. La variació intraescriptor

Per tal d'avaluar la hipòtesi de la variació intraautor i, a tall d'exemple, s'ha aplicat l'anàlisi que es proposa als textos de novel·la (N) dels sis autors considerats. Els resultats es poden observar a la taula 7.

Taula 7. Variació intraautor en N

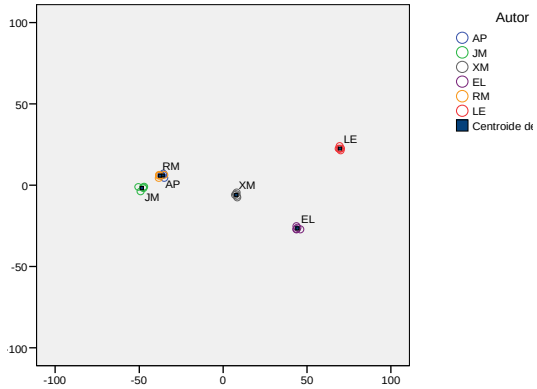
Autor	Bigrames	Trigrames
AP	13 %	22 %
JM	13 %	0 %
XM	18 %	9 %
EL	9 %	9 %
RM	9 %	6,8 %
LE	4,5 %	4,5 %

Tal com il·lustra la taula 7, el percentatge d'etiquetes que varien al llarg del temps és baix, cosa que confirma la hipòtesi segons la qual la variació intraparlant és baixa pel que fa a la variable estudiada. És a dir, que l'ús d'aquestes seqüències de categories gramaticals es manté estable al llarg del temps. Les estructures lingüístiques més constants al llarg del temps són els sintagmes preposicionals, els sintagmes adjectivals, els sintagmes nominals compostos i els sintagmes amb infinitius. En canvi, l'ús dels pronoms sol presentar una major variabilitat.

6.2. La variació interescriptor

Per avaluar la hipòtesi de la variació interautor es van fer diversos estudis en ambdós gèneres textuals mitjançant, primer, els bigrames i, posteriorment, els trigrames. Els resultats de l'anàlisi del gènere novel·la en temps aparent (TA) mitjançant bigrames es mostren en el gràfic 1.

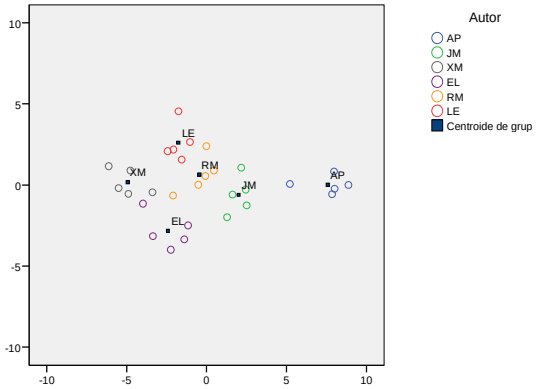
Gràfic 1. Variació interautor en N en TA mitjançant bigrames



En el gràfic 1 es pot observar com la tècnica d'anàlisi mitjançant bigrames agrupa clarament els textos de cada autor, ja que queden concentrats en un centroide i, en la majoria dels casos, clarament separats entre autors en els diferents quadrants del gràfic, tot i que RM, AP i JM se situen més a prop l'un de l'altre. La classificació és correcta ja que hi ha un encert del 100 % dels casos, amb una validació creuada [2] del 80 %.

El gràfic 2, que es presenta a continuació, mostra els resultats de l'anàlisi de bigrames en els fragments de novel·les escrites deu anys després (TR) pels mateixos autors.

Gràfic 2. Variació interautor en N en TR mitjançant bigrames

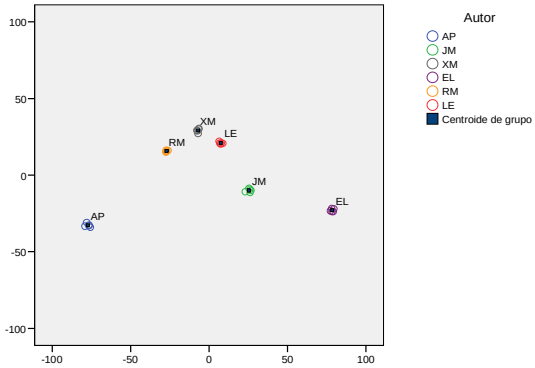


[2] La validació creuada ens permet avaluar l'estabilitat o validesa dels resultats si es modifiqués la mostra.

Es pot observar que, tant en TA com en TR, l'anàlisi de seqüències de categories gramaticals amb els bigrames resulta tenir un gran potencial discriminatori ja que els textos de cada autor es concentren al voltant del seu centroide i separats entre autors en diferents quadrants del gràfic.

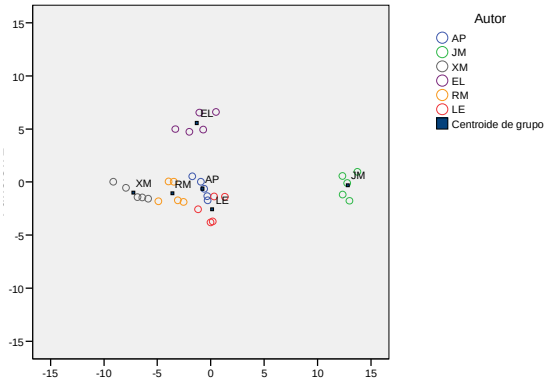
En el gràfic 3 es va dur a terme la mateixa anàlisi però amb seqüències de tres categories gramaticals (trigrames). En aquest cas, els textos no apareixen tan clarament concentrats al voltant del seu centroide però queda igualment delimitada una zona per a cada autor. L'encert de classificació en aquest cas és d'un 96,7 % i una validació creuada del 76,7 %.

Gràfic 3. Variació interautor en N en TA mitjançant trigrames



En l'anàlisi en temps real feta amb textos produïts deu anys després (gràfic 4) s'observa una classificació menys dispersa dels textos respecte del seus centroides i una situació diferenciada dels autors JM i AP respecte del temps aparent, que indicaria una certa variabilitat intraautor. En aquest estudi s'obtenen uns resultats de classificació similars als obtinguts en temps aparent, és a dir, un 100 % d'encert i una validació creuada del 80 %.

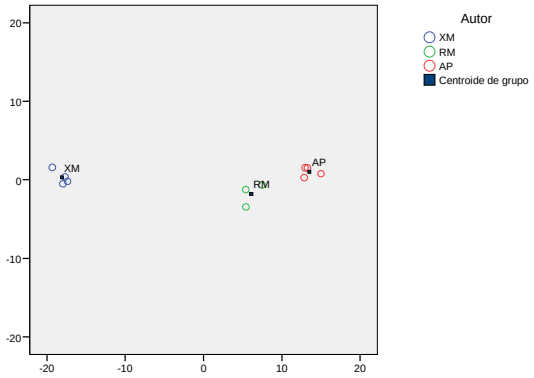
Gràfic 4. Variació interautor en N en TR mitjançant trigrames



Per tant, es pot concloure que en els textos de novel·la, els bigrames semblen tenir un potencial discriminatori més alt que els trigrames.

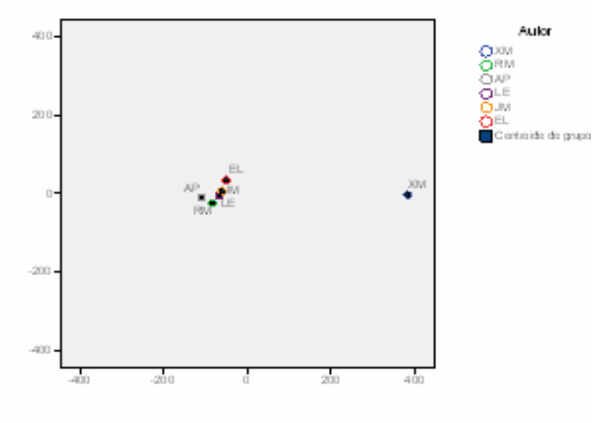
En aquest apart es presenten els resultats de l'anàlisi dels articles d'opinió (AO), en primer lloc, amb bigrames i, posteriorment, amb trigrames. Les anàlisis en TA, tant en bigrames com en trigrames, només es van poder fer amb tres autors perquè no va ser possible aconseguir suficients mostres dels altres autors en dos temps diferents. En el gràfic 5 s'observa la gran diferència existent entre cada autor, ja que els centroides apareixen molt distanciats els uns dels altres, sobretot, l'autor XM. L'encert de classificació és d'un 100 % i la validació creuada d'un 90,9 %.

Gràfic 5. Variació interautor en AO en TR mitjançant bigrames amb tres autors



En el gràfic 6, els textos de cada autor es concentren al voltant del seu centroide. No obstant això, tot i que els centroides d'alguns autors apareixen força pròxims, en cap cas es produeix un cavalcament entre textos de diversos autors. En aquest cas s'obté un encert de classificació del 100 % i una validació creuada del 88,9 %.

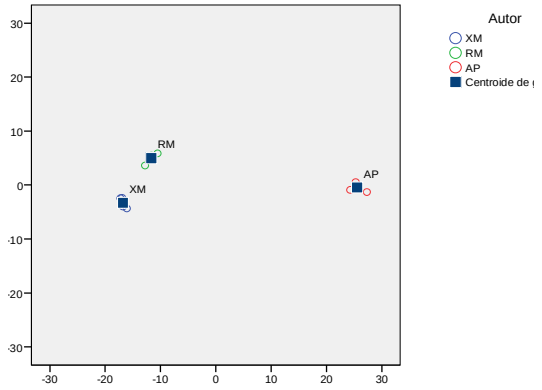
Gràfic 6. Variació interautor en AO en TA mitjançant bigrames amb sis autors



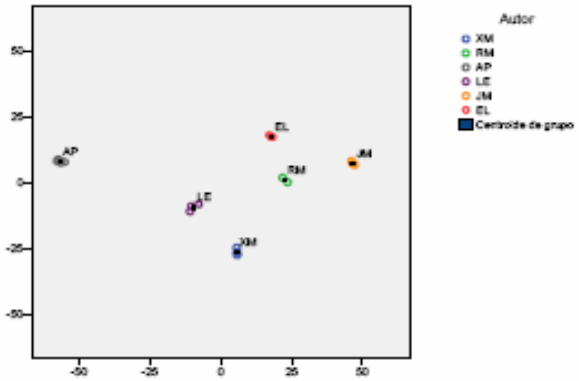
Per tant, podem concloure que en el cas de l'anàlisi amb bigrames, la modificació del nombre d'autors, tal com indiquen les validacions creuades, no modifiquen els resultats d'encert en la classificació, i per tant, l'anàlisi mitjançant seqüències de dues categories gramaticals (bigrames) sembla que és estable i robusta encara que es modifiqui el nombre d'autors.

Pel que fa a l'anàlisi amb seqüències de tres categories gramaticals (trigrames), d'una banda, els resultats mostren que hi ha encert en les anàlisis, tant en TA com en TR. D'altra banda, els resultats amb els trigrames tampoc no semblen veure's afectats per canvis en la grandària de la mostra. Tant en la mostra de tres autors (gràfic 7) com en la de sis autors (gràfic 8), l'encert és aproximadament del 100 % i s'obté una validació creuada superior al 90 %. Els centroides dels autors apareixen molt distanciats els uns dels altres i tots els textos de cada autor s'agrupen al voltant del seu centroide.

Gràfic 7. Variació interautor en A0 en TR mitjançant trigrames amb tres autors



Gràfic 8. Variació interautor en A0 en TA mitjançant trigrames amb sis autors



7. Conclusions

Per tant, els resultats d'aquestes anàlisis permeten concloure que la comparació forense de textos escrits mitjançant les seqüències de categories gramaticals, ja siguin bigrames com trigrames, és una tècnica que ofereix bastants bons resultats de classificació.

L'anàlisi de la variació intraautor permet concloure que els bigrames i els trigrames més utilitzats pels autors varien molt poc al llarg del temps, exactament, entre un 22 % i un 0 %. I, per tant, aquest resultat també permet concloure que les seqüències més freqüentment emprades per cada autor es mantenen al llarg del temps.

Pel que fa a l'anàlisi de la variació

interparlant duta a terme amb textos tant de novel·la com d'articles d'opinió, ha quedat demostrat que els n-grames classifiquen correctament un alt percentatge de textos al voltant del seu autor, concretament amb un encert de classificació entre el 90 % i el 100 % i amb una validació creuada entre un 70 % i un 94,4 %. Aquests resultats també permeten concloure que aquest alt percentatge d'encerts no es veu afectat per la grandària de la mostra.

A més a més, les anàlisis dutes a terme tenint en compte el gènere textual dels textos que composen el corpus d'anàlisi mostren que la bona capacitat classificatòria de les seqüències de categories gramaticals no es veu afectada per aquesta variable independent.

Finalment, aquests estudis permeten confirmar que l'interès i els beneficis dels resultats d'aquesta tècnica basada en seqüències de categories gramaticals (n-grames) són clars i molt útils per a la comparació forense de textos escrits i donen lloc a una determinació o atribució forense d'autoria més fiable. Sens dubte caldrà continuar avaluant aquesta tècnica amb un corpus més ampli d'escriptors i de mostres.

8. Referències bibliogràfiques

BAAYEN, R.H. i VAN HALTEREN, H. et al. (1996). «Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution». *Literary and Linguistic Computing*, 11, 3, 121-131.

BURROWS, J. et al. (1987). *Computation into Criticism: A study of Jane Austin's Novels and an Experiment in Method*. Oxford: Clarendon Press.

CHASKI, C. E. (2001). «Empirical evaluations of language-based author identification techniques. Forensic Linguistics». *Forensic Linguistics. The International Journal of Speech, Language and the Law*, 8, 1, 1-65.

CICRES BOSCH, J. (2007). *Aplicació de l'anàlisi de l'entonació i de l'alineació tonal a la identificació de parlants en fonètica forense*. Tesi doctoral sense publicar. Universitat Pompeu Fabra.

COULTHARD, R. M. (2004). «Author identification, idiolect, and linguistic uniqueness». *Applied Linguistics*, 25, 4, 431-447.

FEIGUINA, O. i HIRST, G. (2007). «Authorship attribution for small texts literary and forensic experiments.» Estudi presentat a International Workshop on Plagiarism Analysis, Authorship Identification and Near-Duplicate Detection. 30th Annual International ACM SIGIR Conference (SIGIR '07).

FOSTER, D. (2001). *Author Unknown: on the Trail of Anonymous*. 2a ed. London: Macmillan.

GRANT, T. (2007). «Quantifying evidence in forensic authorship analysis». *Forensic Linguistics. The International Journal of Speech, Language and the Law*, 14, 1, 1-25.

GRANT, T. (2008). «Approaching questions in forensic authorship analysis». J. Gibbons et al. (eds.), *Dimensions of Forensic Linguistics*, 215-229. Amsterdam/Philadelphia: John Benjamins.

GRANT, T. i BAKER, K. (2001). «Identifying reliable, valid markers of authorship: a response to Chaski». *Forensic Linguistics. The International Journal of Speech, Language and the Law*, 8, 1, 66-79.

HALLIDAY, M.A.K. (1989). *Language, context and text. Aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

HÄNLEIN, H. (1999). *Studies in Authorship Recognition: A Corpus-Based Approach*. Frankfurt: Peter Lang.

HOEY, M. (2005). Lexical priming: A new theory of words and language. Abingdon: Routledge.

HOLMES, D. I. (1994). «Authorship attribution». *Computers and the Humanities*, 28, 2, 87-106.

HOLMES, D. I., ROBERTSON, M. i PAEZ, R.

(2001). «Stephen Crane and the New-York Tribune: A case study in traditional and non-traditional authorship attribution». *Computers and the Humanities*, 35, 3, 315-331.

LOVE, H. (2002). *Attributing authorship: An introduction*. Cambridge: Cambridge University Press.

MCMENAMIN, G. (2001). «Style markers in authorship studies». *Forensic Linguistics. The International Journal of Speech, Language and the Law*, 8, 2, 93-97.

MCMENAMIN, G. (2002). *Forensic Linguistics: Advances in Forensic Stylistics*. London: CRC Press.

MCMENAMIN, G. (2004). «Disputed authorship in US Law». *Forensic Linguistics. The International Journal of Speech, Language and the Law*, 11, 1, 73-82.

SPASSOVA, M. i GRANT, T. (2008). «Categorizing Spanish written texts by author gender and origin by means of Morpho-Syntactic Trigrams: some observations on method's feasibility of application for linguistic profiling». Estudi presentat a Curriculum, language and the law Inter-University Centre, University of Zagreb, Dubrovnik (Croàcia).

SPASSOVA, M. S. (2009). *El potencial discriminatorio de las secuencias de categorías gramaticales en la atribución forense de autoría de textos en español*. Tesi doctoral sense publicar. Universitat Pompeu Fabra.

SPASSOVA, M. S. i TURELL, M. T. (2007). «The use of morpho-syntactically annotated tag sequences as forensic markers of authorship attribution». TURELL, M.T., SPASSOVA, M. S. i CICRES, J. (ed.), *Proceedings of the Second European IAFL Conference on Forensic Linguistics/Language and the Law*, 229-237. Barcelona: Publicacions de l'IULA.

TIERSMA, P. M. i SOLAN, L. M. (2002). «The linguist on the witness stand: forensic linguistics in American courts». *Language*, 78, 221.

TURELL, M. T. (2004a). «Textual kidnapping revisited: the case of plagiarism in literary translation». *Forensic Linguistics. The International Journal of Speech, language and the Law. Forensic Linguistics*, 11, 1, 1-26.

TURELL, M. T. (2004b). «The disputed authorship of electronic mail: linguistic, stylistic and pragmatic markers in short texts». Estudi presentat a 1st European IAFL Conference on Forensic Linguistics/language and Law. Cardiff University.

TURELL, M. T. (2010a). «Los retos de la lingüística forense en el siglo XXI: In Memoriam Enrique Alcaraz Varó». Homenaje a Enrique Alcaraz Varó. *Revista Alicantina de Estudios Ingleses*.

TURELL, M. T. (2010b). «The use of textual, grammatical and sociolinguistic evidence in forensic text comparison». *Forensic Linguistics. The International Journal of Speech, Language and the Law*, 17, 2, 211-250.

WOOLLS, D. i COULTHARD, R. M. (1998). «Tools for trade». *Forensic Linguistics. The International Journal of Speech, Language and the Law*, 5, 1, 33-57.