Ferramentas de Recuperación de Textos para Bibliotecas Dixitais: Lematización*

Marisa Moreda Leirado
Ángeles S. Places
Eloy Vázquez Fontenla
Miguel R. Penabad
Universidade da Coruña
lmoreda@udc.es, asplaces@udc.es,
eloy_vazquez_fontenla@yahoo.es, penabad@udc.es

Resumo:

Un dos servizos máis interesantes das bibliotecas dixitais é o que permite a busca de documentos polo seu contido, quere dicir, o que permite buscar aqueles textos que traten dun certo tema. Para que as bibliotecas poidan implementar servizos deste tipo é preciso que existan recursos e ferramentas de recuperación de textos (corpora, dicionarios electrónicos, lematizadores, analizadores morfolóxicos, etc.) desenvolvidas para o idioma en que estean escritos os documentos da biblioteca.

A cantidade e a calidade dos recursos e ferramentas que estean desenvolvidos depende do idioma de que se tratar. O inglés está á cabeceira de todos, e aquí na Península as bibliotecas dixitais de textos escritos en galego son as que teñen máis complicado desenvolveren servizos de busca por contido, xa que non existen até o momento as ferramentas e os recursos de apoio apropiados.

Neste artigo presentamos unha ferramenta de recuperación de textos que foi desenvolvida para o galego, grazas á colaboración de investigadores en Filoloxía Galego-Portuguesa e Informática da Universidade da Coruña. Trátase dun lematizador que foi presentado por primeira vez en 2002, e que nos últimos anos foi optimizado, completado e probado con corpora de diferente natureza para ser usado en servizos de busca por contido de bibliotecas dixitais.

Palabras chave:

Bibliotecas Dixitais, Recuperación de Textos, Lematización.

Abstract:

The ability to search documents by content, i. e., to look for documents dealing with a certain subject, is one of the most interesting services offered by a Digital Library. In order to offer these services, digital libraries need resources and text retrieval tools (such as corpora, *electronic dictionaries, stemmers, or morphological analyzers), which must be developed for the language in which the documents of the library are written.*

* Este traballo está parcialmente financiado por MCYT (PGE e FEDER) a través do proxecto de investigación ref. TIC2003-06593.

The quantity and quality of the developed resources and tools depend on the used language. English has always had a great advantage in this field. On the contrary, in the Iberian Peninsula, Digital Libraries devoted to texts written in Galician have difficulties to develop content search services, since there are not enough tools and resources to do these implementations yet.

This paper shows a Text Retrieval tool for the Galician language, built through a collaboration between Galician–Portuguese Philology and Computer Science researchers from the University of A Corunna. This tool is a stemmer that was first introduced in 2002, and it has been optimized, completed and tested during last years. We have used several different corpora to perform the tests, in order to accurately incorporate content search services in Digital Libraries.

Key words:

Digital Libraries, Text Retrieval, Stemming.

1. Introdución

As primeiras bibliotecas dixitais xurdiron como réplicas das bibliotecas convencionais, no sentido de que se desenvolveron coas mesmas funcionalidades: armacenábase un catálogo que recollía os datos bibliográficos das obras, das revistas etc. (xenericamente documentos) da biblioteca e desenvolvíanse funcionalidades de busca sobre dito catálogo. Así proliferaron as bibliotecas dixitais, sobre todo na Internet (Brisaboa e Fernández 2001).

Máis adiante comezouse a armacenar, á parte dos datos bibliográficos dos documentos, tamén as páxinas dixitalizadas e os textos dos mesmos co fin principal de preservalos e facelos accesíbeis de maneira máis sinxela para un público máis amplo. Foi ao ter dispoñíbel o texto dos documentos cando xurdiu a necesidade de desenvolver procuras máis complexas que as puramente bibliográficas. Á parte de manter a posibilidade de recuperar (isto é, de achar como resultado de procuras) os documentos cun certo título, escritos nunha certa data ou publicados nun determinado lugar (procura polos *metadatos* dos documentos), xorde a necesidade de recuperar documentos que traten sobre certo tema (procura por contido). Investigadores e investigadoras de moi diversas áreas: antropoloxía, historia, filoloxía, socioloxía etc., decatáronse decontado das posibilidades deste novo servizo, debido á axuda que suporía para os seus traballos de investigación, mais é evidente a súa utilidade para todo o público en xeral (Fernández e Saavedra 2004).

Moitas bibliotecas dixitais actuais teñen implementados servizos básicos de procura por contido que, en xeral, non van máis lonxe da busca de documentos que conteñan certa palabra ou cadea de caracteres. Deste xeito, de querermos recuperar aqueles documentos que traten sobre cazarías, poderiamos dar como palabra de procura, por exemplo, "*cazaría*". Bibliotecas dixitais que tiveran unha implementación básica da busca por contido, devolveríannos como resultado da procura uni-

camente aqueles documentos que contiveran exactamente esa palabra. Porén, probabelmente tivésemos interese en recuperar tamén os textos en que aparecesen palabras como: "*cazar*", "*cazador*", "*cazarías*", ou calquera outra variación morfolóxica da palabra de procura. Aínda máis, en ocasións, precisariamos ter a posibilidade de ver a nosa busca ampliada, de forma automática, e poder acceder a documentos con palabras como "*lanza*", "*escopeta*" etc. Quere dicir, documentos que, sen conter exactamente a palabra procurada, tivesen algún dos seus sinónimos, antónimos ou vocábulos relacionados, porque, aínda que indirectamente, tratarían do tema en que estamos interesados/as.

A busca por contido coñécese formalmente como *Text Retrieval* (recuperación de textos), e abrangue un amplo abano de métodos e técnicas orientados a permitir que se recuperen documentos que traten sobre certo tema. Calquera posibilidade de procura máis complexa que a simple busca dunha palabra concreta dentro do texto do documento precisa de ferramentas de recuperación de textos específicas para o idioma en que estean escritos ditos documentos. No exemplo anterior, para poder recuperar a partir da palabra de procura "*cazaría*" documentos con calquera variación morfolóxica da mencionada palabra precisaríase un lematizador. Do mesmo xeito, para poder recuperar documentos que falen indirectamente do tema en que estamos interesados/as serían necesarios dicionarios de sinónimos, antónimos, analizadores morfolóxicos etc.

As técnicas de *Text Retrieval* están en diferente nivel de desenvolvemento segundo a lingua de que se tratar. O inglés é, sen dúbida, a lingua que conta con máis ferramentas. O portugués e español teñen aínda un longo camiño por andar, mais xa contan con dicionarios electrónicos e ferramentas de lematización. No que atinxe ao galego está case todo por facer, aínda que hai equipos de investigación na Universidade de Vigo (http://webs.uvigo.es/sli/) e no Centro Ramón Piñeiro para a Investigación en Humanidades (http://www.cirp.es/) que están a traballar en diferentes frontes. Tamén, desde o ano 2002 investigadores en Filoloxía Galego-Portuguesa e Informática (http://rosalia.dc.fi.udc.es/lbd) da Universidade da Coruña estamos a traballar no desenvolvemento de ferramentas e recursos de recuperación de textos para o galego.

Neste traballo presentamos un lematizador para o galego. Unha primeira versión deste foi presentada xa no congreso SPIRE'02 (Brisaboa et alii 2002: 91-97), e nos últimos anos foi optimizado, completado e probado con corpora de diferente natureza para ser usado en servizos de busca por contido de bibliotecas dixitais. O que se pretende é que, a partir da palabra de busca introducida polo/a usuario/a, sexa a biblioteca dixital a que, usando a ferramenta de lematización, amplíe a busca automaticamente a todas as palabras da familia morfolóxica do termo procurado. Para iso, na biblioteca dixital realízase un proceso de lematización das palabras dos documentos, previo ás consultas dos usuarios/as, de maneira que, realmente, unha

procura consiste en buscar o lema da palabra introducida polo utilizador na listaxe de palabras lematizadas de cada documento. Finalmente, o que se lle amosa ao/á usuario/a é o texto orixinal dos documentos (en que aparece a palabra de procura nalgunha das súas variacións morfolóxicas).

O resto do artigo organízase como se expón a seguir: no apartado 2 exponse polo miúdo o que é a lematización, explicando os principais problemas e as aproximacións que existen en informática para levala a cabo de maneira automática e descríbese o algoritmo de lematización que temos desenvolvido para o galego. O apartado 3 fala das características especiais do galego que inflúen na configuración do algoritmo. O apartado 4 describe os resultados obtidos da aplicación do algoritmo de lematización do galego a un corpus formado por textos de diferentes áreas. Finalmente, como anexo, amósase unha boa parte das regras que forman o algoritmo de lematización para o galego, incluíndo as listaxes de excepcións de cada unha delas.

2. A lematización

A lematización consiste en representar mediante un único termo (lema) todas as posibilidades flexivas dunha palabra. Desde o punto de vista lingüístico, un lema é un termo que representa e unifica todos os elementos dun conxunto de palabras morfoloxicamente similares (Crystal 2000). Así, *camion-* sería o lema de *camioneiro*, *camiois*, *camións*, *camiós* etc; *garraf-,* o de *garrafón*, *garrafa*, *garrafiña*, etc.; ou *and-,* o de *andaría*, *andase*, *andar* etc.

O proceso de lematización pode realizarse mediante un algoritmo[1] que utilice regras gramaticais de derivación morfolóxica do idioma en cuestión, ou mediante un dicionario informatizado que asocie a cada posíbel variación morfolóxica o representante desta. O que os diferencia é a dificultade de elaboración entre un e outro, pois mentres que para a creación dun dicionario é preciso un grande esforzo de recopilación por ter que introducir cada palabra e o seu lema manualmente, a técnica baseada nun algoritmo permite realizar a lematización simplemente declarando unha serie de regras lingüísticas.

Os resultados de ambas as dúas técnicas son similares dentro dunha marxe de erros razoábel. No que se referir aos dicionarios informatizados, o principal problema co que nos encontramos é a ambigüedade semántica, que só podería ser evitada realizando unha análise semántica de cada entrada. Canto aos algoritmos, ademais desta ambigüedade, presentan dous problemas básicos: o *overstemming* e o *understemming*.

[1] Secuencia de instrucións que, aplicadas sobre un dato chamado "de entrada", transfórmano noutro dato calificado "de saída".

- *Overstemming*: o mal funcionamento da lematización provoca que palabras que na realidade deberían de se agrupar baixo diferentes lemas, por non se corresponder o contido semántico de cada unha delas, sexan agrupadas baixo unha mesma raíz. Por exemplo, se a "*macheta*" se lle elimina o sufixo apreciativo *-eta* obteriamos o lema "*mach-*" que coincidiría co lema de calquera forma flexionada do verbo "*machucar*" (*machuquei*, *machucaría*, etc) e coas palabras derivadas de "*macho*" (*machismo*, *machista*, etc). O *overstemming* provoca un mal funcionamento no sistema de busca por contido, xa que fai que recupere documentos non relevantes á procura.
- *Understemming*: execución errónea do algoritmo que resulta da obtención de diferentes formas canónicas para palabras que deberían agruparse baixo unha mesma raíz por teren o mesmo significado. Así, se a un termo como "*mazar*" se lle elimina a vogal temática "a" e o morfema substancial "r" obteriase o lema "*maz-*", que é diferente ao que ao que se facilitaría se o termo a lematizar é "*mace*" ou "*macei*", cando na realidade simplemente son distintas variantes flexivas do mesmo verbo. Isto fai que o sistema de busca non atope documentos que si están relacionados.

Un dos primeiros algoritmos de lematización foi desenvolvido para o inglés, no ano 1980, por Martin Porter. Polo nome do autor é coñecido como "Algoritmo de Porter" (Porter 1980), e pode ser definido como un autómata de estados finitos que inclúe un grupo de regras que se empregan para a eliminación de terminacións morfolóxicas e flexivas de palabras en inglés, sendo a súa idea básica a redución de plural a singular como forma de normalizar termos. Este algoritmo basicamente eliminaba o "s" final de cada palabra, mais como é evidente isto non é suficiente cando falamos de linguas como as románicas, onde as variacións flexivas e morfolóxicas son maiores.

A partir de entón, foron aparecendo adaptacións do Algoritmo de Porter para diferentes linguas (español, portugués, francés, etc.), axustándoo no posíbel ás regras do idioma de que se tratar.

No caso do español, introduciron pequenas modificacións para adaptar o Algoritmo de Porter á lingua hispana. Isto pode verse, por exemplo, en http://www.udlap.mx/~is112924/ [IS346/Tarea1.html. Porén, para o Portugués realizáronse adaptacións substanciais, tendo en conta todos os sufixos que operan nesa lingua, e acrecentando así mesmo unha lista de excepcións onde se inclúen aquelas palabras que non deben ser lematizadas.

Para o galego, partiuse do algoritmo deseñado por Viviane Moreira e Christian Huyck para o portugués (Moreira e Huyck 2001: 186-193), mais mellorándoo e reducindo ao máximo o *overstemming* e o *understemming*.

2.1. Algoritmo de lematización para o galego

O algoritmo de lematización para o galego está constituído por regras, tantas como sufixos existen na nosa lingua. Cando se propón un termo para lematizar, o algoritmo comproba que regra debe aplicarlle a dito termo tendo en conta os sufixos de que consta.

Para a construción destas regras guiámonos pola *Gramática da Lingua Galega* (vol. II, III) (Freixeiro 1999, 2000), e polo *Vocabulario Ortográfico da Lingua Galega* (VOLGA) (http://www.linux-galicia.org/diccionario/volga_revisado.zip). Na táboa 1 apreséntase a sintaxe das regras, así como un exemplo.

Táboa 1. Sintaxe xeral e exemplo das regras do algoritmo de lematización.

Sintaxe das regras	Exemplo de regra
"Sufixo que hai que mudar",	"eiro",
"Tamaño mínimo da raíz",	"3"
"Sufixo substituto",	""
"Listaxe de excepcións"	{canteiro, mareiro, peleiro}

Os compoñentes das regras son os seguintes:

"**Sufixo que hai que mudar**": terminación que se elimina ou, nalgún caso, se substitúe. É a primeira comprobación que lle hai que facer a un termo a lematizar, e así, no caso do exemplo, aplicaríase esta regra ás palabras acabadas en *-eiro*.

"**Tamaño da raíz**": número que se refire á lonxitude da raíz despois de terse eliminado o sufixo. A regra aplicarase só se a dimensión da raíz resultante é igual ou maior que esta medida. Deste xeito, a regra utilizaríase para a palabra *palleiro* por ser a raíz resultante de tres caracteres (*pall-*), mais nunca a *abeiro*, pois a lonxitude da base (*ab-*) despois da eliminación do sufixo sería menor á indicada na regra. Isto evita a eliminación de terminacións que na realidade non son sufixos, senón que forman parte do propio lema.

"**Sufixo substituto**": é o sufixo por que se substitúe o "sufixo que hai que mudar". Non todas as regras teñen este compoñente, así, no exemplo proposto, a terminación *-eiro* é simplemente eliminada, mais noutras regras como a de *-ós* realízase unha substitución por *-ón* (*xamós>xamón*).

"**Listaxe de excepcións**": relación de palabras para as que a regra non se debe aplicar. A regra do exemplo nunca se lle aplicaría a palabras como "*canteiro*", xa que forma parte da listaxe de excepcións. De non estar nesta listaxe, eliminariase o sufi-

xo e *canteiro* quedaría lematizado na raíz *cant-*, producíndose *overstemming* (coincidiría co lema do verbo "*cantar*").

Así, se aplicamos a regra do exemplo á palabra "*zapateiro*" obteriamos o lema *zapat-*, por ter a súa raíz máis de tres caracteres e non estar presente a palabra na listaxe de excepcións.

Parte das regras do algoritmo para o galego amósanse no anexo deste artigo, e a listaxe completa está dispoñíbel en http://bvg.udc.es/recursos_lingua/stemming.jsp.

As regras organízanse en etapas, dependendo do tipo de sufixos que tratar. Dentro de cada etapa, as regras vanse examinando secuencialmente, non podendo aplicarse máis de unha de cada vez. A orde das regras dentro de cada etapa fai que se comproben antes e, por tanto, se eliminen antes os sufixos máis longos, asegurando así que se aplique a regra máis adecuada á terminación da palabra. Por exemplo, nas regras do plural o sufixo *-ais* compróbase antes que o sufixo *-s* e nas de sufixos apreciativos compróbase o sufixo *-deiro* antes que *-eiro*. De non ser así, a palabra "*panadeiro*" lematizaría en "*panad-*" e non en "*pan-*" como debe ser (tras a eliminación da vogal final nunha etapa posterior).

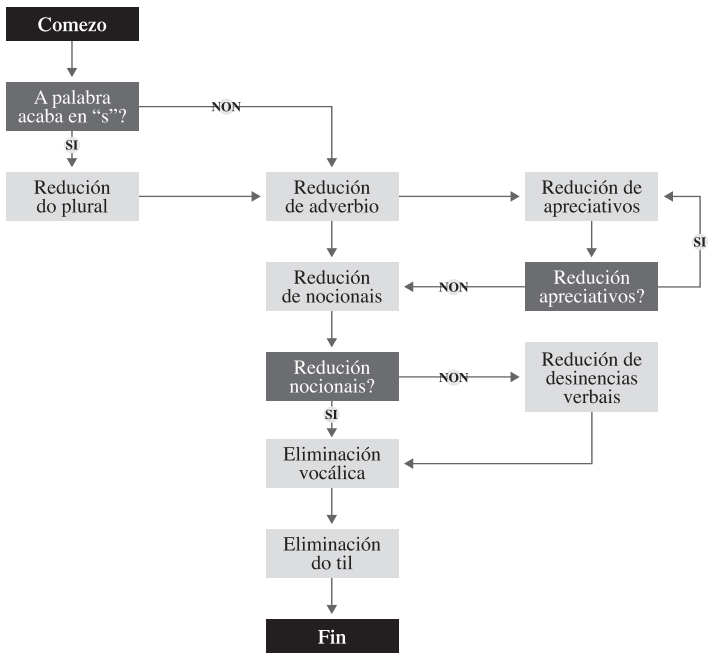

Figura 1. Diagrama de fluxo do algoritmo de lematización.

O algoritmo consta de sete etapas en total. A orde en que estas se executan non é aleatoria, así como tampouco as condicións que é preciso que se dean para que unha palabra pase por unha etapa ou outra, tal e como se define no diagrama de fluxo da Figura 1. A continuación describimos cada unha das sete etapas de que consta o algoritmo:

Etapa 1. Redución das desinencias de plural. Nesta etapa só se comproban as palabras acabadas en *-s*. O algoritmo comproba se o termo a lematizar acaba nalgunha das terminacións propostas na listaxe, e de ser así múdaa pola que está indicada, agás no caso do morfema *-s* que é eliminado sen substituírse por ningún outro (sempre e cando esa palabra non estea na listaxe de excepcións). Así, se o termo a lematizar é "*normais*", o sistema muda *-ais* por *-al*, e facilítanos a forma singular da palabra (*normal*), que logo será tratada a través de regras doutras etapas até chegar ao seu lema.

Etapa 2. Redución dos sufixos adverbiais. Esta etapa só contén unha regra, pois unicamente existe un sufixo capaz de formar adverbios, *-mente*. (*xentil>xentilmente*).

Etapa 3. Redución dos sufixos apreciativos (diminutivos, aumentativos, pexorativos e intensificadores). Nesta etapa elimínanse o conxunto de sufixos apreciativos, é dicir, aqueles que non teñen a capacidade de mudaren a categoría da palabra sobre a que actúan. Unha das características deste tipo de sufixos é a súa morfoloxía recursiva, que permite a acumulación de sufixos sobre a mesma base, como por exemplo, en "*gordochiño*". Neste caso o algoritmo na terceira etapa substitúe o sufixo *-iño* por *-o* (*gordochiño>gordocho*), a seguir, volta a revisar esta etapa e elimina a terminación *-ocho* (*gordocho>gord*), obtendo así o lema da palabra. Até o momento o algoritmo de lematización para o galego é o único capaz de recoñecer máis de un sufixo deste tipo nunha mesma palabra.

Etapa 4. Redución dos sufixos nocionais. Estes sufixos mostran unha grande tendencia á lexicalización, resultando con frecuencia dificultosa a distinción entre a base e o sufixo. Moitas palabras que conteñen un sufixo nocional non deben ser reducidas ao lema orixinal, xa que o sentido da palabra variou. Por este motivo, e para evitar o overstemming, as regras que se executan nesta etapa soen ter moitas excepcións. Por exemplo, no vocábulo "*lanzal*" a diferenza entre a base (*lanza*) e o sufixo nocional (*-al*) está esquecida polos/as falantes, e cando unha persoa procura aqueles textos que conteñan o adxectivo "*lanzal*" (*esvelto*) probabelmente non estea interesada naqueles que falen sobre lanzas, e é por iso que esta palabra está incluída na listaxe de excepcións da regra do sufixo *-al*.

Etapa 5. Redución de terminacións verbais. Neste paso obterase a raíz do verbo, após a eliminación da vogal temática e os morfemas de tempo e número. Así, por exemplo, a P6 do copretérito do verbo cantar, "*cantaban*", redúcese a "*cant-*".

Etapa 6. Redución vocálica. Neste caso redúcense as vogais que, tras pasar por todas as fases anteriores, aínda se manteñen. É o caso, por exemplo, de "*movedizo*", que despois da extracción do sufixo nocional *-dizo* fica como resultado "*move*", sendo na realidade a raíz deste "*mov-*". Outra solución sería considerar *-edizo* como sufixo, mais isto ampliaría do punto de vista cuantitativo o número de terminacións en cada etapa. Dentro deste grupo inclúense tamén aquelas palabras que varían a súa raíz segundo foren coas vogais "a,o" ou "e,i". Estamos a nos referir aos dígrafos "gu" e "qu" que deben ser substituídos por "g" e "c" (*cheguemos>chegu->cheg-*; *marquei> marqu->marc-*). Introducimos estes aquí por unha razón de rendibilidade, pois, por seren estes casos minoritarios, non sería produtivo crear unha etapa en exclusiva para eles.

Etapa 7. Eliminación do til. Este último paso é necesario porque as raíces das palabras galegas poden ir acentuadas ou non dependendo do sufixo que tiveren. Por exemplo, palabras como "*práctica*" e "*practicamente*", que teñen a mesma base, darían resultados diferentes, nun caso obteriamos "*práct-*" e noutro "*pract-*". Despois de pasar por esta fase ambas as dúas palabras se reducen ao lema "*pract-*".

3. Particularidades do galego

O galego presenta unha serie de características que fan que o algoritmo deseñado sexa especialmente complexo. Unha das máis importantes é o feito do galego ter un número inusualmente elevado de sufixos, e isto fai que o número de regras do lematizador aumente (de forma proporcional ao número de sufixos).

Os motivos para a aparición destas particularidades son principalmente históricos. A marxinación sufrida polo noso idioma, durante séculos transmitido principalmente de forma oral, propiciou a aparición de moitas variantes dialectais para unha mesma palabra, así como a inclusión de castelanismos. Por exemplo, para as palabras acabadas en *-ón* temos en galego tres terminacións diferentes para a formación do plural: *-óns* (na parte occidental da Galiza), *-ós* (na parte central) e *-ois* (na parte oriental), e mesmo, solucións coincidentes co portugués como *-ões*. Tamén contamos coa presenza en textos galegos de castelanismos, non só léxicos, senón tamén morfolóxicos, como palabras acabadas en *-l* que fan o plural en *-les* (*animales*, *normales*). Tanto as variantes dialectais como os castelanismos non poden ser esquecidos na construción dunha ferramenta que vai traballar sobre textos galegos de todas as épocas, debido á súa frecuencia de aparición.

Ademais, durante moito tempo non existiu unha institución pública nin autoridade individual que, co seu prestixio, lograse impor unha normativa. Por este motivo, os escritores e escritoras do século XIX mantiveron un comportamento libre de ataduras normativas, resolvendo os problemas que se manifestaban con criterios pro-

pios. Aínda hoxe, cunha normativa xa revisada (12 de xullo de 2003), continúa a haber discrepancias ortográficas. Por esta razón, se só temos en conta a normativa oficial (por exemplo, o plural *camións* no exemplo anterior), moitas palabras dos textos ficarían sen lematizar, ou o resultado sería unha raíz errónea.

A consecuencia desta situación encontrámonos que debemos manter máis dunha variante morfolóxica da mesma palabra, co que, como xa se indicou, o número de sufixos (e por tanto de regras) utilizados para a elaboración do algoritmo de lematización para o galego é significativamente superior a outros idiomas romances, como poden ser o portugués ou o español. Aínda que isto dificulta a elaboración do lematizador, cremos que é necesario, pois na construción de calquera corpus de textos galegos (especialmente os literarios) atoparemos estas variacións.

Mais non sempre é posíbel manter todas as variantes morfolóxicas dunha palabra. Por exemplo, como singular da palabra "bons" (e por tanto como lema) debemos seleccionar entre "*bon*" ou "*bom*", mais non ambos, xa que o algoritmo produce un único lema para unha palabra dada. Para realizarmos este tipo de seleccións fixemos unha serie de estudos estatísticos que nos dan a frecuencia de aparición das palabras nos textos galegos. Para este traballo empregamos principalmente dous corpus: o Tesouro Informatizado da Lingua Galega (TILGA) (http://www.ti.usc.es/TILGA/) e a Biblioteca Virtual Galega (http://www.bvg.udc.es). No exemplo anterior escollemos a palabra "*bon*", baseándomonos nos resultados que nos mostran os corpus: "*bon*" aparece en 3616 ocasións no TILGA fronte ás 6 en que se encontra "*bom*". Resultados similares amósanse tamén na BVG.

Ao realizar os estudos estatísticos sobre estes dous corpora, encontrámonos con que estes tamén foron de utilidade para descubrir palabras ou variacións que, nun principio, non foron consideradas e que si deberían selo. Comprobamos, por exemplo, que debíamos incluír o plural en *-ás* para as palabras acabadas en *-al*, pois aínda que non é unha terminación moi común actualmente, si tivo un uso frecuente, especialmente no século XIX, e en concreto a forma "*reás*", como demostra a súa aparición en 1197 documentos do TILGA e en 21 obras da BVG.

4. Resultados empíricos

Para comprobarmos a efectividade deste algoritmo aplicámolo a un corpus monolingüe en galego composto por documentos de diferentes xéneros (literarios, xornalísticos e xurídicos) tirados da BVG (http://bvg.udc.es), do xornal *A Nosa Terra* (http://www.anosaterra.com) e do *Diario Oficial de Galicia* (http://www.xunta.es/dog/dog.nsf). O corpus resultante ten un tamaño de 42'1 MB, sendo maioritarios os documentos de tipo literario (26'8 MB).

Táboa 2. Vocabulario dos textos en galego antes e despois da lematización.

Arquivos	Tamaño (MB)	Palabras diferentes	Lemas obtidos	Porcentaxe
Literatura	26'80	231.291	93.295	40'33%
Xornalismo	7'47	56.452	24.723	43'79%
Xurídico	7'83	68.510	43.882	64'05%
TOTAL	42'10	356.253	161.900	45'44%

O feito de non existiren para a nosa lingua outras ferramentas semellantes ao algoritmo de lematización, imposibilita as comparacións co fin de comprobarmos a súa efectividade, e por iso se confrontan os resultados expostos na táboa 2 cos que se obtiveron con lematizadores para linguas como o español ou o portugués con corpus similares. Tras a execución destes algoritmos os resultados son os que se amosan nas táboas 3 e 4 (Brisaboa 2002: 13-24).

Táboa 3. Vocabulario dos textos en español antes e despois da lematización.

Arquivos	Tamaño (MB)	Palabras diferentes	Lemas obtidos	Porcentaxe
Literatura	88'0	305.309	129.437	42'40%
Xornalismo	7'6	61.966	26.520	41'80%
Xurídico	9'6	49.312	19.965	40'50%
TOTAL	105'0	416.587	175.922	42'22%

Táboa 4. Vocabulario dos textos en portugués antes e despois da lematización.

Arquivos	Tamaño (MB)	Palabras diferentes	Lemas obtidos	Porcentaxe
Literatura	15'0	116.838	40.495	34'65%
Xornalismo	35'0	136.573	56.263	41'20%
Xurídico	1'4	10.765	5.590	51'90%
TOTAL	51'4	264.176	102.348	38'74%

O primeiro que debemos destacar é o feito de que, proporcionalmente, o número de entradas antes de se realizar o proceso de lematización é significativamente superior no corpus de galego. Isto non sorprende se se consideran as posíbeis variacións léxico-morfolóxicas que unha palabra pode ter na nosa lingua.

Unha vez realizado o proceso de lematización, o tamaño dos vocabularios foi reducido significativamente nos tres casos. Porén, os algoritmos para o galego e o portugués compórtanse de maneira distinta dependendo da natureza dos textos a lematizar, tendo tendencias similares en canto ao número de lemas obtidos. Non obstante os resultados do algoritmo para o español son independentes da natureza dos corpus. Isto é debido a que a complexidade dos algoritmos para o galego e o portugués é maior que a do algoritmo para o español (hai que lembrar que a lematización consiste en eliminar sen máis certas terminacións das palabras).

Tanto no portugués como no galego, os textos xurídicos son os que menos se lematizan, sendo o galego o que deu as porcentaxes máis altas (64'05%). Isto débese a que os documentos dos que se partiu están tirados do DOG, onde aparece unha grande cantidade de datas, abreviaturas e nomes propios que esta ferramenta non lematiza. Como se pode observar o que máis lematiza neste tipo de textos é o español (40'50%) o que demostra que se produce *overstemming*, pois neste tipo de documentos o número de palabras lematizadas ten que ser menor. No caso do portugués, ao igual que no galego, mais en menor medida, prodúcese menos lematización en relación aos textos dos outros xéneros.

Na táboa 5 amósase una proba feita co lematizador para o galego utilizando un fragmento dun texto do coruñés Ramón Armada Teixeira, tirado da súa obra teatral *Non mais emigración,* do ano 1886.

Táboa 5. Exemplo dun texto galego lematizado.

Texto orixinal	**Texto lematizado**
Pedide cabritiños	ped cabr
Á Virxen d'o Cristál,	A virx d'o cristal
Qu'o meu amor non fuxa,	Qu'o meu am non fux
N-a vida, d'o lugár.	N-a vid d'o lugar

Unha vez realizada a lematización sobre este fragmento pódese observar que o sistema inspeccionou correctamente os sufixos, procedendo á súa eliminación (*cabritiños>cabrito>cabr-*) así como as desinencias verbais (*pedide>ped*, *fuxa>fux*). As

palabras que pertencen a categorías pechadas como as preposicións (*d'o, N-a),* os pronomes (*meu*), etc, non son lematizadas.

Así unha vez lematizado o texto, cando unha persoa buscase por exemplo aqueles documentos que falen sobre "*cabras*" obtería como resultado da busca o documento en que se inclúe o texto proposto (texto orixinal) sen ter que explicitar a variante morfolóxica que realmente aparece (*cabritiños*) nin outras que pode ter a palabra "*cabra*".

5. Conclusións e traballo futuro

Un requisito imprescindíbel para a implementación dun servizo de busca por contido nunha biblioteca dixital é que o idioma en que estean escritos os documentos que armacena conte cos recursos e ferramentas de recuperación de textos axeitados (corpora, lematizadores, analizadores morfolóxicos etc.).

Neste artigo presentamos un algoritmo de lematización para o galego, ferramenta imprescindíbel no desenvolvemento de servizos de busca por contido nas bibliotecas dixitais.

O algoritmo desenvolvido está baseado en regras e listaxes de excepcións, e presenta unha mellora significativa con relación ao prototipo creado no 2002, pois, agora é capaz de tratar tanto a recursividade morfolóxica, que supuña un problema por estaren os nosos textos repletos de construcións deste tipo, así como as diferentes formas flexivas dunha mesma palabra que se adscriben a unha ou outra área lingüística da Galiza. Tamén é de destacar os bos resultados obtidos en relación aos lematizadores existentes para outras linguas da Península, concretamente español e portugués, cando se executan sobre textos de características semellantes. Isto é debido á maior precisión con que o algoritmo para o galego detecta e elimina sufixos para a obtención do lema.

Finalmente, cómpre facer referencia á grande riqueza léxica do galego. Aínda que para evitar o *overstemming* se crean as listaxes de excepcións, debido a esta abundancia lexical é imposíbel sinalar todos os casos, polo que non é posíbel eliminar no cen por cen este tipo de erros. Mais, por estar orientado este lematizador á recuperación da información, e non se tratar dun puro exercicio de lingüística, o que se pretende é que dentro dunha marxe razoábel de acertos, palabras semellantes sexan reducidas a lemas idénticos. Por iso consideramos, como se amosa nas probas empíricas, que dentro desa marxe razoábel o lematizador aquí presentado funciona de forma correcta.

Referencias bibliográficas

Brisaboa, N. R. / Fariña, A. / Navarro, G. / Iglesias, E. L / Paramá, J. R. / Esteller, M. F. (2002): "Compresión de textos en Lenguas Romances", en Brisaboa, N. R. (ed.): *Ingeniería del Software*: 169-180 (Colombia: AECI).

Brisaboa, N. R. / Fernández, C. (2001): "Introdución ás Bibliotecas Dixitais", *Revista Galega de Filoloxía,* 2: 27-51 (A Coruña: Baía Edicións).

Brisaboa, N. R. / Callón, C. / López, J. R. / Places, A. S. / Sanmartín, G. (2002): "Stemming Galician Texts", en Laender, A. / Oliveira, A.: *Proceedings of the 9th International Symposium, String Precessing and Information Retrieval (SPIRE'02)* (*Lisboa, 11/13-9-2002*): 91-97 (Berlín: Springer-Verlag).

Crystal, D. (2000): *Diccionario de Lingüística y Fonética* (Barcelona: Octaedro).

Fernández, C. / Places, A. S. (2004): *As bibliotecas dixitais* (Santiago de Compostela: Laiovento).

Freixeiro, X. R. (1999): *Gramática da Lingua Galega. III. Semántica* (Vigo: A Nosa Terra).

Freixeiro, X. R. (2000): *Gramática da Lingua Galega. II. Morfosintaxe* (Vigo: A Nosa Terra).

Moreira, V. / Huyck, C. (2001): "A Stemming Algorithm for the Portuguese Language", en Navarro, G.: *Proceedings of the 8th International Symposium on String Processing and Information Retrieval (SPIRE'01)* (*Chile, 13/15-11-2001*): 186-193 (USA: IEEE Computer Society).

Porter, M. (1980): www.tartarus.org/~martin/PorterStemmer

Anexo: Regras do algoritmo de lematización para o galego

ETAPA 1. REDUCIÓN DAS DESINENCIAS DE PLURAL

Sufixo	Tam.	Subst.	Exemplo	Excepcións
ns	1	n	bons→bon	luns, furatapóns, furatapons
ós	3	ón	xamós→xamón	
ões	3	ón	balões→balón	
ães	1	ão	capitães→capitão	mães, Magalhães
ais	2	al	normal→normais	cais, tais, mais, pais, ademais, namais, lapis
áis	2	al	Amais	cáis, táis, máis, páis, ademáis, namáis
éis	2	el	papéis→papel	
eis	2	el	posíbeis→posíbel	

Sufixo	Tam.	Subst.	Exemplo	Excepcións
óis	2	ol	espanhóis→espanhol	escornabóis
ois	2	ol	caracois→caracol	escornabois
ís	2	il	cadrís→cadril	país
is	2	il	cadris→cadril	menfis, pais, Kinguís,
les	2	l	males→mal	ingles, marselles, montreales, senegales, manizales,Móstoles, Nápoles
res	3	r	mares→mar	petres, Henares, Cáceres, Baleares, Linares, Londres, Mieres, Miraflores, Mércores, venres
ces	2	z	luces→luz	frances, escoces
zes	2	z	luzes→luz	
ises	3	i	leises→lei	
ás	1	al	animás→animal	más
ses	2	s	gases→gas	
s	2		casas→casa	barbadés, Barcelonés, Cantonés, gabonés, Llanés, medinés, pés, escocés, escocês, francês, Barcelonês,Cantonês , revés, macramés, Reves, barcelones, cantones, gabones, Llanes, Magallanes, medines, escoces, frances, xoves, martes, aliás, pires, lápis, cais, mais, mas, menos, férias, pêsames, crúcis, país, Cangas, Atenas, Asturias, Canarias, Filipinas, Honduras, Molucas, Caldas, mascareñas, Micenas, Covarrubias, psoas

ETAPA 2. REDUCIÓN DOS SUFIXOS ADVERBIAIS

Sufixo	Tam.	Subst.	Exemplo	Excepcións
mente	4		felizmente→feliz	experimente, vehemente

ETAPA 3. REDUCIÓN DOS SUFIXOS APRECIATIVOS (46 regras en total)

Sufixo	Tam.	Subst.	Exemplo	Excepcións
mente	4		felizmente→feliz	experimente, vehemente
díssimo	5		cansadíssimo→cansad	
díssima	5		cansadíssima→cansad	
dísimo	5		cansadísimo→cansad	
dísima	5		cansadísima→cansad	
abilíssimo	5		amabilíssimo→ama	
abilíssima	5		amabilíssima→ama	
abilísimo	5		amabilísimo→ama	
abilísima	5		amabilísima→ama	
íssimo	3		fortíssimo→fort	
íssima	3		fortíssima→fort	
ísimo	3		fortísimo→fort	
ísima	3		fortísima→fort	
ésimo	3		centésimo→cent	
ésima	3		centésima→cent	
érrimo	4		paupérrimo→paup	
érrima	4		paupérrima→paup	
ana	2		charlatana→charlat	argana, badana, , banana, boleeana, brahmana, cabana, canana, caravana, catana, choupana, espadana, faciana, filigrana, galbana, hossana, iguana, lantana, lesbiana, macana, maiorana, mariguana, marihuana, mediacana, membrana, mesana, nirvana, obsidiana, palangana, pavana, persiana, pestana, porcelana, pseudomembrana, roldana, sábana, salangana, saragana, ventana
ão	3		garrafão→garraf	abalão, acordeão, aldrabão, alerão, alinhão, ambão, bombão, calção, campão, canalão, cantão, canhão, centão, ciclão, colhão, colofão, copão, coração, cotão, cupão, petão, tirão, tourão, turão, versão, zubão, zurrão

Sufixo	Tam.	Subst.	Exemplo	Excepcións
om	3		garrafom→garraf	abalom, acordeom, alciom, aldrabom, alerom, aliñom, ambom, bombom, calzom, campom, canalom, cantom, cañom, centom, ciclom, collom, colofom, copom, corazom, cotom, cupom, petom, tirom, tourom, turom, unciom, versiom, zubom, zurrom
án	3		charlatán→charlat	ademán, agremán, alcavarán, alcorán, astracán, bambán, bardán , barragán, barregán, capitán, cordobán, corricán, cotián, curricán, faisán, furacán, fustán, gabán, gabián, galán, ganapán, gañán, harmatán, iatagán, lavacán, mazán, mazapán, mourán, pasamán, rabadán, refrán, sasafrán, serán, serrán, solimán, tabán, temperán, temporán, titán, tobogán, verán, volcán, volován
…				

ETAPA 4. REDUCIÓN DOS SUFIXOS NOCIONAIS (61 regras en total)

Sufixo	Tam.	Subst.	Exemplo	Excepcións
idade	3		vistosidade→vistos	acridade, autoridade, amenidade, calidade, comunidade, escuridade
edade	3		solidariedade→solidari	
dade	3		lealdade→leal	bondade
adeiro	3		cantadeiro→cant	
adeira	3		cantadeira→cant	
edeiro	3		tendedeiro→tend	
edeira	3		tendedeira→tend	bandeira
…				

ETAPA 5. REDUCIÓN DE TERMINACIÓNS VERBAIS (169 regras en total)

Sufixo	Tam.	Subst.	Exemplo	Excepcións
aba	2		amab→am	
abade	2		andabade→and	
ábade	2		andábade→and	
abamo	2		chorabamo→chor	
ábamo	2		chorábamo→chor	
aban	2		moraban→morab	
ache	2		andache→and	
ade	2		andade→and	
ai	2		cantai→cant	
am	2		cantam→cant	
amo	2		cantamo→cant	
an	2		varran→and	
ando	2		cantando→cant	
ar	2		cantar→cant	azar, bazaar, patamar
ara	2		cantara→cant	arara, prepara
ará	2		cantará→cant	alvará, bacará, bacarrá
arade	2		lembrarade→lembra	
árade	2		convidárade→convid	
aram	2		cantaram→cant	
arám	2		enviarám→envi	
aramo	2		bañaramo→bañar	
áramo	2		cantáramo→cant	
arán	2		enviarán→envi	
…				

ETAPA 6. REDUCIÓN VOCÁLICA

Sufixo	Tam.	Subst.	Exemplo	Excepcións
gue	2	g	segue→seg	
que	2	c	marque→marc	
a	3			
e	3			
o	3			
â	3			
ã	2			amanhã, arapuã, fã, divã, manhã
ê	3			
ô	3			
á	3			
é	3			
ó	3			