C A D E R N O S D E L I N G U A

ANO 2008-2009

30-31

R E A L
ACADEMIA
G A L E G A

Colaboracións e correspondencia:
Xosé Luís Regueira Fernández
Paseo da Quinta, 23
15897 Santiago de Compostela

Subscrición e intercambio: Dirixirse a
Editorial Galaxia, S. A.
r/. Reconquista, 1 - 36201 Vigo

Edita: Real Academia Galega
r/. Tabernas, 11 - 15001 A Coruña

I.S.S.N.: 1130-5924

Título clave: Cadernos de Lingua
Título abreviado: Cad. Ling.

Depósito Legal: C-1.642-1992

Deseño: SINMÁS COMUNICACIÓN VISUAL S.L.
Imprime: Galigraf Galicia

ÍNDICE

ARTIGOS

RECENSIÓNS

AVALIACIÓN DUN ETIQUETADOR AUTOMÁTICO ESTATÍSTICO PARA O GALEGO ACTUAL: XIADA

EVA DOMÍNGUEZ NOYA
Centro Ramón Piñeiro para a Investigación en Humanidades
edomin@cirp.es

FCO. MARIO BARCALA RODRÍGUEZ
Centro Ramón Piñeiro para a Investigación en Humanidades e Universidade de Vigo
fbarcala@cirp.es

MIGUEL ÁNGEL MOLINERO
Universidade da Coruña
mmolinero@udc.es

RESUMO
Neste traballo avaliamos, dende o punto de vista lingüístico, un etiquetador automático estatístico, desenvolto conxuntamente polo Centro Ramón Piñeiro para a Investigación en Humanidades e o Grupo COLE das Universidades de Vigo e A Coruña, destinado a etiquetar os documentos do *Corpus de Referencia do Galego Actual* co obxecto de proporcionar recursos e ferramentas para a análise lingüística computacional do galego actual.

PALABRAS CHAVE
Corpus, galego, etiquetación, lematización, avaliación.

ABSTRACT
"Evaluation of an statistical automatic labelling machine for present Galician: XIADA"

We evaluate, from a linguistic point of view, a statistical automatic labelling machine, which is explained together by the

Center Ramón Piñeiro on Humanities Research and the COLE Group of Vigo and La Coruña Universities, and which also set aside for labelling the papers of Present Galician Reference Corpus so as to provide tools and resources for the computational linguistic analysis of Present Galician.

KEYWORDS
Corpus, Galician, tagging, lemmatization, evaluation

1. Introdución

1.1. O corpus

O *Corpus de Referencia do Galego Actual* (CORGA)[1] é un corpus documental que en decembro de 2008 contén 23 millóns de formas, integrado por distintos tipos de textos, -xornais, semanarios, revistas, ensaios e textos de ficción (novela, relato curto e teatro)-, que abrangue temporalmente dende o ano 1975 ata a actualidade.

Todos os documentos do corpus están codificados segundo o estándar XML[2] (*eXtensible Markup Language*) e cada un deles está constituído por unha *cabeceira* e un *corpo*. Na *cabeceira* incluímos a información bibliográfica pertinente xunto cunha caracterización temática, e no *corpo* aparece o texto estruturado formalmente segundo o tipo de documento de que se trate. Por exemplo, o XML permítenos considerar un xornal un único arquivo XML que está organizado en múltiples noticias as cales, á súa vez, conteñen obrigatoriamente un corpo e opcionalmente un titular, resumo e/ou pé de foto. A maiores, cada un destes elementos está constituído por parágrafos (texto comprendido entre dous puntos e á parte) e estes son segmentados en oracións (secuencia textual separada do resto do texto por un punto, punto e coma, dous puntos, etc.). Esta disposición deta-

[1] http://corpus.cirp.es/corga
[2] http:/www.w3.org

llada posibilita realizar consultas sobre a totalidade da noticia ou sobre unha unidade estrutural concreta (titular, resumo, pé de foto ou corpo). Ademais, ao ter asignada como mínimo unha área temática que caracterice cada noticia, é posible delimitar os ámbitos nos que unha palabra ou expresión se emprega.

O formato XML facilitou diferenciar formalmente os distintos tipos de textos que integran o corpus pero tamén permitiu a marcación de fragmentos que aparecen nunha lingua distinta ao galego -deste xeito evitamos indexalos no sistema de buscas e impedimos que engorden os datos totais do corpus no referente a número de palabras-; marcar a presenza de poemas e de cada un dos seus versos[3]; marcar a existencia de entrevistas e as intervencións dos distintos interlocutores[4] ou sinalar a aparición dun determinado vocábulo nunha gráfica ou nunha táboa, etc.

Dende o 2007 o sistema de buscas do CORGA que está dispoñible na rede incorpora unha nómina de autores e obras que lle permite ao usuario:

1) Obter datos específicos sobre a estrutura textual do corpus. Para iso pódese empregar un ou varios criterios de busca, sendo posible combinar estes ou os seus valores segundo as necesidades do usuario: autor, título da obra, área temática, ano, etc.

2) Obter a frecuencia normalizada dunha ocorrencia fronte á global; deste xeito poden obterse datos relativos á frecuencia dunha expresión de busca con respecto ao subconxunto de documentos sobre os que se realiza a consulta.

---

[3] Decidimos non incluír no deseño do CORGA poemarios, porén non é infrecuente a aparición dalgún poema, sobre todo en ensaios literarios e noticias de opinión e cultura. A caracterización de certas oracións ou fragmentos como pertencentes a un poema pódelle servir ao usuario para desbotar ou explicar certas ocorrencias.

[4] Non só nas entrevistas e coloquios senón tamén nas obras teatrais remitimos cada alocución ao personaxe ou interlocutor que lle corresponde. Deste xeito, os nomes non suman voces para o cómputo de formas totais e proporciónamoslle ao usuario información que lle pode ser útil na súa consulta (por exemplo, determinadas ocorrencias poden aparecer só nun interlocutor).

Conscientes de que para realizar consultas máis avanzadas e dar un salto cualitativo nas posibilidades de busca é fundamental que os textos do CORGA estean lematizados e etiquetados, paralelamente á construción do corpus e á incorporación a este de novos documentos, no Centro Ramón Piñeiro para a Investigación en Humanidades estase desenvolvendo unha ferramenta denominada *Etiquetador e lematizador do galego actual*[5] cuxos resultados iniciais queremos presentar aquí. Con ela pretendemos asignarlle a cada unidade léxica o lema e etiqueta que lle corresponde segundo o contexto no que se localice.

Dende o punto de vista lingüístico, para a construción do analizador e etiquetador automático foi necesario seguir os pasos seguintes: determinar un sistema de etiquetas, deseñar a estrutura dun lexicón e construílo en consonancia con esta e elaborar certas regras lingüísticas que axudasen no recoñecemento e segmentación das distintas unidades léxicas.

1.2. O conxunto de etiquetas

O etiquetario ou *tagset* co que traballamos foi desenvolvido no marco do proxecto *Construción, etiquetación e lematización do Corpus de Referencia do Galego Actual* como unha ferramenta para a análise computacional e a anotación automática dos textos do CORGA. Para a súa elaboración seguimos as recomendacións de EAGLES [Leech e Wilson, 1994] e os textos básicos da lingüística galega[6]. Así, este sistema de etiquetación presenta unha estrutura xerárquica na que se identifica, no primeiro nivel, a categoría morfolóxica, e nun segundo, os atributos gramaticais pertinentes que constitúen o conxunto básico de atributos para cada categoría.

En primeiro lugar delimitamos as clases de palabras existentes en galego e, a continuación, establecemos os atributos gramaticais pertinentes que permitisen o recoñecemento mor-

---

[5] Ferramenta á que denominamos XIADA. Véxase http://corpus.cirp.es/xiada

[6] Véxase o apartado de Bibliografía complementaria.

folóxico de calquera palabra. Primamos a descrición morfolóxica e reducimos a información sintáctica á caracterización de determinadas categorías cuxos elementos integrantes poden funcionar como determinantes, non determinantes, nucleares ou modificadores.

Como se pode observar no etiquetario[7] de XIADA que reproducimos na táboa 1, na primeira columna recóllense as diferentes clases de palabras; nas seguintes columnas recóllense os atributos e os valores destes e, por último, nas celas incorpórase un número que indica a posición que o atributo, se é pertinente, ocupa dentro da etiqueta final.

A distribución dos atributos dentro de cada categoría é simétrica, de tal xeito que todos os elementos pertencentes a unha categoría dada posúen o mesmo número de atributos e os seus valores ocupan a mesma posición dentro da cadea resultante. Con todo, non todos os elementos pertencentes a unha mesma categoría morfolóxica responden positivamente a todos os atributos que a caracterizan. Así, na categoría verbal, o atributo xénero só é un trazo pertinente para a caracterización do participio. Para o resto de elementos do paradigma verbal a aplicación dese atributo non é pertinente. Por este motivo creamos un valor "0" co significado de "Non aplica". Hai tamén ocasións nas que a atribución dun valor inequívoco é imposible polo que debemos neutralizar o valor dalgún ou de varios atributos. En XIADA marcamos esta ambigüidade co valor "a" que representará o xénero, número, persoa, posuidor ou valor dependendo da clase de palabra e da posición que ocupe na cadea de etiquetaxe.

Combinando as clases de palabras delimitadas e os atributos que as caracterizan, este sistema de etiquetación presenta arredor de 400 etiquetas posibles aínda que polo momento no subcorpus desambiguado só constatamos 375 distintas.

[7] Dispoñible en http://corpus.cirp.es/xiada

1.3. O lexicón

Unha vez tivemos claro cal ía ser o noso etiquetario necesitabamos un lexicón onde almacenar a información de etiqueta e lema que lle correspondía a cada palabra.

Dado que a introdución e xestión dos datos debía ser práctica, estimamos que maioritariamente as entradas do lexicón tiñan que ser caracterizadas morfoloxicamente desagregándoas en *lema*[8], *raíz*[9], *subetiqueta*[10] e *grupo de derivación*[11]. Esta estrutura esixiu un estudo formal detallado dos integrantes de cada unha das clases gramaticais variables para agrupalos en modelos e deste xeito construír a gramática formal que nos permitise analizar e tamén xerar as formas flexionadas e conxugadas do galego moderno.

O lexicón de XIADA, ademais de conter a información morfolóxica que permite a identificación e caracterización morfolóxica plena de calquera palabra galega, responde aos pará-

---

[8] É o representante das distintas formas flexionadas ou conxugadas que se integran nun paradigma. No noso sistema a clase morfolóxica é un elemento constitutivo do lema polo que todo lema está asociado a unha clase de palabra. Así, para "galego", por exemplo, teremos dous lemas: un, asociado á categoría substantivo e outro á de adxectivo.

[9] Basicamente podemos considerar que a raíz é o lema sen morfemas gramaticais. A raíz e o lema coinciden nas entradas das clases invariables e nas dos substantivos e adxectivos que só presentan flexión no número. En ningún caso se pode identificar esta raíz co lexema da gramática tradicional posto que nós, para o seu establecemento, non temos en conta os morfemas derivativos senón só os flexivos.

[10] Caracterización morfolóxica básica da entrada no lexicón.

[11] Conxuntos de terminacións para as que se proporcionan uns valores. Por exemplo o grupo G1 consta de:

o	-> masculino singular
a	-> feminino singular
os	-> masculino plural
as	-> feminino plural

Deste xeito, para todos os substantivos e adxectivos que teñen flexión de xénero e forman o plural engadindo -s introducimos só o lema e a raíz correspondente, sen necesidade de ter que inserir todas as formas do paradigma.

metros seguintes: indicación sobre a normativa oficial, ámbito léxico e procedencia da forma:

a) Normatividade. Para un correcto recoñecemento das formas existentes nos textos reais contemporáneos galegos necesitamos que o lexicón sexa amplo e que non só estea integrado por lemas normativos. Se este fose o caso, moitos dos documentos do CORGA non poderían ser etiquetados ou seríano a medias. É fundamental, polo tanto, darlles cabida tamén a dialectalismos, hiperenxebrismos, castelanismos, etc. porque estes aparecen nos textos. Debemos, así a todo, poder diferencialos e para iso introducimos este criterio na caracterización de lemas, formas e incluso algunha desinencia.

b) Tipo de léxico. Cremos que podía ser útil diferenciar na estrutura do lexicón módulos segundo o tipo de léxico que conteñan que poidan ser combinados. Para iso dispuxemos no deseño da estrutura, e obviamente implementamos na construción do lexicón, dun campo máis no que se caracteriza o lema como pertencente ao léxico común ou ao técnico-científico, especificando neste último o ámbito no que se clasifica: administración, economía, medicina, etc.

c) Fonte. Nos textos actuais preséntanse formas que non aparecen nos dicionarios-vocabularios preceptivos da lingua galega ben porque son termos técnicos para os que se acaba de propoñer unha denominación ben porque se documentan cun uso categorial distinto, non obstante, a súa introdución no lexicón é imprescindible para o recoñecemento e caracterización dos textos. Coa indicación da fonte achegamos datos concretos sobre a localización do lema e a fiabilidade da forma.

Recapitulando, o lexicón de XIADA, ademais de conter a información morfolóxica que permite a identificación e caracterización morfolóxica plena de calquera palabra galega actual, proporciona información sobre a normativa oficial, o ámbito léxico e a procedencia da forma[12].

---

[12] Así, por exemplo, cando introducimos unha unidade como "galego" no lexicón temos que proporcionar a seguinte información:

Unha vez deseñada a estrutura do lexicón, implementámolo coas entradas do *Vocabulario Ortográfico da Lingua Galega* (VOLG) e coas 100 000 formas máis frecuentes do CORGA, o que se traduce en algo máis de 55 000 lemas.

Na seguinte táboa amosamos a información correspondente ás tres grandes clases de palabras variables -substantivos, adxectivos e verbos- sobre o número de lemas que figura no lexicón xeral de XIADA e a cantidade de formas que é posible xerar con eses lemas. Debemos aclarar que as formas dos diminutivos, aumentativos ou superlativos sintéticos regulares son remitidos ao lema base, non conforman un lema á parte. Así, "gravísimo" ou "caladiños" están asociados respectivamente aos lemas "grave" e "calado". Polo que se refire ao cómputo das formas, nas verbais, non están incluídas as posibles combinacións con pronomes enclíticos.

CATEGORÍA	LEMAS	FORMAS
Substantivos	29 806	64 646
Adxectivos	13 715	45 449
Verbos	6 779	461 185

1.4. O preprocesador - etiquetador

Lingüisticamente, unha vez elaborados os modelos formais que acollen as raíces e grupos de derivación e implementado o lexicón computacional, estamos en disposición de etiquetar automaticamente calquera texto galego contemporáneo. Trátase de executar o programa que o equipo informático do

lexicón	sublexicón	raíz	subetiqueta	grupo	lema	categoría-lema	normativa-lema	fonte
xeral	principal	galeg	Sc	G1	galego	S	si	volga_2004
xeral	principal	galeg	A0	G1	galego	A	si	volga_2004

proxecto está desenvolvendo para traballar con arquivos codificados en XML, formato no que están os textos do CORGA.

Para proceder a etiquetar un documento, primeiro é necesario delimitar as distintas unidades léxicas que este contén, tarefa que lle corresponde ao preprocesador. Aínda que poida parecer sinxelo, non é tarefa trivial indicarlle a un programa como segmentar formas verbais con pronomes enclíticos ou separar os formantes dunha contracción. Seguimos traballando para que cometa menos erros pero é inevitable que nalgúns casos de ambigüidades falle de vez en cando. É sobre todo un traballo informático pero no que o equipo lingüístico participa coa construción de regras que gobernan o comportamento do preprocesador. Por exemplo, a caracterización como pronome átono acusativo de terceira persoa para "no, na, nos, nas" só será válida se vai enclítico a unha forma verbal rematada en ditongo.

No noso caso, ademais, despois de delimitar as unidades léxicas presentes nos textos asígnalle a cada unha os lemas e etiquetas posibles. Así, a saída do documento preprocesado proporcionaranos o texto analizado contendo para cada palabra gráfica ou *token* todas as posibles análises morfolóxicas. Non obstante, para realizar unha consulta nun corpus etiquetado non necesitamos todas as análises potenciais dunha forma senón só a que presenta en cada contexto, tarefa esta que realiza o etiquetador escollendo a etiqueta e o lema correctos.

Para poder empregar con certas garantías o etiquetador automático estatístico, necesitamos adestralo desambiguando manualmente un subconxunto de textos. O subcorpus escollido para o adestramento inicial foi extraído do *Corpus de Referencia do Galego Actual* e constou de 200 000 palabras procedentes de noticias xornalísticas do ámbito económico.

Sabemos que as probabilidades de acerto do etiquetador aumentan se o corpus de adestramento é amplo, pero tamén estamos seguros de que os resultados positivos se incrementarán en proporción á riqueza e variedade temática do subcorpus desambiguado manualmente, de xeito que o noso obxectivo máis próximo é desambiguar á man unhas 200 000 formas para

cada un dos grandes bloques do CORGA: prensa, ensaio e ficción.

A mediados do 2008 iniciamos unha segunda fase de desambiguación manual dun subcorpus, extraído tamén do CORGA, que estaba constituído polo número 1166 do ano 2005 do semanario *A Nosa Terra* e polas noticias da versión electrónica do día 17 de marzo de 2007 do xornal *Galicia Hoxe*. En total, 58 766 palabras.

Durante a revisión e corrección da versión etiquetada automaticamente do semanario apreciamos unha alta porcentaxe de acerto na etiquetación, motivo polo que valoramos a eficacia do etiquetador co seguinte documento escollido para desambiguar. As 89 noticias e as 22 212 palabras das que consta o arquivo GH2007-03-17.xml son suficientes para realizar unha proba significativa sobre a efectividade do etiquetador empregado.

Antes de proceder a expoñer os resultados obtidos, queremos chamar a atención sobre dous feitos que consideramos fundamentais para valorar na súa xusta medida o etiquetador utilizado:

1) As porcentaxes de acerto que presentan os etiquetadores existentes para unha lingua como o castelán ou o inglés -para o galego non hai estatísticas- son do 95%/97% [Graña, 2000: 137-165]. Con todo, debe terse en conta que nos resultados que presentan esoutros traballos inclúense soamente erros de etiquetación -non de lematización-, e, ademais, en ningún caso eses etiquetadores realizan segmentacións das unidades léxicas presentes no texto. Estes dous factores, non lematización e non segmentación textual, cómpre telos presentes na comparativa final.

2) O tamaño do etiquetario que empregamos -arredor de 400 etiquetas distintas- é bastante máis grande ca o empregado pola maior parte de etiquetadores. Segundo Jurafsky [Jurafsky, 2000: 296-297], os etiquetarios máis usados para o inglés son o

Penn Treebank[13] que consta de 45 etiquetas e é o que se usou no corpus *Brown*[14]; o *C5* de 61 etiquetas empregado polo programa CLAWS[15] para etiquetar o *British Nacional Corpus* (BNC)[16] e o *C7*, de 146 etiquetas, utilizado para etiquetar un subconxunto do BNC.

Só uns poucos sistemas de etiquetación teñen un etiquetario equiparable en tamaño ao que aquí presentamos [Graña, 2000: 139 e Sampson, 1994] (Véxase táboa 1).

2. Avaliación da etiquetación automática realizada sobre as 22 212 formas gráficas das noticias do xornal GH2007-03-17

Cómpre advertir que fomos rigorosos na análise efectuada e recollemos e clasificamos todos aqueles casos nos que a solución proporcionada polo etiquetador automático non coincidiu coa solución escollida polo lingüista revisor. Tamén é preciso aclarar que os erros cometidos polo etiquetador debido á presenza de erratas, faltas de acentuación ou outros erros lingüísticos, existentes todos eles nos documentos orixinais, foron recollidos e rexistrados no apartado que lles correspondía segundo o tipo de fallo detectado.

Todos os erros que se produciron no documento etiquetado automaticamente intégranse nalgunha das epígrafes seguintes:

a) Non se lle asigna lema á unidade léxica aínda que si se propón unha caracterización morfolóxica. Non se pode, con todo, responsabilizar ao etiquetador dos fallos que se producen por non asignar lema -de momento non ten tarefas específicas de lematización de formas descoñecidas, simplemente, para cada forma, ten asociado un lema a unha etiqueta-. Os erros ti-

---

[13] Véxase http://www.cis.upenn.edu/~treebank

[14] http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/ corpora/list/ private/brown/brown.html

[15] http://ucrel.lancs.ac.uk/claws/

[16] http://www.natcorp.ox.ac.uk

pográficos, presentes nos documentos analizados, que provocan a aparición de palabras inexistentes contabilizáronse neste apartado.

b) Prodúcese unha segmentación errónea á hora de separar os constituíntes da unidade.

c) Caracterízase erroneamente un elemento (*token)* cun atributo que non lle corresponde.

d) Érrase ao clasificar a unidade categorialmente.

e) Engánase ao atribuír o lema.

f) Non se proporciona a análise morfolóxica axeitada porque esta non figura no lexicón. O lematizador só lle asigna lema a aquelas formas que están implementadas nos lexicóns cos que traballa e para que o etiquetador proporcione unha análise determinada dunha forma gráfica calquera é preciso que na gramática formal que este emprega estea recollida dita análise.

g) Non se produce un recoñecemento morfolóxico atinado porque o texto orixinal presenta erros ortográficos ou lingüísticos, entendendo por estes últimos, basicamente, desviacións da norma de tipo morfosintáctico.

Un dos problemas que atopamos ao traballar con textos reais é a presenza, en bastantes ocasións, de formas ou construcións morfosintácticas que non seguen a normativa oficial[17], por exemplo, en canto ao xénero dos substantivos, ao uso do til diacrítico ou ao emprego dos pronomes átonos. Outro dos problemas derivados de non manipular o texto orixinal xórdenos coa presenza dos erros tipográficos que dan lugar a unha unidade -imposible nese contexto- coincidente con outra xa existente no lexicón. Pois ben, estes dous tipos de erro provocan que unha unidade, á que lle correspondería unha etiqueta *x*, coincida no lexicón formalmente con outra unidade cuxa etiqueta é *z*, polo que o etiquetador non vai acertar coa análise.

Na táboa 2 (véxase final do artigo) amosamos todos os erros detectados na etiquetación automática das 22 212 formas

[17] Nunhas ocasións percíbese que ese uso é intencional e noutras, que é froito do descoñecemento do galego normativo.

gráficas das que consta a versión electrónica do xornal *Galicia Hoxe* do día 17 de marzo de 2007 introducido en CORGA, distribuídos por sección e organizados segundo a anterior clasificación.

Só en tres seccións -Cultura-Sociedade, Opinión e Televisión- a porcentaxe de acerto descende do 94,5% ata situarse no seu punto máis baixo no 92,1%. Unha das razóns que explican esta diferenza é a abundancia de nomes propios multipalabra -sobre todo na sección de TV, títulos de películas, series, etc.- que non son correctamente segmentados e, en consecuencia, provocan unha concatenación de erros.

Outro dos factores diferenciais radica no estilo e lingua propios das noticias de cultura e opinión que se asemellan máis á narración que á crónica xornalística. É probable que cando dispoñamos do corpus de adestramento correspondente á ficción, se o empregamos para etiquetar este tipo de noticias, vexamos unha melloría considerable[18].

A continuación examinaremos con detalle algúns dos erros cometidos polo etiquetador segundo a clasificación antes proposta. Comezaremos polos que non son achacables ao programa.

2.1. Non se lle asigna lema á unidade léxica

Os *tokens* incluídos baixo esta epígrafe -122 en total que representan o 0,54% do total- presentan o lema baleiro aínda que son etiquetados correctamente na maioría dos casos. Esta circunstancia dáse porque, a pesar de que o lematizador-etiquetador non é capaz de asignar un lema se este non figura no lexicón ben por omisión, ben por ser errata, castelanismo, estranxeiris-

[18] O etiquetador que utilizamos foi adestrado cun corpus xornalístico de 200.000 palabras da temática económica formado, sobre todo, por noticias das seccións de Economía, Galicia, España e Internacional. A nosa idea é constituír un subcorpus desambiguado manualmente por cada xénero presente no CORGA de xeito que os poidamos combinar e escoller aquel que máis se adecúa ao tipo de texto que queremos etiquetar automaticamente.

mo, etc., si é capaz de aventurar unha caracterización morfolóxica tendo en conta o contexto no que ese *token* se insire.

O etiquetador non é quen de detectar nin de corrixir unha errata -tampouco é ese o seu cometido- polo que ante os erros tipográficos cometidos polo(s) autor(es) da(s) noticia(s) vaise comportar como se non o fosen e vai analizar o texto que se lle proporciona. Posto que un acento mal colocado ou unha letra que sobra ou falta converten a palabra pretendida noutra, existente ou non na nosa lingua, o lematizador-etiquetador vai actuar de modo distinto segundo o caso. Se a unidade resultante coincide con outra que teña no lexicón, outórgalle a etiqueta e o lema que a el lle consta; se non a atopa, non lle vai asignar ningún lema. Por este motivo algunhas erratas, como se poderá observar nos exemplos, foron catalogadas neste apartado.

Nas seguintes oracións o contexto determina que a unidade que destacamos subliñándoa fose analizada correctamente:

> <expresión>Tahoces acusou a Xunta de "crear inseguridade xurídica" e de "atentar contra as regras do xogo", xa que considera que o ofrecemento de presenza pública "actuou como criterio convalidante" de propostas que "non cumprían outros requisitos".</expresión> [A0ms] [GH2007-03-17/19; *Galicia*]

> <expresión>Non obstante, o país está envolto nunha "pseudo" guerra civil por mor da violencia interconfesional con atentados terroristas a diario e máis de 650.000 iraquís mortos en catro anos.</expresión> [A0fs] [GH2007-03-17/70; *Internacional*]

> <expresión>Algo vai mal cando unha sociedade se infantiliza ata estes extremos e chamamos para todo a papa Estado (ou papa Xuíz).</expresión> [Vpi30s]; [GH2007-03-17/86; *Opinión*]

Algunhas outras formas descoñecidas para o lematizador pero ben analizadas son as seguintes: *chanchullos; vertebrador; reapertura; promedia; enfatizou; galescolas; xugosos; PDC; desbaldimento; vozarrón; comisión-homenaxe; catátrofe; emprendemento; FEFN; INWA; pedofilia; pedófilo; chats; armonización; bici; ornanismos; linguística; eúscaro; viceconselleiro; plurilinguismo; anfitrón; NOAA; esiste; catoliscismo; RG; blog; histórico-contemporánea; performer, etc.*

No entanto, equivócase en formas como *garantista, CIG-Ensino, reponsables, FEFN, Réplica-rede, calatán, plurilingue, monolingue, biodioversidad, represaliados, socio-cultura, PS3, SETI@Home, XrossMediaBar, pasquines, etc.*

2.2. Erros ortográficos, lingüísticos ou desviacións morfolóxicas da norma existentes nos textos orixinais

A incidencia deste apartado é das menores -12 casos que representan o 0,05% dos erros totais-. Con todo, para ser rigorosos, debemos reflectila aínda que, máis tarde, cando sexa o momento de avaliar verdadeiramente o etiquetador, deberemos eliminar estas análises do seu debe pois son erros do autor textual, non da ferramenta.

Baixo esta epígrafe recollemos só os fallos ortográficos ou lingüísticos cometidos polos autores das noticias que provocan que, por coincidencia gráfica con algunha forma existente no lexicón, o etiquetador actúe de xeito erróneo. Non nos referimos ás grallas ortográficas que causan a aparición de novas unidades léxicas (*reponsables* é claramente un erro de tecleado por *responsables*) senón soamente a aqueles erros ortográficos ou desviacións morfosintácticas da norma que fan que o *token* resultante coincida na súa grafía con un xa existente no lexicón pero para o cal se lle outorgan análises diferentes.

De todos eles, o que máis chama a atención é o que, por partida dobre, se produce nunha noticia da sección *Campus*, en concreto na noticia GH2007-03-17/29.etq.xml do CORGA:

> <expresión>Así, en Infantil matricularanse uns 22.000 nenos, que abranguen unhas 14.000 familias; uns 2.140 en Primaria, dunhas 1.430 familias; uns 13.260 niso, dunhas 8.800 familias; e uns 5.110 en Bacharelato, de 3.400 familias.</expresión>
>
> <expresión>Os prazos para formalizar a matrícula prolongaranse do 20 ao 30 de xuño en Infantil e Primaria, e do 25 de xuño ao 10 de xullo niso e Bacharelato.</expresión>

O noso etiquetador segmenta e caracteriza ben os dous elementos integrantes da contracción pero, como é lóxico, non

consegue dar coa análise correcta. Obviamente, queríase facer referencia á etapa educativa ESO pero a casualidade de que antes houbese unha preposición "en" e logo a sigla ESO provoca que a secuencia "en ESO" se convirta na contracción da preposición e mais o demostrativo neutro, resultado probable da aplicación dun tradutor ou un corrector automático.

O etiquetador tampouco consegue atinar cando o xénero do substantivo empregado difire do que a el lle consta no lexicón:

> <expresión>Despois tivo lugar a mesa redonda no Ateneo da capital ourensá, presentada polo humorista Xose Lois González O Carrabouxo, quen asegurou que se trata de homenaxe "necesario e bo", porque "aínda a calor da presenza de Peña na rúa do Paseo non se foi".</expresión> [GH2007-03-17/26; *Galicia*]

cando se emprega unha forma singular por unha plural:

> <expresión> "Hai unha pregunta que o presidente do Goberno debe responder porque é o que lle interesa aos navarros:</expresión> [GH2007-03-17/63; *España*]

ou cando, en vez dun substantivo, se utiliza unha unidade que formalmente se integra no paradigma verbal:

> <expresión> Despois de todo o mundo policial é un xenero en sí mesmo, o chamado thriller e ata ten entidade propia no mundo das letras con ese xénero chamado novela detectivesca pero... ¿o dos médicos?</expresión> [GH2007-03-17/89; *TV*]

2.3. Non figura no lexicón con ese valor

O 0,47% dos erros -106 dos 1260 totais- débense a que o etiquetador non proporciona entre as alternativas posibles a etiqueta que lle corresponde ao *token* segundo o contexto no que está inserido. En realidade, non deberiamos contabilizar os fallos recollidos baixo esta epígrafe no debe do etiquetador -senón no do equipo lingüístico- posto que se no lexicón non consta unha determinada análise o etiquetador non nola pode proporcionar.

Estas deficiencias morfolóxicas son emendables, basta con implementar para cada unha das entradas afectadas as características formais das análises das que carece.

Vexamos a continuación como estas carencias inciden en diversos aspectos da caracterización global.

Afecta á identificación da clase de palabra en casos como:

> <expresión>A Consellería de Medio Rural cuadriplicará este ano os cursos para mellorar a preparación dos máis de 4.600 efectivos do operativo de extinción de incendios forestais, co obxectivo de mellorar a súa eficacia, despois de que o propio departamento recoñecese que a actual formación era "insuficiente e deficitaria". [analizouno como adxectivo no sitio de substantivo] [GH2007-03-17/1; *Galicia*]

> <expresión>A creación do Cegadi e da figura do asistente persoal, estratexias pioneiras</expresión> [consideraronos substantivos en vez de sigla e adxectivo, respectivamente] [GH2007-03-17/32; *Cultura-Sociedade*]

> <expresión>Este foro internacional permitiu tamén recordar que a explotación sexual infantil está moi ligada á situación de desigualdade da que seguen sendo vítimas moitas mulleres en todo o mundo.</expresión> [Identificouno como participio en vez de adxectivo] [GH2007-03-17/37; *Cultura-Sociedade*]

atínguelles aos atributos:

> <expresión>No entanto, admitiu que esta modernización non se levará a cabo "na súa totalidade" este ano, cun sistema de "alta eficiencia".</expresión> [Clasificouna como locución conxuntiva en lugar de adverbial] [GH2007-03-17/6; *Galicia*]

> <expresión>Máis de 650.000 iraquís faleceron catro anos despois da famosa foto de Bush, Blair e Aznar que marcou o inicio do conflito polas "armas de destrución masiva".</expresión> [Tratouno como adverbio nuclear en vez de modificador] [GH2007-03-17/70; *Internacional*]

repercute no lema:

> <expresión>A produtora do inimitable "Loco de la colina" alega que non poderá seguir "desenvolvendo a súa profesión" por

> unha doenza</expresión> [Remitiuno ao lema "produtor" en vez de a "produtora"] [GH2007-03-17/87; *Opinión*]

e finalmente pode tamén implicar a segmentación, a etiquetación e o lema da unidade, como nos dous exemplos que seguen[19]:

> <expresión>O Consello de Ministros deu onte o visto e prace a obras de urxencia, por valor de case 1,5 millóns de euros, que foron aprobadas polo Ministerio de Medio Ambiente para restaurar zonas afectadas polos incendios forestais que se produciron o pasado verán.</expresión> [GH2007-03-17/3; *Galicia*]

> <expresión>O descubrimento dunha placa conmemorativa nunha céntrica rúa de Ourense e unha mesa redonda centraron onte os actos conmemorativos en honra ó nacionalista ourensán Xosé Enrique Rodríguez Peña, que faleceu o verán pasado e que foi ex conselleiro no goberno tripartito de Fernando González Laxe, ex deputado e ex candidato do BNG á alcaldía ourensá.</expresión> [GH2007-03-17/26; *Galicia*]

2.4. Segmentacións erróneas

Os 228 casos recollidos -1,02%- converten este apartado no terceiro máis deficiente do etiquetador. Porén, se lembramos a existencia das contraccións, das formas verbais con pronomes enclíticos, os nomes propios e numerais multipalabra e, sobre todo, a grande cantidade de locucións que presenta o galego, podemos afirmar que esa porcentaxe é irrisoria. Con todo, a medida que o corpus de adestramento se vaia incrementando, dispoñamos dunha gramática formal de propios e dunha regra de construción dos numerais multipalabra irán tamén diminuíndo as segmentacións fallidas.

Como cremos que é interesante e que dá unha idea bastante aproximada da valía deste etiquetador, deseguido clasifi-

[19] Consideramos que "visto e prace" e "mesa redonda" son substantivos multipalabra pero ata o momento non introducimos no lexicón nin estudamos a gramática formal de máis unidades multipalabra que as locucións.

caremos e comentaremos as diferentes clases de fallos segmentais que se producen.

2.4.1. Con numerais

Ante un numeral multipalabra o preprocesador consulta o lexicón, comproba se algunha das regras introducidas lle afecta e, ao ver que non, segméntao en tantas unidades como numerais distintos haxa implicados. É dicir, para un caso como o seguinte, en vez de identificar o numeral "vinte e dous", considera que son dous numerais coordinados pola conxunción "e" e ofrece unha análise para "vinte" e outra para "dous"[20].

> <expresión>Un mozo de vinte e dous anos, residente en Dos Hermanas (Sevilla), levou ó seu pai ó Xulgado porque este se negaba a subirlle a súa paga para extras, que chegaba ós 150 euros ó mes (a familia xa lle cubría ademais os gastos básicos).</expresión> [GH2007-03-17/86; *Opinión]*

2.4.2. Con propios

De entrada, queremos indicar que ata agora no lexicón de XIADA non hai nomes propios nin nos documentos do CORGA se marcan de ningún xeito, polo que o etiquetador só dispón dunha pequena lista externa de antropónimos e topónimos e dunha serie de propios que foron detectados automaticamente procesando os textos, extraéndoos a partir de unidades que aparecen en maiúscula no interior das oracións. Temos previsto modificar e completar o módulo dos propios pero polo momento non acometemos esta tarefa.

A maioría dos erros producidos neste apartado débense á presenza de elementos de enlace -preposicións, contraccións, conxuncións e signos de puntuación- no interior do propio multipalabra que provocan que a unidade se rompa ante a súa aparición. Vexamos algúns exemplos:

Asociación pola Defensa Ecolóxica de Galicia

---

[20] Xa se fixeron progresos para que o sistema recoñeza os numerais multipalabra, polo que en versións posteriores do analizador apreciarase unha melloría considerable neste apartado.

George W. Bush

Creación e Difusión Cultural

Pacto Local en Galicia: xestión dos municipios no futuro

Outro dos factores que impiden unha correcta identificación e caracterización do propio é o formato no que este aparece. Se as iniciais do propio multipalabra están en maiúsculas e non hai elementos do apartado anterior, o recoñecemento é atinado pero se só presenta maiúsculas o primeiro elemento, a segmentación, e en consecuencia a etiquetación, vai ser defectuosa. Exs.:

Lei orgánica de educación

Un conto de inverno

Lei orgánica para a igualdade efectiva de mulleres e homes

2.4.3. Con contraccións

Nas dúas secuencias seguintes recollemos os dous únicos casos do subcorpus avaliado nos que o etiquetador erra na segmentación do token "á", e en consecuencia na súa caracterización morfolóxica. Considera que é o substantivo en vez da contracción. A primeira oración aparece no titular e a segunda no corpo da noticia.

> <expresión>As temperaturas no hemisferio norte de decembro a febreiro foron case un grao superiores á media do século XX</expresión> *[GH2007-03-17/45; Cultura-Sociedade]*

> <expresión>O organismo do Goberno de Estados Unidos indicou na súa páxina web que no período que vai de decembro a febreiro, as temperaturas foron de 0,72 graos centígrados superiores á media do século XX.</expresión> *[GH2007-03-17/45; Cultura-Sociedade]*

Na próxima oración dáse o caso contrario; o etiquetador deu como válida a alternativa da contracción en lugar do substantivo:

> <expresión>¿Fíxoselle un gran favor ó xénero feminino, ou cortáronselles as ás para chegar ós postos máis altos como está demostrando a diario coa súa puxanza, tesón, traballo e vontade?</expresión> *[GH2007-03-17/81; Opinión]*

A confusión entre o pronome átono proclítico de primeira plural e a contracción da preposición e o artigo plural, *token* "nos", só se produce nunha ocasión:

> <expresión>Por iso, se non se consegue maioría absoluta nas eleccións e se aspira a ser un partido no goberno "non nos queda máis remedio" que pactar, aínda que "nuns poderemos e noutros non".</expresión> *[GH2007-03-17/23; Galicia]*

Esta análise en concreto, chamounos especialmente a atención e decidimos indagar un pouco máis. A nosa sorpresa foi considerable ao ver que a análise era impecable se retirabamos as comiñas. Confirmamos algo que xa intuïamos con outros erros: a presenza ou ausencia de comiñas altera o recoñecemento. Este signo ortográfico altera a gramática "natural" da oración e iso dificulta as tarefas de recoñecemento e caracterización.

Finalmente, debemos constatar neste apartado outra deficiencia na etiquetación automática realizada. Trátase da forma "daquela" para a que dá como válida a análise de contracción e non a de adverbio que lle corresponde:

> <expresión>O caso é que, ironicamente, ese pai ausente preocupábase daquela polo incerto futuro que nos esperaba e agora pensa que as mulleres temos xa gañada a batalla.</expresión> *[GH2007-03-17/80; Opinión]*

Tomamos nota dos erros e en próximas versións veremos se este comportamento se corrixe ao aumentar o corpus de adestramento ou se é necesario estudar outras alternativas para a súa solución.

2.4.4. Con formas verbais con pronomes enclíticos

Os seis casos que imos tratar aquí son os únicos, de todo o xornal, nos que se produce unha segmentación errónea e, en consecuencia, tamén etiquetación fallida onde está implicada unha posible forma verbal cun pronome enclítico.

Na metade dos casos desglosa a unidade en compoñentes cando correspondería un só elemento. No primeiro deles, a inusitada aparición de formas verbais en primeira persoa singular nos textos xornalísticos quizais explique a escolla a favor da terceira persoa do singular do presente de indicativo do verbo ler co pronome de acusativo singular de terceira -[Vpi30s ler] / [Raa3ms o]- en vez da primeira persoa do presente de indicativo [Vpi10s ler]:

> <expresión>Onde leo a nova non se ofrecen máis datos:</expresión> [GH2007-03-17/86; Opinión]

No segundo exemplo deste subapartado a confusión prodúcese entre dúas posibles análises: unha como infinitivo de "rever" seguido do enclítico de terceira acusativo feminino e outra como terceira persoa do singular do presente de indicativo de "revelar". Obviamente, escolleu a alternativa non válida.

> <expresión>Un libro da EGAP revela que nas deputadas priman as divorciadas e as solteiras</expresión> [GH2007-03-17/52; Cultura-Sociedade]

No último caso, a brevidade da secuencia e o non haber nada antes -é un titular- confunde o etiquetador que se decanta por un imperativo seguido de clítico de acusativo -[V0p20s ler] / [Raa3fp o]- no sitio da análise correcta como substantivo -[Scfp lea]-:

> <expresión>Leas interminables</expresión> [GH2007-03-17/17; Galicia]

Só nun caso o etiquetador opta pola análise simple da unidade en lugar de considerar dous elementos. Así, na seguinte expresión a unidade "fana" é considerada forma verbal de "fa-

nar" -[Vpi30s fanar]- en vez de terceira do plural do presente de indicativo de "facer" e mais o pronome de terceira acusativo feminino singular -[Vpi30p facer] / [Raa3fs o]-.

> <expresión>O secretario xeral da Asociación Unificada da Garda Civil (AUGC), Joan Miquel Perpinyá, mostrou onte a súa satisfacción polos dous proxectos de lei aprobados polo Consello de Ministros xa que, na súa opinión, "cambian absolutamente a Garda Civil, modernízana e fana máis eficaz".</expresión> [GH2007-03-17/67; España]

No seguinte exemplo as categorías implicadas son as mesmas, verbo e pronome, e na forma verbal só difire o atributo de número. É na caracterización do pronome onde se dan máis diferenzas. O etiquetador ten que escoller se está diante de "recorda" seguido de "nos" ou de "recordan" e mais "os". A análise do propio que precede á unidade en cuestión non lle proporciona ao etiquetador ningunha información sobre o número do substantivo, fundamental neste caso, e optou pola segmentación equivocada.

> <expresión>Caxigueiro recórdanos a escasa protección das nosas costas a través de A linguaxe da memoria, un gravado cos nomes dos máis perigosos afundimentos do litoral galego: Prestige, Urquiola, Mar Exeo.</expresión> [GH2007-03-17/12; Galicia]

No último caso desta serie prodúcese unha cadea de erros: o etiquetador decide que "díxome" é un substantivo porque antes del cataloga "médico" como adxectivo e, en consecuencia, a palabra seguinte considéraa relativo en vez de conxunción:

> <expresión>"O médico díxome que debín de perder o DIU, que como é unha cousa pequena puiden perdelo polo servizo sen decatarme", recordou.</expresión> [GH2007-03-17/41; Cultura-Sociedade]

Cómpre facer aquí unha puntualización. A estrutura sintáctica recoñecida polo programa non é inusual no corpus de adestramento:

Artigo + Adxectivo + Substantivo + Relativo: 64 casos

Artigo + Substantivo + Verbo + Conxunción: 183 casos

con todo, a análise non é a axeitada polo que será preciso estudar que factores determinan este comportamento e ver como poderemos corrixilo.

2.4.5. Con locucións

No lexicón de XIADA constan 583 locucións, das cales só 252 están caracterizadas como seguras, é dicir, a súa aparición non presenta ambigüidades segmentais (ex. *non obstante, a carranchapernas, a eito, etc.*). Nas 331 restantes existe ambigüidade e o etiquetador ten que decidir se está ante unha locución ou non (ex. *en breve, con todo, etc.*)

No subcorpus avaliado figuran 235 ocorrencias de locucións. De todas elas, o etiquetador equivocouse en 51 casos, o que supón o 21,7% de erro. Vexamos a súa distribución:

	Galicia	Campus	Cultura	España	Internacional	Economía	Opinión	TV
Nº de palabras	5859	1015	6489	2523	821	761	3795	949
Nº de locucións totais	52	11	74	23	11	5	47	12
Nº de locucións falladas	8	1	13	4	0	2	21	2
Porcentaxe de erro	15,3%	9,09%	17,56%	17,39%	0%	40%	44,68%	16,66%

Unha vez máis comprobamos como na sección de Opinión conseguimos os peores resultados. De todos xeitos, as cifras obtidas estannos indicando que debemos actuar de inmediato sobre este campo, creando regras precisas que desfagan o maior número posible das ambigüidades segmentais existentes e facilitarlle así o labor ao etiquetador. É dicir, incidir, a través de re-

gras, naquelas locucións que foron caracterizadas como inseguras.

No exemplo seguinte o etiquetador desagregou a locución nos seus constituíntes e non a tratou como unha unidade multipalabra:

> <expresión>Barreras seguirá á espreita por se Navantia se abre ó civil en breve</expresión> [GH2007-03-17/13; Galicia]

Este erro podería corrixirse cunha regra na que se indique que só se a continuación aparece un substantivo non a considere locución adverbial: en breve + S -> non La0

No caso seguinte o erro producido é xustamente o contrario: considera locución o que debería ser desglosado en distintas unidades:

> <expresión>Din os analistas que o camiño do bilingüismo efectivo faise máis curto día a día.</expresión> [GH2007-03-17/84; Opinión]

Como locución prepositiva, "camiño de" equivale a "cara a, en dirección a" e para que teña este valor non pode ir precedido de determinante [D], adxectivo [A] ou calquera outra forma que funcione como determinante. É dicir, o substantivo "camiño" non pode aparecer precedido de artigo, adxectivo ou ningunha outra forma cuxa etiqueta teña na segunda posición un "d" correspondente ao valor "determinante" do atributo valor. A regra complícase cos numerais onde debemos lembrar que ese atributo ocupa a terceira posición: [D, A, Ed, Td, Md, Id, N?d, Gd] + camiño de -> non Lp0

2.5. Erros na asignación dos atributos

No referido aos atributos é interesante destacar que a confusión prodúcese maioritariamente só sobre un deles, sendo os valores do xénero (masculino, feminino, ambiguo, non aplica) os que máis problemas lle causan ao etiquetador mentres que no caso de errar en dous atributos, estes son xénero e número (singular, plural, ambiguo, non aplica).

Na táboa seguinte aclaramos a nosa afirmación coas cifras resultantes da extracción e clasificación dos fallos referentes aos atributos:

	Galicia	Campus	Cultura	España	Internacional	Economía	Opinión	TV
Nº de palabras	5859	1015	6489	2523	821	761	3795	949
Nº de erros totais	295	51	375	134	40	39	257	75
Nº de erros nos atributos	107	20	113	40	26	11	63	22
Nº de erros en 1 atributo	89	17	97	33	21	7	51	17
Nº de erros en 2 ou máis atributos	18	3	16	7	5	4	12	5
Porcentaxe de erro sobre o total de erros	36,27%	39,21%	30,13%	29,85%	65%	28,2%	24,51%	29,33%
Porcentaxe de erro nos atributos	1,82%	1,97%	1,74%	1,58%	3,16%	1,44%	1,66%	2,31%

A táboa que segue recolle os atributos en conflito, distribuídos por sección:

	Galicia	Campus	Cultura	España	Internacional	Economía	Opinión	TV
valor (nuclear | modificador)	6	------	4	8	4	------	10	3
valor (determinante | non determinante)	7	5	4	1	------	------	9	4
xénero	38	9	60	13	9	5	22	7
número	9	2	10	4	5	1	4	2
número do posuidor	12	1	12	3	3	1	1	---
propio | común	10	------	3	2	------	------	------	---
caso [pronome]	4	------	1	------	------	------	------	---
tipo [locución ou pronome]	2	------	1	------	------	------	1	---
persoa [3ª por 1ª]	------	------	2	------	------	------	2	---
tempo	------	------		1	------	------	2	---
modo	1	------	------	1	------	------	------	1
xénero e número	15	3	14	7	5	4	6	3
tempo e modo	2	------	2	------	------	------	2	1
caso e xénero	1	------	------	------	------	------	3	---
tempo, modo e persoa	------	------	------	------	------	------	1	1

Deseguido recollemos todos os *tokens* que se corresponden con fallos nos atributos correspondentes a *número do posuidor, caso, tipo, persoa, tempo, modo, tempo e modo, caso e xénero* e, finalmente *tempo, modo e persoa*. Para os demais atributos *-valor,*

xénero, número, tipo no substantivo e *xénero e número*- eliximos unha mostra representativa por non ser posible incluír todos os casos.

Como se pode deducir polos exemplos recollidos, os erros na asignación de atributos só se producen cando a forma implicada presenta na súa etiqueta algún valor ambiguo, para cuxa caracterización só o contexto –e non sempre– nos aclara que valor, dos posibles, lle corresponde.

Atributo implicado	Exemplos de formas gráficas nas que se comete o fallo
valor (nuclear | modificador)	máis, non, antes, mesmo, menos, mellor, seriamente, etc.
valor (determinante | non determinante)	aquel, outros, calquera, oito, dúas, primeira, 2.140, 5.110, etc.
xénero	axentes, importante, que, lle, portavoces, dezaoito, eu, demais, etc.
número	que, PXOM, 80, Galicia, 2002, José Saramago, prevén, etc.
nº do posuidor	seu, seus, súa, súas
propio | común	Alcaldía, Barreras, Barreiro, Bugallo, Saramago, Falanxe, etc.[1]
caso [pronome]	nos, me
tipo [locución ou pronome]	ela, nós, mentres tanto
persoa [3ª por 1ª]	tiña, era, tería
tempo	ampliamos, lía, cría
modo	estea
xénero e número	que, ninguén, PXOM, Navantia-Fene, Méndez Romeu, etc.
tempo e modo	privatizara, advertira, fora, pedira, nacera
caso e xénero	me, te
tempo, modo e persoa	mira, parece

[21] Todas estas unidades aparecen a comezo de oración, o que implica a presenza de maiúscula inicial e a ambigüidade no tipo, polo que o etiquetador ten que decidir se o substantivo é propio ou común.

2.6. Erros na atribución de categoría

Se nos centramos nos problemas de recoñecemento automático da clase de palabra morfolóxica, comprobamos que as confusións producidas son, na maioría dos casos, lóxicas dada a proximidade de certas clases ou explicables pola limitación que presenta o etiquetador, baseado nun modelo de Markov [Markov, 1913: 153-162] de grao 2, ao ter en conta conceptualmente só a forma e etiqueta das dúas palabras anteriores e posteriores. A seguinte táboa recolle os erros producidos por asignar unha categoría incorrecta. Na primeira columna preséntanse as clases que entran en conflito[22] e nas seguintes as seccións onde se producen eses erros e exemplos extraídos dos documentos:

	Galicia	Campus	Cultura	España	Internacional	Economía	Opinión	TV	Exemplos
adverbio | conxunción							4		nin, pois
adverbio | indefinido	2		2			1	7	1	algo, máis, tal, mesmo, pouco, etc.
adverbio | numeral			1						primeiro
adverbio | pronome	3								el
adxectivo | adverbio	1						2		libre, medio
adxectivo | verbo (participio)	14	2	6	5	2	1	1	1	postas, afectados, achegadas, etc.
adxectivo | substantivo	21		31	7		2	27	6	efectivos, tácticas, técnicas, etc.
artigo | indefinido	1					1	1		uns, un
artigo | numeral	1	1	3	1			1	1	unha, un

[22] Anotamos as dúas clases de palabras implicadas, a incorrecta -asignada polo etiquetador- e a correcta -asignada polo lingüista- pero a orde de aparición non significa que todos os casos se dean na dirección apuntada senón que integran fallos nos dous sentidos. Por exemplo, a confusión entre conxunción e relativo presenta algún caso en todas as seccións do xornal, non obstante nunhas ocasións caracterízase como conxunción o que é relativo e noutras como relativo cando é conxunción.

artigo | preposición	8	2	6		1	1	4		a
artigo | pronome	4		2	1			2		o, os, as, a
conxunción | interrogativo	3		1			2	2	1	que, como
conxunción | numeral			1						segundo
conxunción | preposición			2	1					segundo, senón
conxunción | relativo	22	3	30	18	1	3	19	2	que, como, cando
conxunción | pronome	7	1	3	3			5	1	se
indefinido | adxectivo								1	propia
indefinido | numeral			1						un
numeral | sigla	1								A-8
preposición | adverbio			1				1		ata
preposición | pronome	1		1						a
preposición | substantivo		1							cabo
relativo | interrogativo	1		2	1		3			que, como, quen, onde
sigla | numeral			2				1	2	XX, Leb2
substantivo | adverbio			1	2					ben, si, non
substantivo | indefinido							1		propios
substantivo | numeral							1		segundos
substantivo | sigla			8					3	Cegadi, Igape, M.I.R., C.S.I., etc.
substantivo | verbo	5		7	2			10	1	perda, poder, lixo, conta, etc..
verbo | adxectivo			1				1		turbia, digna

2.7. Erros na atribución do lema

Como podemos observar na táboa xeral de erros, os 6 casos recollidos baixo esta epígrafe só constitúen o 0,02% de fallos, non obstante o noso obxectivo é acadar o 100% de efectividade.

Debemos distinguir neste apartado de erros na adxudicación do lema dous tipos ben diferenciados:

1) O etiquetador falla na atribución do lema pero acerta coa etiqueta. Estes erros foron contabilizados no apartado e) da

táboa xeral onde se recollen os erros producidos ao atribuír o lema.

2) O etiquetador equivócase ao asignar o lema porque fallou na etiquetación. Estes errores incluímolos nesta sección para a súa explicación -non nas estatísticas, onde van ao apartado c), correspondente a erros nos atributos- porque son os únicos casos nos que o erro na etiquetación provoca un erro na asignación do lema.

Vexamos as diferencias entre estes dous tipos de equivocacións cos casos reais extraídos do corpus avaliado.

1) O etiquetador falla na atribución do lema pero acerta na etiqueta: a etiqueta é idéntica e a caracterización só difire no lema[23].

Un *token* pode corresponder a máis dun lema pertencente á mesma categoría gramatical, -factor que non teñen en conta os etiquetadores-, pero o normal é que o contexto nos aclare a cal corresponde.

Na noticia 12 da sección *Galicia* do subcorpus avaliado temos un caso no que o contexto non nolo aclara:

> <expresión>Algunhas das creacións que conforman a mostra 'ArtNatura', exposta no claustro do Rosalía de Castro</expresión> [GH2007-03-17/12; Galicia]

É unha das ocorrencias de moitas formas verbais coincidentes dos derivados de *poñer* e *pór* nos que o problema de asignación de lema non é do etiquetador senón do lingüista. Por que lema decantarnos? A solución tomada foi considerar que o lema non marcado ía ser o de *poñer* e que só se no resto do documento aparecía algunha das formas distintivas dos derivados de *pór* optariamos por este.

Nos seguintes casos non atopamos ningunha explicación racional e comprobamos que o etiquetador ante unha forma

[23] O etiquetador para as súas escollas só ten en conta a forma e a etiqueta, nunca os lemas.

cunha etiqueta unívoca remisible a máis dun lema escolle un aleatoriamente.

Isto ocorre coas formas do tema de pretérito *fomos* e *foi*, coincidentes para os verbos *ser* e *ir* nas que opta sempre por atribuílas ao lema *ser*.

> <expresión>Pola súa parte, Miguel Anxo Fernández Lores quixo destacar que o BNG "é a única forza deste concello que ten un proxecto claro de cidade, que fomos executando nestes últimos oito anos de goberno, con resultados positivos á vista de todos".</expresión> [GH2007-03-17/20; Galicia]

> <expresión>Despois tivo lugar a mesa redonda no Ateneo da capital ourensá, presentada polo humorista Xose Lois González O Carrabouxo, quen asegurou que se trata de homenaxe "necesario e bo", porque "aínda a calor da presenza de Peña na rúa do Paseo non se foi".</expresión> [GH2007-03-17/26; Galicia]

Paradoxalmente, para *fun* remítea ao verbo *ir* e non a *ser*:

> <expresión>fun coordinador desa "normalización lingüística", durante anos.</expresión> [GH2007-03-17/81; Opinión]

Por último, dúas mostras na clase dos substantivos. Na primeira, a forma "cores", coa etiqueta substantivo común feminino plural -Scfp- pode remitir ao lema "cor" ou a "core" e o etiquetador decide, sen ningún criterio, que é "core". Na segunda, para a forma "raíña", presente nun refrán, o etiquetador considerou que era o feminino de "rei" e non o sinónimo de "raiola":

> <expresión>Caen as árbores da Amazonía, desécanse as lagoas azuis e os encoros apertados entre montañas, morren plantas coñecidas, apáganse as cores da infancia.</expresión> [GH2007-03-17/85; Opinión]

> <expresión>Marzo, marzán, cara de can: pola mañá sol da raíña e pola noite afeitan as barbas cunha fouciña.</expresión> [GH2007-03-17/79; Opinión]

2) O etiquetador equivócase ao asignar o lema porque fallou na etiquetación.

Na seguinte oración atribúe a forma *prevén* ao lema *previr* e non a *prever* como lle corresponde:

> <expresión>Non obstante, a decisión non satisfai a todos e onte mesmo CCOO xa advertiu de que existe "preocupación e temor" entre os traballadores de Fadesa, polo seu futuro laboral, pois prevén que "en dous ou tres anos" os novos donos desmantelen os 450 empregos da Coruña.</expresión> [GH2007-03-17/36; Cultura-Sociedade]

A explicación do fallo radica en primeira instancia na distancia existente entre o verbo e o seu suxeito e en segunda na estatística: no corpus de adestramento figuran 11.752 formas verbais en terceira persoa do singular fronte a 3.998 de terceira plural.

Algo semellante aconteceu coas formas "cría" e "lía", ambas terceira persoa do singular do copretérito de indicativo dos verbos "crer" e "ler", para as que o etiquetador optou pola etiqueta que as caracterizaba como presente dos verbos "criar" e "liar", respectivamente. Consultamos o corpus de adestramento e cremos que as 5.521 ocorrencias de Vpi30s fronte ás 467 de Vii30s influíron no resultado.

> <expresión>Vicente Risco cría que os debuxos de vexetais que antes facían os nosos canteiros nas tampas dos defuntos, poden ter algunha relación coa resurrección, pois as árbores teñen varias vidas, tantas como anos...</expresión> [GH2007-03-17/79; Opinión]

> <expresión>Aquela nena que lía os xornais da época coa sensación de estar a vivir unha páxina da historia sabía que o camiño estaría condicionado por nacer muller, e non poucas veces imaxinou as vantaxes de ser home.</expresión> [GH2007-03-17/80; Opinión]

É a mesma hipótese que consideramos para as tres ocorrencias de *estea*, presente de subxuntivo de *estar*. Para todas elas remite ao lema *estear* (=apoiar; escampar; secar) porque a eti-

queta correspondente ao presente de indicativo -Vpi30s- aparece en 5.521 ocasións mentres que a de presente de subxuntivo -Vps30s- só o fai en 544:

> <expresión>Medio Rural considerou necesario que esta formación estea acompañada dunha divulgación das medidas postas en marcha por este departamento na loita contra os incendios forestais: o Decreto de xuño e o proxecto de Lei de Prevención e Defensa.</expresión> [GH2007-03-17/4; Galicia]
>
> <expresión>Para a AUGC, segundo declarou a Efe Perpinyá, "é un día feliz" por "estas dúas normas, que farán que o persoal estea máis motivado".</expresión> [GH2007-03-17/67; España]
>
> <expresión>Telecinco acaba de lograr que unha teleserie que estaba moi mal de saúde, M.I.R., estea a recuperarse logo dun cambio de día e de subministrar varias doses de paciencia.</expresión> [GH2007-03-17/89; TV]

3. Conclusións

Se descartamos os erros non atribuíbles ao etiquetador -por non proporcionar lema ou equivocarse ao asignalo, por erros lingüísticos textuais e por non figurar unha análise no lexicón- vemos como as porcentaxes de acerto do etiquetador soben, situándose a media nun 95,44%.

	Galicia	Campus	Cultura	España	Internacional	Economía	Opinión	TV	Total
palabras	5859	1015	6489	2523	821	761	3795	949	22 212
erros totais	246	37	303	108	35	36	195	54	1014
segmentacións erróneas	43	7	79	27	5	11	44	12	228
erros nos atributos	107	20	113	40	26	11	63	22	402
erros na categoría	96	10	111	41	4	14	88	20	384
porcentaxe de erro	4,19%	3,64%	4,66%	4,28%	4,26%	4,73%	5,13%	5,69%	4,56%
porcentaxe de acerto	95,81%	96,36%	95,34%	95,72%	95,74%	95,27%	94,87%	94,31%	95,44%

Esta cifra permítenos ser moi optimistas sobre a porcentaxe de acerto que acadará o noso etiquetador no futuro posto que é moi alentador que, cunha morfoloxía bastante complexa e un etiquetario de 375 etiquetas distintas reais, consigamos xa un 95% de acerto para o galego actual, sabendo, ademais, que dispoñemos dunha marxe de mellora considerable nos numerais e propios multipalabra e no recoñecemento das locucións.

Finalmente, queremos comparar os resultados do noso etiquetador cos doutros etiquetadores. Para iso debemos descartar todos os fallos relacionados coa lematización e a segmentación. Os primeiros, porque nos resultados doutros etiquetadores non se recollen os erros de lematización; os segundos, porque eses etiquetadores non realizan ningún tipo de segmentación das distintas unidades léxicas presentes nun texto.

Por iso, para empregar os mesmos criterios de validación na comparativa, só podemos ter en conta os erros producidos en calquera punto da etiqueta, incluídos os que se cometen por non figurar no lexicón con ese valor.

Como se pode observar na última táboa, onde reflectimos só os erros relacionados coa etiquetación, a porcentaxe final de acerto que acadamos con esta versión do etiquetador é semellante ás dispoñibles para o inglés ou o castelán: 95,99%.

	Galicia	Campus	Cultura	España	Internacional	Economía	Opinión	TV	Total
palabras	5859	1015	6489	2523	821	761	3795	949	22 212
erros totais	228	30	247	95	32	27	181	52	892
erros nos atributos	107	20	113	40	26	11	63	22	402
erros na categoría	96	10	111	41	4	14	88	20	384
por non estar con ese valor	25	---	23	14	2	2	30	10	106
porcentaxe de acerto	96,11%	97,05%	96,2%	96,24%	96,11%	96,46%	95,24%	94,53%	95,99%

Referencias bibliográficas

Graña Gil, J. (2000): *Técnicas de Análisis Sintáctico Robusto para la Etiquetación del Lenguaje Natural*, Tese doutoral, Departamento de Computación, Universidade da Coruña, 2000.

Jurafsky, D. e J.H. Martin (2000): *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2000.

Leech, G. e A. Wilson (1994): "Morphosyntactic Annotation", Draft-Work In Progress, EAGLES Document EAG-CSG/IR.T3.1. en N. Calzolari e J. M. McNaught (eds.), *EAGLES Interim Report* EAG-EB-IR-2, 1994.

Sampson, G. (1994): *The SUSANNE corpus, release 3, 04/04/1994.* School of Cognitive & Computing Sciences, University of Sussex, Falmer, Brighton (England).

Markov, A. A., (1913): "An example of statistical investigation in the text of Eugene Onyegin illustrating coupling of tests in chains" en *Proceedings of the Academy of Sciences, St. Petersburg,* vol. 7 of VI, 1913, pp. 153-162.

Bibliografía complementaria

Alarcos Llorach, E. (1994): *Gramática de la lengua española*, Madrid, Espasa Calpe, 1994.

Álvarez, R., H. Monteagudo e X.L. Regueira (1986): *Gramática Galega,* Vigo, Galaxia, 1986, 4ª ed.

Álvarez, R. e X. Xove (2002): *Gramática da lingua galega*, Vigo, Galaxia, 2002.

Álvarez, R. (1994): "Secuencias de pronomes átonos en galego moderno" en Ramón Lorenzo (ed.), *Actas do XIX Congreso de lingüística e filoloxía románicas,* pp. 247-265, vol. VI, A Coruña, Fundación Pedro Barrié de la Maza Conde de Fenosa, 1994.

Bosque, I. (1991): *Las categorías gramaticales. Relaciones y diferencias*, Madrid, Síntesis, 1991.

Bray, T.; J. Paoli e C.M. Sperberg-McQueen (eds.): *Extensible Markup Language (XML)* dispoñible en http://www.w3.org/TR/REC-xml [Consulta: 16/11/2006]

Burnage, G. e D. Dunlop (1993): "Encoding the British National Corpus", en J. Aarts, P. de Haan e N. Oostdijk (eds.), *English Language Corpora: Design, Analysis and Exploitation*, Amsterdam, Rodopi, 1993.

Burnard, L. (1997): "The Text Encoding Initiative's. Recommendations for the Encoding of Language Corpora: Theory and Practice" [Seminario de Industrias de la Lengua, Soria, 14-18 de xullo de 1997]

Chacón Calvar, R. e M. Rodríguez Alonso (1993): *Diccionario crítico de dúbidas e erros da lingua galega*, Sada, Ediciós do Castro, 1993.

Costa Casas, X.X., M. Anxos González Refoxo, C.C. Morán Fraga e X.C. Rábade Castiñeira (1998): *Nova Gramática para a aprendizaxe da língua*, A Coruña, Vía Láctea, 1988.

Díaz Regueiro, M. (1992): *Os verbos galegos*, Consellería de Educación e Ordenación Universitaria. Dirección Xeral de Política Lingüística, Santiago de Compostela, 1992.

Freixeiro Mato, X.R. (2000): *Gramática da lingua galega. Morfosintaxe*, A Coruña, A Nosa Terra, 2000.

Ide, N. (coord): "Corpus Encoding Standard. Versión 1.4." dispoñible en http://www.cs.vassar.edu/CES [Consulta: 16/11/2006]

Instituto da Lingua Galega e Real Academia Galega (2003): *Normas ortográficas e morfolóxicas do idioma galego*, Santiago de Compostela, ILG/RAG, 1993, 11ª ed.; 1997, 16ª ed.; 2003, 18ª ed.

Instituto da Lingua Galega e Real Academia Galega (1989): *Vocabulario Ortográfico da Lingua Galega. Versión provisional*, [Edición fotocopiada], Santiago de Compostela, 1989.

González González, M. e A. Santamarina Fernández (coords.) (2004): *Vocabulario Ortográfico da Lingua Galega*, Santiago de Compostela / A Coruña, ILG/RAG, 2004.

Morel, J; S. Torner; J. Vivaldi; Ll. de Yzaguirre e M.T. Cabré (1998): *El Corpus de l'IULA: Etiquetaris*, Barcelona, Univer-

sitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada, 1997; 2ª ed. revisada e corrixida, 1998.

Nicolás Rodríguez, R. (1993): *Diccionario dos verbos galegos*, Pontevedra, Edicións do Cumio, 1993.

Reis, R. e J.J. Dias de Almeida (1998): "Etiquetador morfo-sintáctico para o Português" en M.R. Braga Marquillas e M.A. Mota (eds.), *Actas do XIII Encontro Nacional da Associação Portuguesa de Linguística,* pp. 209-221, vol. II, APL-Colibri, Lisboa, 1998.

Sánchez León, F. (1997): "El etiquetado del Corpus de Referencia del Español Actual (CREA)" [Seminario de Industrias de la Lengua, Soria, 14-18 de xullo de 1997]

TÁBOAS

			m-masculino			**1**-primeira
			f-feminino	**s**-singular		**2**-segunda
		c- coordinante	**a**-masculino ou feminino	**p**- plural		**3**-terceira
		s- subordinante	**n**-neutro	**a**- singular ou plural	**c**-comparativo **s**-superlativo	**a**-primeira ou terceira
		0- non aplica	**0**-non aplica	**0**-non aplica	**0**-non aplica	**0**-non aplica
Categoría	**Tipo**	**Subtipo**	**Xénero**	**Número**	**Grao**	**Persoa**
Substantivo (**S**) 1	**c**-común **p**-propio 2		m / f / a / **0** 3	s / p / a / 0 4		
Adxectivo (**A**) 1			m / f / a 3	s / p / **a** 4	c / s / 0 2	
Verbo (**V**) 1			m / f / **0** 5	s / p / **0** 6		1 / 2 / 3 / a / 0 4
Preposición (**P**) 1						
Conxunción © 1	**c**-coordinante **s**-subordinante 2					
Adverbio (**W**) 1	**n**-nuclear **m**-modificador **a**-nuclear ou modificador **r**-relativo **g**-interrogativo-exclamativo 2					
Artigo (**D**) 1	**d**-determinado **i**-indeterminado 2		m / f 3	s / p 4		
Demostrativo (**E**) 1			m / f / **n** 3	s / p 4		
Relativo (**T**) 1			m / f / a 3	s / p / **a** 4		
Posesivo (M) 1			m / f / a 5	s / p / **a** 6		1 / 2 / 3 3
Indefinido (I) 1			m / f / a / **0** 3	s / **p / a / 0** 4		
Numeral (**N**) 1	**c**-cardinal **o**-ordinal 2		m / f / a 4	s / p 5		
Pronome (R) 1	**t**-tónico **a**-átono 2		m / f / **n / a** 5	s / p / **a** 6		1 / 2 / 3 4
Exclamativo-Interrogativo (**G**) 1			m / f / a 3	s / p / a 4		
Locución (**L**) 1	**a**-adverbial **p**-prepositiva **c**-conxuntiva 2	c / s / 0 3				
Interxección (Y) 1						
Signo de puntuación (**Q**) 1						
Categoría Periférica (**Z**) 1	**f**-fórmula **a**-abreviatura **g**-sigla **s**-símbolo **o**-outros tipos 2		m / f / a / **0** 3	s / **p / a / 0** 4		

Táboa 1: Etiquetario de XIADA

Caso	Tempo Verbal	Modo	Posuidor	Valor
a-acusativo **d**-dativo **i**-impersoal **p**-preposicional **r**-nominativo ou preposicional **l**-acusativo ou dativo **o**-outros casos **f**- formante léxico **n**-nominativo	**p**-presente **i**-copretérito **l**-antepretérito **f**-futuro **e**-pretérito **c**-pospretérito **0**-non aplica	**i**-indicativo **s**-subxuntivo **m**-imperativo **f**-infinitivo **x**-xerundio **p**-participio	**s**-singular **p**-plural **a**-singular ou plural	**n**-non determinante **d**-determinante **a**-determinante ou non determinante
	p / **i / l / f / e** / c / 0	i / s / m / f / x / p		
	2	3		
				d / n / a
				2
				d / n / a
				2
			s / p / **a**	d / n / a
			4	2
				d / n / a
				2
				d / n / a
				3
n / a / d / i / p / r / l / f / o				
3				
				d / n / a
				2

	Galicia	Campus	Cultura	España
palabras	5859	1015	6489	2523
erros	295	51	375	134
erros ao non dar lema	18 casos | 0,30%	12 casos | 1,18%	47 casos | 0,42%	10 casos | 0,39%
segmentacións erróneas	43 casos | 0,73%	7 casos | 0,68%	79 casos | 1,21%	27 casos | 1,07%
erros nos atributos	107 casos | 1,82%	20 casos | 1,97%	113 casos | 1,74%	40 casos | 1,58%
erros na categoría	96 casos | 1,63%	10 casos | 0,98%	111 casos | 1,71%	41 casos | 1,62%
erros ao atribuír lema	3 casos | 0,05%	------	------	------
por non figurar no lexicón con ese valor	25 casos | 0,42%	------	23 casos | 0,35%	14 casos | 0,55%
por erros lingüísticos ou desviacións da norma	2 casos | 0,03%	2 casos | 0,19%	2 casos | 0,03%	1 caso | 0,03%
porcentaxe de erro	5,03%	5,02%	5,77%	5,31%
porcentaxe de acerto	94,97%	94,98%	94,23%	94,69%

Táboa 2: Distribución de erros clasificados por tipo e sección

Internacional	Economía	Opinión	TV	Total
821	761	3795	949	22 212
40	39	257	75	1260
3 casos ǀ 0,36%	------	24 casos ǀ 0,63%	8 casos ǀ 0,84%	122 casos ǀ 0,54%
5 casos ǀ 0,60%	11 casos ǀ 1,44%	44 casos ǀ 1,15%	12 casos ǀ 1,26%	228 casos ǀ 1,02%
26 casos ǀ 3,16%	11 casos ǀ 1,44%	63 casos ǀ 1,66%	22 casos ǀ 2,31%	402 casos ǀ 1,8%
4 casos ǀ 0,48%	14 casos ǀ 1,83%	88 casos ǀ 2,31%	20 casos ǀ 2,10%	384 casos ǀ 1,72%
------	------	3 casos ǀ 0,07%	------	6 casos ǀ 0,02%
2 casos ǀ 0,24%	2 casos ǀ 0,26%	30 casos ǀ 0,79%	10 casos ǀ 1,05%	106 casos ǀ 0,47%
------	------	3 casos ǀ 0,07%	2 casos ǀ 0,21%	12 casos ǀ 0,05%
4,87%	5,12%	6,77%	7,9%	5,67%
95,13%	94,88%	93,23%	92,1%	94,33%