# La identificación d'autoría n'asturianu / *Autorship identification in Asturian Language*

Pablo Suárez García
Data, Decisions and language - Daedalus[1]

**Resume:** L'asturianu sumóse nestos caberos años al carapiellu de les llingües minorizaes qu'aprovecharon la evolución teunolóxica actual col envís de meyorar la so situación, que yera de mano precaria. En particular, el desendolcu per parte l'Academia de la Llingua Asturiana (ALLA) de la coleición *Termast* (TERMAST), el desenvolvimientu per parte la Universidá d'Uviéu de proyeutos tan interesantes como *Eslema* (ESLEMA) o ETLEN (ETLEN), o la creación del grupu *Softastur* (SOFTASTUR) o la rede social *Ximiélgame* (XIMIÉLGAME) son bones amueses d'esti fechu ensin precedentes. Nesta mesma llinia y, en particular, dientro la faza que suel conocese como Llingüística Computacional o Inxeniería Llingüística, pretendemos amosar l'aplicación de delles téuniques d'análisis d'autoría a dellos casos concretos de la lliteratura asturiana

***Pallabres clave:*** nueves teunoloxíes, llingüística computacional, llingua asturiana, termast, etlen, softastur, ximiélgame.

**Abstract:** Asturian Language has in recent years joined the group of minority languages who took advantage of current technological development in order to improve their situation, which was initially precarious. In particular, the development by the Asturian Language Academy (ALLA) of the *Termast* collection (TERMAST), the development by the University of Uviéu in interesting projects such as *Eslema* (ESLEMA) or ETLEN (ETLEN), or the creation of the *Softastur* group (SOFTASTUR) or the social network *Ximiélgame* (XIMIELGAME) are good examples of this unprecedented event. Within this area and, in particular, within the sector that is generally known as Computational Linguistics and Language Engineering, we aim to demonstrate the application of some authorship analysis techniques to specific cases of Asturian literature.

***Keywords:*** new technologies, computational linguistics, Asturian Language, termast, etlen, softastur, ximiélgame.

[1] www.daedalus.es, *spin-off* de la Universidá Politécnica de Madrid. Email: psuarez@daedalus.es.

## 1. Entamu

La idea d'aplicar métodos cuantitativos (non siempre matemáticos) al problema de reconocer l'autor d'un testu anónimu o apócrifu (atribución d'autoría, estilometría) nun ye nuevu, sinón que remonta pelo menos a la fin del sieglu XIX, a los trabayos de Morgan y Mendenhall, que propunxeron calcular les llonxitúes medies de pallabres en trabayos d'estremaos autores y comparales col oxetu d'establecer l'autoría de dellos otros (Basile 2008: 1). La estilometría tien, darréu, una hestoria perllarga, y a lo llargo décades científicos de mui diferentes campos tuvieron interesaos nesti tema.

Dende los primeros trabayos de Mendenhall fasta l'actualidá hebo un desplazamientu de los métodos emplegaos na atribución, moviéndose l'interés dende indicadores sofitaos en pallabres (pernaturales, desque les pallabres son en ciertu sen los componentes básicos del llinguaxe), a métodos nos que nun se tienen en cunta estructures morfosintáutiques del testu, sinón qu'esti ye tratáu como una cenciella secuencia de símbolos (Markov, Shannon) (Basile 2008: 2). En particular, ún de los elementos más emplegaos na actualidá son los llamaos n-grames, que sedrán descritos detenidamente más alantre nesti trabayu. Too ello permite, ente otres coses, independizar, polo menos parcialmente, los procedimientos d'identificación d'autoría de la llingua concreta cola que se ta trabayando. En cualquier casu, la estadística sobro pallabres (fundamentalmente les pallabres 'funcionales' de la llingua), sigue siendo un métodu auxiliar (en conxunción con otros) perfeutamente válidu anguaño.

Sofitándonos en delles téuniques d'estilometría emplegaes na actualidá, el presente estudiu[2] tien por oxetu, ente otres coses:

1) Valorar la viabilidá de la identificación d'autor dientro'l corpus testual de poemes n'asturianu en llinia *Caveda y Nava* (CAVEDAYNAVA)[3]. Ello inxer la valoración d'estremaes téuniques y la so comparanza col envís de meyorar la estimación per aciu la variación de los parámetros afayadizos.

2) Estudiar la posibilidá d'espardimientu d'esti procedimientu d'identificación d'autoría a los mensaxes na rede social asturiana *Ximiélgame* (XIMIÉLGAME) que comparten col anterior corpus la dificultá inherente a los 'testos percurtios'.

3) Aplicar estremaes téuniques d'autoría col envís d'algamar dalguna conclusión d'interés en dellos casos polémicos dientro la hestoria de la lliteratura asturiana, como son: l'autoría'l poema *Píramo y Tisbe* del sieglu XVII (Antón de Ma-

[2] Esti estudiu presentóse parcialmente como comunicación en Congresu de Filoloxía Románica *Filología Románica Hoy* entamáu pola Universidá Complutense de Madrid y la *Revista de Filología Románica* los díes 3-5 de payares de 2011 na Facultá de Filoloxía d'esta universidá.

[3] Encamentamos al llector la consulta de los autores y poemes nomaos a lo llargo esti trabayu na escelente HLLA asoleyada pola ALLA.

rirreguera / Benito de la Uxa), o'l poema *Al niñín Xesús* del sieglu XIX (Fernández de Castro / Xuan María Acebal). O tamién la deteición de posibles fragmentos non orixinales del autor en dellos poemes de Marirreguera (sieglu XVII), pente medies de téuniques de plaxu intrínsecu.

4) Valorar la meyora que nes téuniques de deteición d'autoría ye a ufiertar l'emplegu del analizador sintáuticu de llingua asturiana *Eslema* (ESLEMA-ANALIZADOR) desendolcáu pola Universidá d'Uviéu nestos caberos años.

Los apartaos qu'inxerimos darréu tendrán el siguiente envís. N'apartáu 2 vamos faer una pequeña llista de trabayos anteriores d'otros autores que tienen un calter asemeyáu al del nuesu artículu. N'apartáu 3 vamos centranos na descripción del esquema xeneral d'un sistema d'identificación d'autoría. N'apartáu 4 vamos faer una esbilla de les principales carauterístiques qu'un sistema d'esti calter emplega. N'apartáu 5, falaremos de lo que ye propiamente'l clasificador de l'arquiteutura, y fadremos tamién una esbilla d'ún d'ellos. N'apartáu 6 dase cuenta del *Casu d'estudiu*, esto ye, de les estremaes xeres qu'a mou d'esperimentos foron feches acordies colos oxetivos afitaos enantes. Finalmente, n'apartáu 7 preséntense les conclusiones del trabayu.

## 2. Trabayos anteriores d'esti calter

Trabayos asemeyaos al d'esti artículu puen atopase en Basile (2008), Belabbes (2008), Girón (2005, 2010), Holmes (1992), Hoorn (1999), Malyutov (2007), McCombe (2002). Trabayu específicu d'una llingua concreta ye, por exemplu, Tas (2007), que tien por oxetu d'estudiu'l turcu. En particular, estudios aplicaos a testos curtios, qu'enzarren, evidentemente, una mayor dificultá, fundamentalmente referíos a correos electrónicos, atopámoslos en Corney (2002, 2003, 2007), De Vel (2007), y Fissette (2010). Una clasificación llingüística sofitada en métodos de deteición d'autoría atopámosla en Benedetto (2002). Aplíquense téuniques de deteición de plaxu en Kimler (2002), Seaward (2009), Suárez (2010). L'usu d'anotación sintáutica (usu d'un analizador sintáuticu) como mediu de meyora d'un sistema d'atribución d'autoría atopámoslu en Van Halteren (1996).

## 3. Esquema xeneral d'un sistema d'identificación d'autoría

Tou sistema d'identificación d'autoría inxer davezu dos sosistemes principales:

1. Sosistema d'estraición de carauterístiques.
2. Sosistema d'entrenamientu-clasificación.

El sosistema d'estraición obtién les carauterístiques de cada testu disponible del llamáu corpus d'entrenamientu. Una máquina d'entrenamientu entrénase darréu con esta información. Finalmente, un nuevu testu sedrá clasificáu pol sosistema de clasificación a partir de la información colo que s'entrenó y les carauterístiques estrayíes del propiu testu. Nos apartaos que vienen darréu vamos falar tanto de la esbilla de carauterístiques como del sistema d'entrenamientu-clasificación.

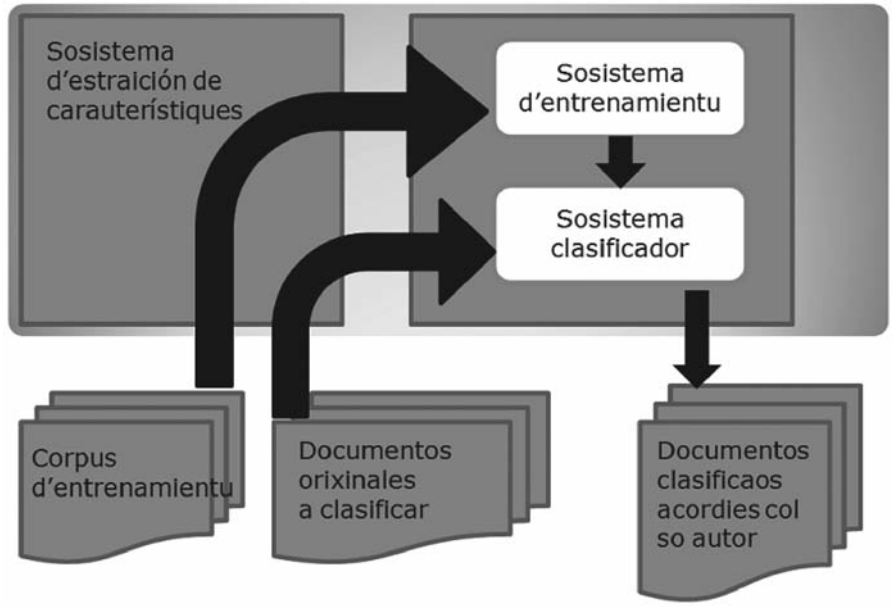


**Figura 1.** Esquema xeneral d'un sistema d'identificación d'autoría.

## 4. Esbilla de carauterístiques

Lo primero que necesita un sistema de clasificación ye la determinación d'un carapiellu de carauterístiques que puedan estrayese del testu y qu'espeyen de dalguna manera, y acordies col nuesu modelu, l'estilu de cada autor. Darréu falamos de les carauterístiques esbillaes por nós, y que son les típiques de la bibliografía actual.

### *4.1. N-grames*

Los n-grames son soconxuntos de n elementos consecutivos d'un conxuntu *T* formáu por una secuencia d'estos elementos. Nel nuesu contestu, el conxuntu *T* nun ye otro qu'un fragmentu de testu. Por *elementos* entiéndense davezu dos coses estremaes: 1) caráuteres; y 2) pallabres. Por poner un exemplu, seya T={*Po-*

*la mor d'aquel brebaxe aciar que-yos dio Circe augusta*[4]}. Si consideramos una implementación per aciu de caráuteres, y afitamos n=3, por exemplu, tendremos que'l conxuntu de n-grames ye: P={«Pol», «ola», «la », «a m», « mo», «mor», «or », etc.}, au se tán teniendo en cuenta los espacios en blanco. Si consideramos les pallabres, y afitamos, por exemplu, n=2, tenemos P={«Pola mor», «mor d'aquel», «d'aquel brebaxe», «brebaxe aciar», «aciar que-yos», «que-yos dio», «dio Circe», «Circe augusta»}. De mano, entiéndese siempre una implementación pente medies de caráuteres.

Demuéstrase que la frecuencia d'estos n-grames carauteriza de dalguna manera al conxuntu *T* (al testu) del que s'estrái. Estes frecuencies son útiles, darréu, como elementos d'identificación d'autoría. Son, de dalguna manera, una xeneralización de les cencielles frecuencies de caráuteres yá emplegaes d'antiguo n'escenarios rellacionaos cola encriptación y l'autoría.

### *4.2. Carauterístiques d'estilu*

Les principales carauterístiques d'estilu emplegaes na bibliografía queden reflexaes nos apartaos que s'ufren darréu (Corney 2003: 147-156). Son valores numbéricos que faen por recoyer de dalgún xeitu l'estilu d'un autor determináu. Toes estes carauterístiques foron implementaes por nós nel nuesu algoritmu y puestes a prueba darréu.

#### *4.2.1. Carauterístiques sofitaes en caráuteres (CSC).*

| ***Escandayu*** | ***Descripción-cálculu*** |
|---|---|
| C0 | Númberu de caráuteres en pallabres |
| C1 | Númberu de caráuteres alfabéticos per númberu total de caráuteres |
| C2 | Númberu de caráuteres en mayúscula per númberu total de caráuteres |
| C3 | Númberu de díxitos per númberu total de caráuteres |
| C4 | Númberu d'espacios en blanco per númberu total de caráuteres |
| C5 | Númberu d'espacios per númberu total de caráuteres |
| C6 | Númberu d'espacios per númberu total d'espacios en blanco |
| C7 | Númberu de tabuladores per númberu total de caráuteres |
| C8 | Númberu de tabuladores per númberu total d'espacios en blanco |
| C9 | Númberu de caráuteres de puntuación per númberu total de caráuteres |

[4] *Odisea*, Cantu X, torna de Xosé Gago, Uviéu, Trabe.

### *4.2.2. Carauterístiques sofitaes en pallabres (CSP).*

| ***Escandayu*** | ***Descripción-cálculu*** |
|---|---|
| P0 | Tamañu mediu de pallabra |
| P1 | Númberu de pallabres estremaes per númberu total de pallabres (bayura de vocabulariu) |
| P2 | Númberu de pallabres funcionales per númberu total de pallabres |
| P3 | Númberu de pallabres curties (llonxitú ≤ 3) per númberu total de pallabres |
| P4 | Númberu d'*hapax legomena* ente númberu total de pallabres |
| P5 | Númberu d'*hapax legomena* ente númberu de pallabres estremaes |
| P6 | Númberu d'*hapax dislegomena* ente númberu total de pallabres |
| P7 | Carauterística R de Guirad (vei Anexu) |
| P8 | Carauterística C de Herdan (vei Anexu) |
| P9 | Carauterística V de Herdan (vei Anexu) |
| P10 | Carauterística K de Rubet (vei Anexu) |
| P11 | Carauterística A de Maass (vei Anexu) |
| P12 | Carauterística U de Dugast (vei Anexu) |
| P13 | Midida de Janenkov y Neistoj (vei Anexu) |
| P14 | Carauterística W de Brunet (vei Anexu) |
| P15 | Carauterística S de Sichel (vei Anexu) |
| P16 | Carauterística H de Honore (vei Anexu) |
| P17 | Carauterística K de Yule (vei Anexu) |
| P18 | Carauterística D de Simpson (vei Anexu) |
| P19 | Entropía (vei Anexu) |

### *4.2.3. Carauterístiques sofitaes en carauterístiques del testu (CSCT).*

| ***Escandayu*** | ***Descripción-cálculu*** |
|---|---|
| T0 | Númberu de llinies en blanco per númberu de llinies |
| T1 | Númberu mediu de pallabres nes oraciones |

### *4.2.4. Carauterístiques sofitaes en pallabres funcionales (CSPF).*

| ***Escandayu*** | ***Descripción-cálculu*** |
|---|---|
| Fi | Frecuencia cola que remanez la pallabra funcional i-ésima |

4.2.4.1. Llistáu de les pallabres funcionales escoyíes.

La frecuencia de les pallabres 'funcionales' (preposiciones, conxunciones, alverbios, determinantes, etc.) de la llingua qu'emplega un autor ye a carauterizalu de dalgún xeitu. El nuesu sistema sofítase nun soconxuntu de les pallabres 'funcionales' de la llingua asturiana. Les pallabres escoyíes por nós foron les siguientes:

> *abaxo, acá, acó, acullá, aculló, alantre, allá, alló, allalantre, allalantrón, allarribón, alredor, arriba, atrás, auquiera, ayuri, ayures, ayundes, cerca, cerque, dayuri, dayures, debaxo, delantre, detrás, dientro, ehí, ellí, embaxo, ende, enfrente, enriba, equí, fuera, lloñe, llonxe, metanes, nenyuri, nenyures, niundes, onde, ondequiera, u, perdayuri, perdayures, uquiera, abenayá, agora, aína, anguaño, antaño, antes, enantes, ayeri, ceo, cuando, cuantagüei, cuantayá, dacuando, depués, dempués, enagora, endagora, enaína, entá, entóncenes, entós, enxamás, güei, llueu, mañana, mentanto, nunca, sero, siempre, tarde, tovía, yá, adulces, albentestate, amodo, ansí, asina, apriesa, depriesa, arré, adré, arrémente, bien, como, damedies, darréu, despacio, davezu, igual, mal, mesmamente, meyor, peor, rápido, sele, selemente, talamente, abondo, afarto, apenes, bastante, bramente, cuanto, cuasi, cuasimente, dafechu, dafechamente, dalgo, daqué, demasiao, enforma, ensembre, malpenes, más, medio, menos, mui, muncho, nada, namái, poco, tan, tanto, claro, daveres, xuru, seguro, seguramente, sí, tamién, non, nun, siquiera, siquieramente, tampoco, acaso, escurque, guapamente, mesmo, pémeque, quiciabes, quiciás, seique, amás, entós, poro, sicasí, au.*

## 5. Esbilla d'un sistema clasificador

Existen estremaes posibilidaes teóriques a la hora d'escoyer un clasificador: clasificadores sofitaos en regles, árboles de decisión, clasificadores bayesianos, redes neuronales, máquines de sofitu vectorial (Support Vector Machines o SVM), etc. (Corney 2003: 43-45; Fissette 2010: 12-17). Los dos primeros esixen la discretización de valores dientro un conxuntu de referencia; les redes neuronales, pela so parte, nun almiten un númberu de parámetros peraltu (Corney 2003: 45). Delles investigaciones paecen suxerir, arriendes, que les máquines de sofitu vectorial son la meyor opción nel ámbitu nel que nos movemos (Corney 2003: 47).

Una SVM «ye un modelu que representa los puntos d'amuesa n'espaciu, xebrando les clases per un espaciu lo más ampliu posible. Cuando les nueves amueses se ponen en correspondencia con dichu modelu, en función de la so proximidá puen ser clasificaes nuna o otra clas» (WIKI s.v. *máquinas de soporte de vectores*). La carauterística fundamental de les SVM ye la de 'separación óptima': «esti tipu d'algoritmos busquen l'hiperplanu que tenga la máxima distancia (marxe) colos puntos que tean más cerca d'elli mesmu. Poro, tamién dacuando se conocen les SVM como *clasificadores de marxe máximu*. D'esti xeitu, los puntos del vector que son escandayaos con una categoría tarán a un llau del hiperplanu y los casos que s'alcuentren na otra categoría tarán al otru llau» (WIKI s.v. *máquinas de soporte de vectores*).

Weka ye una plataforma de software de deprendizax automáticu y minería de datos desendolcáu pola Universidá de Waikato (WEKA). Una implementación de SVM en Weka ye'l clasificador SMO, que ye'l que vamos emplegar en xeneral nes xeres desendolcaes darréu, esceuto na primera, na que se dan resultaos pa otros clasificadores de Weka, col envís de que pueda comprobase la meyora llograda per aciu'l clasificador SMO. El clasificador SMO ye «un algoritmu que resuelve eficientemente'l problema d'optimización que surde mientres l'entrenamientu de máquines de sofitu vectorial» (WIKI s.v. *sequential minimal optimization*), y ye ampliamente emplegáu anguaño con esti envís.

## 6. Casu d'estudiu

Preséntense darréu les xeres que constitúin el nucleu d'esti artículu. Convién tener en cunta qu'estes xeres foron escoyíes un tanto al debalu y nun formen parte d'un trabayu estructuráu. O dicho d'otru mou, escoyimos estes xeres de mou que se pudiera ilustrar d'un xeitu más o menos completu estremaos aspeutos d'aplicación de la identificación d'autoría, ensin que constituyan una secuencia de xeres empobinaes a una única fin.

### *6.1. Xera 1*

L'envís de la xera 1 foi valorar la viabilidá de la identificación d'autor dientro'l corpus testual de poemes n'asturianu en llinia *Caveda y Nava* (CAVEDAYNAVA).


Proyectu
Caveda y Nava
www.cavedaynava.org

**Introducción**

El proyectu Caveda y Nava ye una iniciativa d'ARAZ que tien l'apoyu de la Conseyería d'Educación y Cultura del Principáu d'Asturies. El so oxetivu ye dixitalizar y facer que tean disponibles en formatu testu -na so fase inicial- los testos de tola lliteratura n'asturianu anterior a 1950.

Nel añu 2.000, na primer fase, entamó la introducción de testos d'autores incluyíos na recopilación "Poesías selectas en Dialecto Asturiano", del propiu Caveda y Nava, l'estudiosu y escritor asturianu que-y da títulu al proyectu y "Los Nuevos Bablistas", de García Rendueles.

Yá tienes na rede los primeros frutos d'esti trabayu y nos prósimos díes dirás viendo inclusiones nueves. Como puedes albidrar, esti proyectu básase na aplicación al casu de la lliteratura asturiana del conocíu mundialmente Proyectu Gutenberg.

**Figura 2.** Páxina d'aniciu del proyeutu *Caveda y Nava*.

#### *6.1.1. Xera 1.1*

Na primer xera tuviemos en cuenta los testos ensembre de la biblioteca virtual *Caveda y Nava*. Esta primer xera valdrános, arriendes, pa calibrar el nuesu clasificador.

En primer llugar, descompunxemos aleatoriamente'l conxuntu de testos de la llibrería *Caveda y Nava* nun grupu d'entrenamientu (123 testos) y nun grupu de test (10 testos). Emplegando tan solo n-grames, con n=2, K=15 (númberu máximu de n-grames más frecuentes per testu que s'aceuten n'algoritmu), y clasificador Naive Bayes (clasificador bayesián), el percentax total d'aciertos foi del 40%.

Sicasí, emplegando una validación cruciada, con agrupamientu z=10, el percentax d'aciertos amiyó fasta un percentax del 12'782%, esto ye, tan solo 17 testos correutamente clasificaos d'un total de 133. La validación cruciada ye más significativa que'l casu anterior (esbilla d'un corpus d'entrenamientu y ún de test al azar), y sedrá lo qu'emplegaremos con preferencia de magar nesti trabayu. Consiste n'agrupar los datos en z conxuntos de la mesma cardinalidá, y emplegar socesivamente los posibles conxuntos de z-1 soconxuntos como corpus d'entrenamientu y el conxuntu restante como corpus de test. El promediu de toes estes operaciones ye lo que nos ufre la validación cruciada.

Darréu, fixemos variar dellos de los parámetros del nuesu clasificador, acordies colo que se recueye na tabla que s'inxer darréu, col envís de meyorar el resultáu.

| *agrupamientu* | *téunica* | *clasificador* | *percentax d'aciertos* |
|---|---|---|---|
| z=10 | 2-grames K=15 | Naive Bayes | 12'7820 % (17/133) |
| z=10 | 2-grames K=15 + CSP | Naive Bayes | 13'5338 % (18/133) |
| z=10 | 2-grames K=15 | SMO | 15'0376 % (20/133) |
| z=10 | 2-grames K=15 + CSC | SMO | 15'0376 % (20/133) |
| z=10 | 2-grames K=15 + CSPF | SMO | 15'0376 % (20/133) |
| z=10 | 2-grames K=15 + CSP | SMO | 15'7895 % (21/133) |
| z=10 | 2-grames K=7 + CSP | SMO | 15'7895 % (21/133) |
| z=10 | 2-grames K=2 + CSP | SMO | 15'0376 % (20/133) |
| z=10 | 2-grames K=25 + CSP | SMO | 16'5414 % (22/133) |
| z=10 | 2-grames K=35 + CSP | SMO | 16'5414 % (22/133) |
| z=10 | 2-grames K=45 + CSP | SMO | 16'5414 % (22/133) |
| z=10 | 3-grames K=25 + CSP | SMO | **18'7970 % (25/133)** |

**Figura 3.** Validación cruciada per aciu d'estremaos parámetros.

El meyor resultáu (d'ente los que testexemos) algamémoslu per aciu de 3-grames, con K=25, activación de CSP y clasificador SMO. El percentax total d'aciertos nesti casu foi del *18'7970%*. Dos son les causes d'esti resultáu tan ruin cuanto esperable: l'emplegu d'un númberu probetayu de testos, que son, arriendes, de llonxitú percurtia; y l'abondanza relativa d'autores. En xeneral, como ye evidente, los clasificadores funcionen meyor cuanto de más información disponen (más testos, más llargos) y cuanto de menos niveles de clasificación se riquen (menos autores).

Nun s'algamaron meyores resultaos amestando carauterístiques de tipu CSC, CSCT nin CSPF, quiciabes tamién como consecuencia del calter percurtiu de los testos. Nel casu de les frecuencies de les pallabres funcionales pudo influyir tamién el calter prenormativu de los testos y la presencia nellos de dialeutalismos de la llingua asturiana. Darréu d'ello, prescindiemos de magar d'estes carauterístiques nel nuesu clasificador.

#### *6.1.2. Xera 1.2*

Nesti casu, darréu, vamos centranos solo nuna pareya d'autores. Vamos tener en cuenta, por exemplu, tan solo los testos disponibles d'Enrique García Rendueles y de Pin de Pría (dos autores de la primer metada'l sieglu XX).

En primer llugar, descomponemos aleatoriamente'l conxuntu de testos de los que disponemos (16 de García Rendueles y 14 de Pin de Pría) nun grupu d'entrenamientu (12 + 10 testos) y nun grupu de test (4 + 4 testos). Emplegando tan solo n-grames, con n=2, K=15, y clasificador Naive Bayes, el percentax total d'aciertos ye del 87'5 %.

Nestes mesmes condiciones, emplegando dafechamente los testos disponibles (30 testos) d'estos dos autores, y habilitando una validación cruciada con un clasificador SMO, con 3-grames, K=25 y CSP, el percentax d'aciertos ye del *76'6667%* (23 sobro 30).

Resulta evidente, si se compara cola xera anterior, y esto ye lo que pretendíemos ilustrar, que la xera de deteutar un autor con testos tan curtios ente un carapiellu numberosu d'autores, ye enforma más abegosa que la de deteutar el verdaderu autor ente un númberu pequeñu d'ellos. Nesti últimu casu, la predicción que faemos paez relevante y útil, y podemos esfotanos nella.

### *6.2. Xera 2*

L'oxetivu de la xera 2 foi estudiar la posibilidá d'espardimientu d'esti procedimientu d'identificación d'autoría a los mensaxes na rede social asturiana *Ximiélgame* (XIMIELGAME) que comparten col anterior corpus la dificultá inherente a los 'testos percurtios'.


**Figura 4.** Interfaz de *Ximiélgame*, au se ve l'accesu al *telégrafu*.

Los datos de *Ximiélgame* obtuviémoslos vía RSS per aciu l'ataque al serviciu http://ximielga.me/mod/thewire/everyone.php, que permite l'accesu a los cables del llamáu 'telégrafu'. A esti envís, implementemos un cliente HTTP nel llinguax de programación *Java* que permitiónos percorrer los allugamientos de los cables del serviciu 'telégrafu'. Parsexemos el resultáu devueltu pol serviciu, obteniendo un conxuntu de datos como'l que s'ufre darréu.

```
1
2  Titulu: dangerouspiper: Canibalismu: Vezu o aición de comer carne humano. Entiendo entós que como la xente rico ye inhumana podemos c
3  Enllaz: http://ximielga.me/pg/thewire/dangerouspiper
4  Creador: Dangerous Piper
5  Fecha: Sun, 04 Sep 2011 12:41:57 +0200
6  Descripcion: dangerouspiper: Canibalismu: Vezu o aición de comer carne humano. Entiendo entós que como la xente rico ye inhumana pod
7
8  Titulu: trayasgaya: preparando una entradina pa espublizar el llunes nel blogue... va de llesbianes, semeyes y Ciañu... #intriga
9  Enllaz: http://ximielga.me/pg/thewire/trayasgaya
10 Creador: trayasgaya
11 Fecha: Sat, 03 Sep 2011 23:13:39 +0200
12 Descripcion: trayasgaya: preparando una entradina pa espublizar el llunes nel blogue... va de llesbianes, semeyes y Ciañu... #intrig
13
14 Titulu: dangerouspiper: Toi en: Cimavilla, Xixón (Comarca de Carreño) - http://acurti.es/c0t #Ayures Xeollocalización n'asturianu ht
15 Enllaz: http://ximielga.me/pg/thewire/dangerouspiper
16 Creador: Dangerous Piper
17 Fecha: Sat, 03 Sep 2011 03:02:32 +0200
18 Descripcion: dangerouspiper: Toi en: Cimavilla, Xixón (Comarca de Carreño) - http://acurti.es/c0t #Ayures Xeollocalización n'asturias
```

**Figura 5.** Fragmentu de los datos forníos pol sirvidor de *Ximiélgame* al nuesu cliente HTTP.

La xera centrémosla nel intentu d'identificar un autor ente una pareya de candidatos. En particular, esbillemos les entraes correspondientes a los autores (escoyíos al azar) *dangerouspiper* (539 cables) y *trayasgaya* (99 cables). En fayendo de xeitu automáticu la escoyeta de los cables, procediemos a aplicar sobro esti carapiellu de cables l'algoritmu d'identificación d'autoría.

Nesti casu, descompunxemos aleatoriamente los cables en dos sogrupos: a) datos d'entrenamientu (*dangerouspiper*, 529 cables; *trayasgaya*, 89 cables): b) datos de test o prueba (*dangerouspiper*, 10 cables; *trayasgaya*, 10 cables). L'oxetivu yera, evidentemente, tratar d'identificar l'autor de los 20 cables de prueba. Emplegando tan solo n-grames de caráuteres, con n=2, K=15 y clasificador Naive Bayes, el nuesu algoritmu danos un percentax d'aciertos del 75% na xera propuesta.

Per otru llau, realizando una validación cruciada cola totalidá de los datos, y con agrupamientu de 10, y les mesmes condiciones del algoritmu que nel casu anterior, clasificáronse correutamente'l *63'9498%* de los cables (408 cables sobro 638), que ye un resultáu abondo bonu.

### *6.3. Xera 3*

L'envís de la xera 3 foi aplicar estremaes téuniques d'autoría col envís d'algamar dalguna conclusión d'interés en dellos casos polémicos dientro la hestoria de la lliteratura asturiana. En particular, centrémonos na autoría del poema *Píramo y Tisbe* del sieglu XVII (ente Antón de Marirreguera y Benito de la Uxa) y na del poema *Al niñín Xesús* del sieglu XIX (ente Fernández de Castro y Xuan María Acebal).

#### *6.3.1. Xera 3.1:* Píramo y Tisbe

Tradicionalmente, esti testu atribúise a Marirreguera, poeta asturianu del sieglu XVII. Sicasí, Xulio Viejo, na so edición crítica de la obra de Marirreguera axudica esta obra a Benito de la Uxa, autor coetaneu de Marirreguera (Marirreguera 1997).

Nesti casu na biblioteca *Caveda y Nava* nun cuntamos con más testos de Benito de la Uxa que'l que se-y atribúi. Nin tampoco s'asoleyaron en nengún otru llugar que nós sepamos fasta güei[5]. Poro, desendolquemos los siguientes procedimientos alternativos:

En primer llugar, empleguemos la totalidá de los testos de la biblioteca virtual *Caveda y Nava*, escluyendo'l testu de *Píramo y Tisbe*, y asignando los testos que nun pertenecen a Marirreguera a un solu autor X. Con estos testos fixemos l'entrenamientu de la nuesa máquina. El nuesu envís ye agora averiguar si'l testu de *Píramo y Tisbe* (qu'emplegaremos como testu de test) pertenez o non a Marirreguera. Nestes condiciones, emplegando un clasificador SMO, con 3-grames, K=25 y CSP, la máquina danos que'l testu de *Píramo y Tisbe* pertenez a Marirreguera con un *95'037%* de probabilidá.

En segundu llugar, empleguemos nuevamente la totalidá de los testos, inda que calteniendo los nomes de tolos autores, esto ye, ensin inxertalos nun mesmu escandayu o alcuñu X. Col mesmu clasificador y el mesmu algoritmu, el resultáu ye nesti casu que d'ente tolos posibles autores (ente los que, como sabemos, nun ta Benito de la Uxa, desgraciadamente) identifícase nuevamente a Marirreguera como autor del testu con un *95'037%* de probabilidá.

[5] Con posterioridá a la redaición d'esti trabayu asoleyóse'l llibru *Poesíes*, de Benito de la Uxa y Antón Balvidares, editáu por Trabe (2012) n'edición de Xuan Carlos Busto, qu'inxer el poema *Sueños de Nabucodonosor* de Benito de la Uxa, inéditu (sacantes dos versos) fasta agora.

Por supuesto, nesti casu, los resultaos, dada la falta de testos disponibles de Benito de la Uxa, son inda más duldosos que nel restu de les xeres nes que trabayemos.

#### *6.3.2. Xera 3.2:* Al niñín Xesús

Güei paez claro qu'esti poema pertenez a Xuan María Acebal (s. XIX). Sicasí, hubo nello dalguna polémica años atrás. De fechu, Xurde Blanco propunxera que l'autor d'esti poema yera Manuel Fernández de Castro (s. XIX) (Blanco 1995; Fernández de Castro 1997: 74 -nota 3-). Sicasí, un poco más sero, Antón García perafitó que'l poema pertenez realmente a Acebal (Fernández de Castro 1997: 74 -nota 6-).

Como corpus d'entrenamientu emplegáronse tolos testos disponibles d'Acebal na biblioteca *Caveda y Nava* (5 poemes), más los tres poemes de Fernández de Castro inxertos na obra de Fernández de Castro (1997), esto ye, *Añada de la Virxe*, *Les oveyines* y *Un ricu avarientu*, desque la biblioteca virtual *Caveda y Nava* nun inxer poemes d'esti autor. Nestes condiciones, emplegando nuevamente un clasificador SMO, con 3-grames, K=25 y CSP, la máquina danos que'l testu de *Al niñín Xesús* pertenez a Xuan María Acebal (frente a Fernández de Castro) con un *99'552%* de probabilidá.

### *6.4. Xera 4*

#### *6.4.1. Deteición de plaxu intrínsecu*

L'oxetivu que nos plantegamos agora ye'l de descomponer un poema en fragmentos, calculando darréu la distancia (en dalgún sen) ente cada fragmentu y el testu completu. Cuando la distancia ye abondo importante (cuando ye superior a una llende afitada de mano), pescánciase que'l fragmentu correspondiente pudiera ser un plaxu.

Ye evidente que podría emplegase tamién una distancia sofitada en n-grames con esti oxetivu. De fechu, el procedimientu ye idénticu al de les xeres feches fasta agora, solo que referíu a fragmentos del mesmu testu en cuenta testos diferentes. Pero nesti casu vamos emplegar una forma estremada de midir la distancia, vamos emplegar la distancia de Lempel Ziv, que ye una implementación de distancia ente dos cadenes a partir de la llamada complexidá de Kolmogorov (ente otros, Basile 2008; Li 1997; Seaward 2009; Belabbes 2008; Malyutof 2007), desque dionos bonos resultaos yá n'otres pruebes no que fai al plaxu intrínsecu (Suárez García 2010). Nun vamos entrar en detalles matemáticos: abástanos conocer que, daes dos cadenes de testu *x* ya *y*, y un algoritmu de compresión de tipu *zip*, que comprime *x* a C(x), la distancia de Lempel Ziv defínese asina cuando l'algoritmu de compresión ye'l de Lempel Ziv:

$$d(x,y) \approx 1 - \frac{llonx(C(x)) + llonx(C(y)) - llonx(C(xy))}{máx\{llonx(C(x)), llonx(C(y))\}}$$

Nesti casu, darréu, nun emplegamos una máquina d'entrenamientu-clasificación, como fasta agora, sinón un simple algoritmu que fragmenta'l testu y calcula sobro los fragmentos la distancia conseñada ente ellos mesmos y el testu completu.

### *6.4.2. Aplicación*

L'oxetivu d'esta xera ye algamar dalguna conclusión d'interés na deteición de posibles fragmentos non orixinales del autor en poemes de Marirreguera (sieglu XVII), pente medies de téuniques de plaxu intrínsecu. Vamos centranos a mou d'exemplu tan solo nel poema *Hero y Lleandro*.

Nesti casu, l'algoritmu configurémoslu pa trabayar con descomposición en párrafos, con distancia de Lempel Ziv y llende fixada por un valor que calculemos como la media más una ponderación (fixada a 3.0) pola *meda* (mediana de les esviaciones absolutes con rellación a la mediana de les amueses). La descomposición fíxose en 43 fragmentos. Aquellos fragmentos que nes condiciones afitaes n'algoritmu son candidatos a fragmentos que nun pertenecerían al propiu autor de la composición son los que s'ufren darréu en forma de tabla. L'allugamientu indica'l númberu de caráuter au entama'l fragmentu (acordies colos testos de Caveda y Nava, desque desaniciada la cabecera informativa), la llonxitú ta espresada tamién en caráuteres, y la distancia refierse al valor calculáu de la distancia de Lempel Ziv.

| ***Fragmentu*** I | ***Allugamientu=4697*** | ***Llonxitú=311*** | ***Distancia=168.0*** |
|---|---|---|---|
| | Echóse andar muy fora de sentido ,<br>Como aquel que d'un palo está ablucado:<br>Perdió les riendes , y el rocín erguido<br>Revoltióse , y apúnxolo d' un llado .<br>Tornó á miralla , y viéndolu embebido ,<br>Dixo el criau : - Señor , vas descuidado<br>Asotripóse y dio una sofronada<br>Y esperó al toru , puestu é na estacada . | | |
| ***Fragmentu*** II | ***Allugamientu=10119*** | ***Llonxitú=333*** | ***Distancia=178.0*** |
| | Arrespondiói : - «Non tengo cenar cosa :<br>Vente aqui cabo min ; lo demás calla»,<br>Hero tapó la cara vergonzosa ,<br>Toda temblando . Al ir desabrochalla<br>Dexó cayer los brazos viciayosa ,<br>Y dixo : - «¿Aquella lluz ? Por Dios ; matalla ,<br>¿ Que ye de min . . . . . Lleandro , ¡ que me muero !<br>Isti ye de mió vida el fin postrero». | | |

| ***Fragmentu*** III | ***Allugamientu=10455*** | ***Llonxitú=314*** | ***Distancia=160.0*** |
|---|---|---|---|
| | Y non fue tal , que nunca más contenta ,<br>Con so amigu dormió á la pata llana ,<br>Y al alba dispertando soñolienta ,<br>Dixo : - «Cuirpo de tal ! ¿non sia mañana . . . . . ?<br>Bona la fixe entós». Va dormilienta<br>Y abriendo un poqueñín una ventana ,<br>Era tan claro q'iba ya la xente<br>Cad' un al so llabor muy dilixente | | |
| ***Fragmentu*** IV | ***Allugamientu=10772*** | ***Llonxitú=318*** | ***Distancia=159.0*** |
| | Ella dixo á Lleandro : - «Mira , amigo ,<br>Puedes estáte aquí sin dalgún vete .<br>Mió pa non vien acá falar conmigo :<br>Miós dames allá están en mió retrete .<br>Voi veles : lluigo torno á estar contigo ;<br>Fasta la nuiche , que querrás volvéte ;<br>Y si te quies quedar puedes quedáte :<br>Faré lo que pudiés por contentáte» . | | |

**Figura 6.** Tabla de fragmentos candidatos a nun pertenecer al propiu autor.

Convién dexar nidio que nos resultaos d'estos cálculos de plaxu intrínsecu nun podemos esfotanos dafechu, darréu qu'un autor siempre podría escribir un fragmentu que fora claramente diferente de los otros de so ensin que se tratara de plaxu dalu. Los resultaos tienen que s'interpretar nel sen de que son fragmentos embaxo sospecha o, como enantes se dixo, 'candidatos' a posibles fragmentos plaxaos.

### *6.5. Xera 5*

El plaxu intrínsecu tien munches utilidaes rellacionaes coles téuniques de clasificación: por exemplu, ye útil na clasificación d'idiomes. Nesta xera, baxemos de la rede la *Declaración de Drechos Humanos* en dellos idiomes (DDHH). Emplegando otra vegada la llamada distancia de Lempel Ziv ente los testos de les estremaes llingües, obtuviemos la tabla de distancies que s'inxer darréu. Estes distancies son una midida de la mayor o menor proximidá ente les estremaes llingües.

| | es | en | fr | it | ast | cat |
|---|---|---|---|---|---|---|
| es | 0.0 | 0.5182913274921215 | 0.44665539231199225 | 0.44522719759234675 | 0.23153838498893325 | 0.30204129436562843 |
| en | 0.5182913274921215 | 0.0 | 0.44400064079225926 | 0.5374985720742422 | 0.5967325705501936 | 0.43634777034270933 |
| fr | 0.44665539231199225 | 0.44400064079225926 | 0.0 | 0.5113628182004153 | 0.47861883466330385 | 0.35105494732631626 |
| it | 0.44522719759234675 | 0.5374985720742422 | 0.5113628182004153 | 0.0 | 0.5323133155208466 | 0.4308664589145338 |
| ast | 0.23153838498893325 | 0.5967325705501936 | 0.47861883466330385 | 0.5323133155208466 | 0.0 | 0.34420935364335126 |
| cat | 0.30204129436562843 | 0.43634777034270933 | 0.35105494732631626 | 0.4308664589145338 | 0.34420935364335126 | 0.0 |

**Figura 7.** Tabla de distancies ente diferentes idiomes.

Una clasificación de families llingüístiques obtúvola Benedetto emplegando un procedimientu asemeyáu al de nueso, como amosamos na figura que s'inxer darréu (Benedetto 2002).

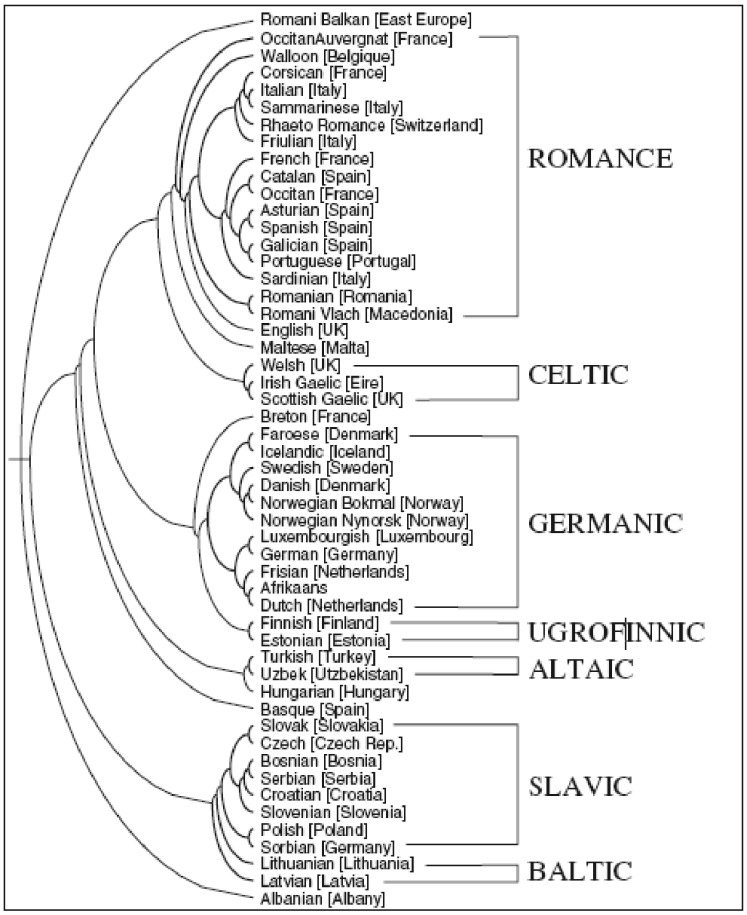


**Figura 8.** Clasificación de families d'idiomes obtenida por Benedetto (2002) a partir d'un procedimientu asemeyáu al emplegáu por nós.

### *6.6. Xera 6*

#### *6.6.1. Analizador sintáuticu Eslema*

L'oxetivu de la xera que vien darréu foi valorar la meyora que nes téuniques de deteición d'autoría ye a ufiertar l'emplegu del analizador sintáuticu de llingua asturiana *Eslema* (ESLEMA-ANALIZADOR) desendolcáu pola Universidá d'Uviéu nestos caberos años.

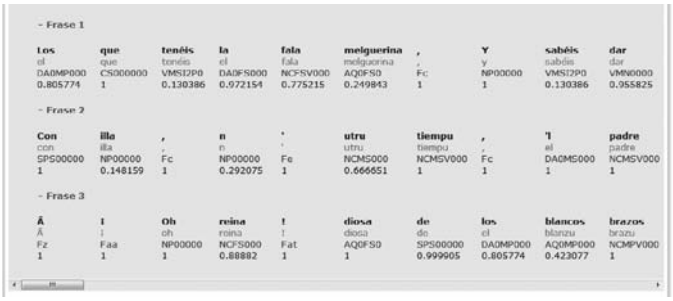


**Figura 9.** Análisis morfolóxicu desambiguáu en contestu forníu pol sirvidor de *Eslema*.

### *6.6.2. Aplicación*

Nesti casu, centrémonos nuevamente na pareya d'autores qu'enantes yá valoremos, la formada por Enrique García Rendueles y Pin de Pría.

Lo que fixemos nesti casu foi analizar caún de los testos per aciu del analizador sintáuticu *Eslema*. Desgraciadamente, tuviemos que faer esti procesu a mano, darréu que l'analizador *Eslema* actual nun ufre una API d'accesu (una Interfaz d'accesu) col envís de forninos de los sos datos de mou afayadizu (automáticu) y, darréu, nun nos foi posible acceder al mesmu per aciu d'un programa (d'un cliente) fechu *ad hoc* por nós que nos permitiera automatizar daqué'l procesu. Desque completáu'l procesu, cada ficheru quedó tresformáu nuna socesión d'escandayos morfosintáuticos acordies colo que se representa na figura qu'apruz darréu.

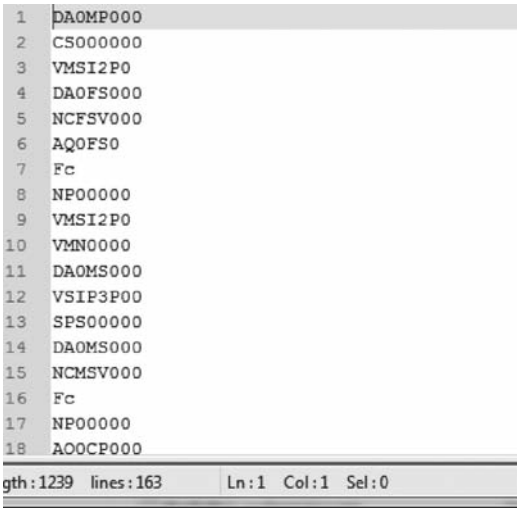


**Figura 10.** Detalle d'ún de los poemes de la coleición tresformáu por nós nuna socesión d'escandayos morfosintáuticos a partir del análisis forníu por *Eslema*.

Nestes condiciones refuguemos dellos testos por mor del análisis claramente incorreutu del nuesu analizador, motiváu principalmente pol calter prenormativu de los testos lliterarios emplegaos, y la so relativa abondanza de dialeutalismos que, dacuando, faen tracamundiar l'escandayáu a *Eslema*. En cualquier casu, y cuntando con estes llendes que de mano yá conocíemos, procediemos a la realización de la xera.

La coleición quedó reducida a 23 testos: 13 de García Rendueles y 10 de Pin de Pría. Darréu empleguemos los testos escandayaos del mesmu xeitu que si foran los testos orixinales. Emplegando un clasificador SMO, con 3-grames, K=25 y CSP, el percentax d'aciertos na validación cruciada ye del *91'3043%* (21 de 23), que ye un resultáu perbonu, abondo meyoráu con rellación a la xera orixinal. Albídrase, arriendes, que l'usu de testos más axustaos a la normativa estándar actual de la llingua asturiana habríen necesariamente de meyorar inda más los resultaos.

Por supuesto, existen estremaes posibilidaes d'emplegu d'esti escandayáu morfosintáuticu (por exemplu, un emplegu de n-grames de pallabra en cuenta de n-grames de caráuter; pa otres posibilidaes vei Van Halteren 1996), que podríen reflexase n'otres estremaes xeres o esperimentos, inda que nós nun fomos más alló, desque cuidamos qu'esta xera amuesa perdafechu la potencialidá del procedimientu.

## 7. Conclusiones

Presentóse un casu d'estudiu con estremaes xeres que pretendíen ilustrar l'aplicación de diferentes téuniques emplegaes na actualidá con rellación a la identificación d'autoría. Estes téuniques aplicáronse a un carapiellu de situaciones diferentes nel contestu de la nuesa llingua, pretendiendo resolver situaciones concretes.

Esperamos que nel futuru tanto téuniques d'esti tipu como otres en rellación cola llingüística computacional seyan a abrise camín y apaeza una investigación real y un desendolcu d'esti campu pa la nuesa llingua, qu'hasta hai poco tuvo ayena dafechu a esta mena de trabayos. En particular, sedría bien prestosa l'apaición de dalgún trabayu en rellación cola nuesa llingua, fasta agora inexistente, nel boletín añal de la *Sociedad Española para el Procesamiento del Lenguaje Natural* (SEPLN), o otres revistes asemeyaes.

### Bibliografía

ACEBAL, Xuan María (1995): *Obra Completa*. Uviéu, Alvízoras.
ALLA = < http://www.academiadelallingua.com/>.

ANTÓN DE MARIRREGUERA (1997): *Fábules, teatru y romances*. Uviéu, Alvízoras.

BASILE, C. et al. (2008): «An example of mathematical authorship attribution», en *Journal of Mathematical Physics* 49:125211-125230.

BELABBES, Sigem *et al*. (2008): «On Using SVM and Kolmogorov Complexity for Spam Filtering», en *Proceedings of the Twenty-First International FLAIRS Conference*.

BENEDETTO, Darío *et al*. (2002): «Language Trees and Zipping», en *Phisical Review Letters*. Vol. 88, númb. 4: 0487021-0487024.

BLANCO, Xurde (1995): *Poemes*. Uviéu, Academia de la Llingua Asturiana.

CAVEDAYNAVA = <http://www.cavedaynava.org/>.

CORNEY, Malcolm (2003): «Analysing E-mail Text Authorship for Forensic Purposes». Tesis. Disponible en: < http://eprints.qut.edu.au/16069/1/Malcolm_Corney_Thesis.pdf >.

CORNEY, Malcolm *et al.* (2002): «Gender-Preferential Text Mining of E-mail Discourse», en ACSAC '02 Proceedings of the 18th Annual Computer Security Applications Conference IEEE Computer Society Washington, DC, USA ©2002.

CORNEY, Malcolm *et al.* (2007): «Identifying the Authors of Suspect E-mail». Disponible en: < http://eprints.qut.edu.au/8021/ >.

DDHH = <http://www.ohchr.org/EN/UDHR/Pages/SearchByLang.aspx>.

DE VEL, Olivier & Malcolm CORNEY *et al.* (2002): «Language and gender author cohort analysis of e-mail for computer forensics», en *Proceedings Digital Forensics Research Workshop*. Syracuse (NY)-USA.

ESLEMA = <http://di098.edv.uniovi.es/apertium/comun/traductor.php>.

ESLEMA-ANALIZADOR = < http://di098.edv.uniovi.es/freeling/comun/analizador.php>.

ETLEN = <http://www.unioviedo.es/etlen/introduccion.html>.

FERNÁNDEZ DE CASTRO, Manuel (1997): *Versión asturiana del Dogma de la Inmaculada y poesía*. Uviéu, Alvízoras.

FISSETTE, Marcia (2010): «Author identification in short texts». Disponible en: <http://www.nici.ru.nl/~idak/teaching/batheses/MarciaFissette_scriptie.pdf>.

GIRÓN, F. J. & J. GINEBRA &. A. RIBA (2005): «Literatura y estadística: el problema de la autoría de Tirant lo Blanc», en BEIO. Boletín de Estadística e Investigación Operativa, vol. 21, 2: 6-9.

GIRÓN, Francisco Javier (2009): «Literatura y Estadística: problemas de autoría». Ciclo de talleres divulgativos «Matemáticas en Acción 2008». Disponible en: <http://www.matesco.unican.es/talleres_matematicas/transparencias20082009/transparencias-Giron.pdf >.

GONZÁLEZ, J. C. & M. CASTELLÓN & M. J. CASTEJÓN (2009): «Técnicas de clasificación en el entorno de Weka para la determinación de cultivos de regadío (cítricos) en Librilla, Murcia (España)», en Salomón Montesinos Aranda & Lara Fernández Fornos (eds.), *Teledetección, agua y desarrollo sostenible. XIII Congreso de la Asociación Española de Teledetección. Calatayud, 23 al 26 de septiembre de 2009*: 20.

GUILLEM-NIETO, Victoria & Chelo VARGAS-SIERRA *et al.* (2008): «Exploring State-of-the-Art Software for Forensic Autorship Identification». Murcia, Servicio de Publicaciones de la Universidad de Murcia. IJES, vol. 8 (1): 1-28.

HLLA = *Historia de la lliteratura asturiana*. M. Ramos Corrada (ed.). Uviéu, Academia de la Llingua Asturiana, 2002.

HOLMES, D. I. (1992): «A Stylometric Analysis of Mormon Scripture and Related Texts», en *Journal of the Royal Statistical Society*. Series A (Statistics in Society), vol. 155, 1: 91-120.

HOORN, Johan F. *et al.* (1999): «Neural Network Identification of Poets Using Letter Sequences», en *Literary and Linguistic Computing*, 14, 3: 311-338.

KIMLER, Marco (2002): «Using Style Markers for Detecting Plagiarism in Natural Language Documents». Tesis. Disponible en: < http://stynalyser.sourceforge.net/thesis.pdf>.

LI, M. & P. VITANYI (1997): *An introduction to Kolmogorov complexity and its applications*. Springer, New York. [2ª ed.].

MALYUTOV, M. B. *et al.* (2007): «Conditional complexity of compression for authorship attribution». SFB 649 Discussion Paper 2007-057. Disponible en: <http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2007-057.pdf>.

MCCOMBE, Niamh (2002): «Methods Of Author Identification». Disponible en: <http://www.scss.tcd.ie/undergraduate/bacsll/bacsll_web/mccombe0102.pdf>.

SEAWARD, Leane & Stan MATWIN (2009): «Intrinsic Plagiarism Detection using Complexity Analysis», en Stein, Rosso, Stamatatos, Koppel, Aguirre (eds.). PAN'09: 56-61.

SEPLN = < http://www.sepln.org/>.

SOFTASTUR = <http://softastur.org/index.php>.

SUÁREZ GARCÍA, Pablo *et al.* (2010): «A plagiarism detector for intrinsic plagiarism». Lab Report for PAN at CLEF 2010. Disponible en: <http://www.webis.de/research/events/pan-10>.

TAŞ, Tufan & Abul Cadir GÖRÜR (2007): «Author Identification for Turkish Texts». Natural Language Processing in *Computer Engineering* (2007).

TERMAST = <http://www.academiadelallingua.com/termast/index.php>.

VAN HALTEREN, Hans & Fiona TWEEDIE & Harald BAAYEN (1996): «Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution». *Lit Linguist Computing (1996)* 11 (3): 121-132.

WEKA = <http://www.cs.waikato.ac.nz/ml/weka/>.

WIKI = < http://es.wikipedia.org/wiki/>.

XIMIELGAME = <http://ximielga.me/>.

## ANEXU

$N$ = númberu total de pallabres nel testu
$V$ = númberu total de tipos (de pallabres estremaes nel testu)
$V_i$ = númberu total de pallabres del tipu i-ésimu.

$$\text{RGuirad} = \frac{V}{\sqrt{N}}$$

$$\text{C}Herdan = \frac{\log_{10} V}{\log_{10} N}$$

$$\text{V}Herdan = \overset{V}{\underset{i=1}{!}} V_i \frac{i^2}{N^2}$$

$$KRubet = \frac{\log_{10} V}{\log_{10}(\log_{10} N)}$$

$$A\text{Maass} = \sqrt{\frac{\log_{10} N \,!\, \log_{10} V}{(\log_{10} N)^2}}$$

$$U\text{Dugast} = \frac{(\log_{10} N)^2}{\log_{10} N - \log_{10} V}$$

$$\text{Janenkov y Neistoj} = 1 - \frac{V^2}{V^2 \log_{10} N}$$

$$W\text{Brunet} = N^{V^{-0.172}}$$

$$\text{SSichel} = \frac{núm b.Hapaxdislegomena}{V}$$

$$\text{HHonore} = \frac{100 \log_{10} N}{1 - \frac{núm b.Hapaxlegomena}{V}}$$

$$K\text{Yule} = 10^4 \left[ -\frac{1}{N} + \sum_{i=1}^{V} V_i \left( \frac{i}{N} \right)^2 \right]$$

$$D\text{Simpson} = \sum_{i=1}^{V} V_i \frac{i}{N} \frac{i-1}{N-1}$$

$$Entropía = \sum_{i=1}^{V} V_i \left( -\log_{10} \frac{i}{N} \right) \frac{i}{N}$$