Una aplicación de RI basada en PLN: el proyecto ERIAL*

Fco. Mario Barcala y **Eva M.ª Domínguez**
Centro Ramón Piñeiro para a Investigación en Humanidades

Miguel A. Alonso y **David Cabrero** y **Jorge Graña** y **Jesús Vilares** y **Manuel Vilares**
Universidade da Coruña

Guillermo Rojo y **M.ª Paula Santalla** y **Susana Sotelo**
Universidade de Santiago de Compostela

Resumen En este artículo se describe el sistema ERIAL, llevado a cabo en el marco del proyecto del mismo nombre, para Recuperación de Información. Tras una primera descripción externa del proyecto (Sección 1), se presenta el entorno modular LEIRA desarrollado con tal fin (Sección 2) y, a continuación (Sección 3), se describen en detalle tanto los recursos lingüísticos integrados en los módulos computacionalmente diferenciados en el entorno en cuestión como los módulos en sí mismos. En la Sección 4, en primer lugar, se describen los corpus elaborados para evaluar el sistema y, en segundo lugar, se valoran los resultados obtenidos, para interrogaciones en gallego o castellano, mediante la activación de diferentes módulos del mismo. La Sección 5, finalmente, explica cómo pretendemos mejorar el sistema en el futuro.

1 Introducción

El proyecto ERIAL, Extracción y Recuperación de Información mediante Análisis Lingüístico, tiene como objetivo desarrollar un entorno para recuperación de información basado en la utilización de técnicas de PLN, un entorno adaptado, además, a las necesidades del grupo Editorial Compostela, que publica el diario *El Correo Gallego/O Correo Galego*, un entorno, por lo tanto, que ha de ser funcional en dos lenguas: castellano y gallego. En el proyecto, por su carácter obviamente interdisciplinar, han participado la Universidad de A Coruña, que aportaba el conocimiento computacional, las Universidades de Santiago de Compostela y Vigo, y el Centro Ramón Piñeiro para la Investigación en Humanidades, que aportaban conocimiento lingüístico respectivamente sobre español y gallego, la empresa Compaq, que proporcionaba equipamiento informático y asesoramiento técnico, y Editorial Compostela, como usuario del sistema. El proyecto ha sido financiado, entre los años 1999 y 2001, por la Secretaría de Estado de Política Científica y Tecnológica (Fondos Europeos para el Desarrollo Regional, FEDER).

2 El sistema ERIAL para RI: el entorno LEIRA y sus módulos

Los módulos de LEIRA se muestran en la figura 1. En primer lugar, un *tokenizador* avanzado trata algunos aspectos lingüísticos complejos al tiempo que realiza una preetiquetación del texto. A continuación, un *etiquetador*, basado en modelos de Markov ocultos, se encarga de realizar la desambiguación de las etiquetas y de obtener los lemas correspondientes a cada palabra. El *generador de familias morfológicas* se utiliza para obtener los conjuntos de palabras relacionadas mediante procesos propios de la morfología derivativa. Una vez etiquetado un texto, se procede a la extracción de los pares de dependencia sintáctica presentes en las frases nominales y en sus variantes sintácticas y morfosintácticas, los cuales son normalizados mediante las familias morfológicas.

La figura 2 muestra cómo el lematizador y las familias morfológicas permiten obtener los términos simples presentes en los documentos, mientras el analizador sintáctico permite indexar los términos multipalabra.

El comportamiento del sistema es similar tanto en el proceso de indexación como en el de consulta, ya que se aplican prácticamente

* La investigación descrita en este artículo ha sido financiada en parte por el Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (TIC2000-0370-C02-01), el Ministerio de Ciencia y Tecnología (HP2001-0044) y la Xunta de Galicia (PGIDT01PXI10506PN).

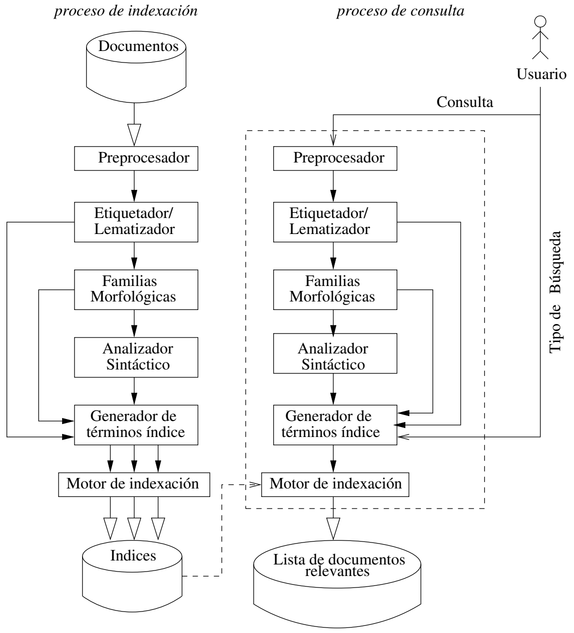

Figura 1: Arquitectura general del sistema

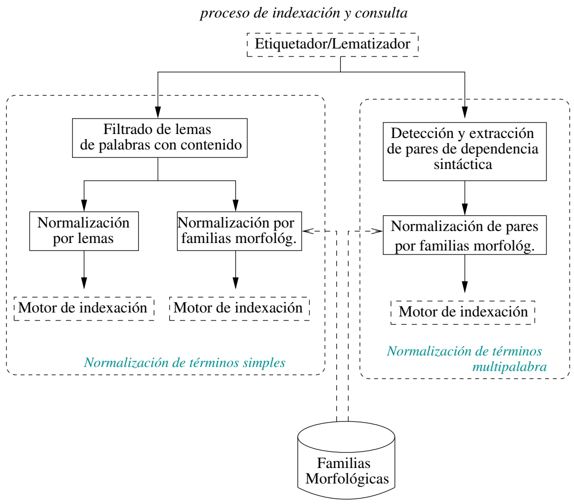

Figura 2: Generación de términos índice

las mismas operaciones tanto a los documentos como a las consultas.

3 La extracción de términos

3.1 Recursos: lexicones y corpus de desambiguación

Uno de los problemas principales al que se tienen que enfrentar los sistemas de procesamiento del lenguaje natural es la resolución de la ambigüedad en todos los niveles. Por ello, en ERIAL hemos producido a partir de lexicones y *tagsets* de orientación general, recursos lingüísticos derivados específicamente adaptados a la aplicación en cuestión, recursos en los que se intenta reducir al máximo el nivel de ambigüedad.

	Español		Gallego	
	lemas	formas	lemas	formas
sustantivos	60.264	131.380	11.706	23.346
adjetivos	22.450	110.236	3.700	5.169
verbos	11.914	363.699	6.789	544.690
adverbios	4.079		3.734	
conjunciones	169		40	
preposiciones	158		43	
determinantes	525		347	
pronombres	617		476	

Tabla 1: Datos acerca de los lexicones empleados.

3.1.1 En español

El lexicón de la aplicación correspondiente a la parte de español contiene 94.628 lemas con información morfológica, así como datos de subcategorización para 1.284 verbos[1], aunque esto último no es utilizado en ninguno de los módulos del sistema. Esta lista de lemas se genera a partir de un lexicón de propósito general, construido de acuerdo con criterios lexicográficos a partir de diccionarios electrónicos y corpus cuya exhaustividad implica un alto índice de ambigüedad evitable (tal como se describe en [1]). Mediante la aplicación de filtros (basados principalmente en datos estadísticos y de uso real) obtenemos lexicones adaptados a la aplicación en los que se resuelven a nivel léxico determinados problemas de ambigüedad que de otro modo hubiesen sido más difíciles de eliminar en niveles superiores de análisis. Por otra parte, cada lema asociado a su correspondiente modelo flexivo es procesado por un conjunto de gramáticas formales morfológicas [1] y utilidades de conversión para generar un lexicón de formas flexivas (véase Tabla 1) en el que además se incluyen todas las posibles combinaciones de verbos con pronombres enclíticos.

Para el entrenamiento del etiquetador (véase Sección 3.2.2) se construyó un corpus desambiguado manualmente en el que fueron incluidos 27 documentos procedentes del corpus de la aplicación, con aproximadamente 16.500 palabras. En la etiquetación de este corpus se empleó el conjunto completo de 918 etiquetas en el que se basa el lexicón, a partir del cual es posible derivar automáticamente

[1]La información sintáctica contenida en el lexicón procede de la *Base de Datos Sintácticos del Español* (BDS), desarrollada por el Grupo de Sintaxis del Español de la USC y disponible en **http://www.bds.usc.es**.

subconjuntos menores más ajustados a los requisitos de una aplicación concreta. En el caso del proyecto ERIAL, junto al conjunto máximo de 918 etiquetas se generaron subconjuntos reducidos de 324 y 218, que logran mejorar los resultados finales de la fase de etiquetación. Estos subconjuntos aplican reducciones que afectan principalmente a las etiquetas de los verbos y, de ellos, fue seleccionado el que contiene 218 etiquetas como *tagset* de trabajo para el módulo del etiquetador (de esas 218 etiquetas, sólo 186 están presentes en el corpus de entrenamiento).

3.1.2 En gallego

Para la elaboración del lexicón se tomó como punto de partida el *Corpus de Referencia do Galego Actual* (CORGA) y se introdujeron los lemas correspondientes a las 70.000 formas más frecuentes de este corpus. Dado que el tamaño del lexicón no resultaba así satisfactorio, fue incrementado a partir del *Vocabulario Ortográfico da Lingua Galega* (VOLG). Para organizar y caracterizar las entradas del diccionario se determinó un sistema de etiquetas (numéricamente no cerrado) que recoge, preferentemente, información morfológica (identificación de categorías de palabras y de categorías morfológicas a ellas asociadas). Así, se decidió que en gallego eran funcionales los adjetivos, adverbios, artículos, conjunciones, demostrativos, exclamativos-interrogativos, indefinidos, interjecciones, posesivos, pronombres personales, posesivos, relativos, signos de puntuación, sustantivos y verbos, siendo pertinentes para una completa clasificación los atributos tipo, género, número, grado, persona, caso, tiempo verbal, modo y número del posesor. Se incluyó además un atributo denominado valor que proporciona información de carácter sintáctico sobre la posición y/o función de determinados elementos. En la construcción del diccionario resultó ser especialmente dificultosa la integración en el sistema de las formas verbales con pronombres enclíticos: modificaciones en la raíz, acentuación divergente, numerosas combinaciones, restricciones en la combinatoria de pronombres y formas verbales, etc. El etiquetario del gallego se elaboró de manera que se pudiese pasar de un sistema amplio a otro reducido fácilmente y, si bien se trabajó teóricamente con 664 etiquetas, por motivos funcionales, en la práctica, el tagset lo constituyeron 284 etiquetas.

En cuanto a la elaboración del corpus desambiguado, se seleccionó un subcorpus de 10 periódicos en versión electrónica que integraban las diferentes noticias de 10 días de *O Correo Galego* del mes de mayo de 1999. Se prescindió de las noticias de la sección "Opinión" por considerar que la lengua utilizada en ellas difería del registro y estilo propiamente periodísticos. Un equipo de cinco personas llevó a cabo la tarea de desambiguar manualmente este subcorpus, constituido por 97.500 palabras, dedicándole una media de 3 a 4 horas durante 20 días. Fue preciso elaborar un manual que facilitase el tratamiento homogéneo de los textos y dictase unas normas de trabajo. Si bien la desambiguación se realizó principalmente desde una perspectiva morfológica —no se consideró la recategorización, por ejemplo— se atendió a la sintaxis para determinar género y número en sustantivos y adjetivos, y persona en las formas verbales.

A los problemas que pueden surgir en una tarea como ésta (dificultades en la caracterización de nombres de organizaciones, de las partículas *como, cuando* e incluso determinar la pertenencia a una categoría u otra de ciertos elementos, etc.) es necesario sumar la falta de fijación de la lengua gallega, la abundancia de variantes dialectales y las numerosas interferencias del español que provocaron que la etiquetación manual de los textos se hiciese sumamente compleja.

3.2 Herramientas

A continuación describiremos de manera resumida las principales características de los módulos que componen el sistema LEIRA.

3.2.1 Preprocesador

Muchas aplicaciones de procesamiento del lenguaje natural presuponen que los textos ya están correctamente segmentados en *tokens*, unidades de información de alto nivel que identifican los componentes individuales de los textos. Esta creencia es errónea puesto que el concepto lingüístico de palabra no siempre coincide con su realización gráfica. Por ello hemos desarrollado un segmentador avanzado [5] que realiza tareas de preetiquetación de los textos sobre secuencias de caracteres no ambiguas tales como *números* o *fechas*, al tiempo que trata fenómenos como *abreviaturas*, *acrónimos*, *contracciones*, *pronombres enclíticos*, *identificación de expresiones* —por lo cual entendemos la unión de todas las palabras que componen locuciones, expresiones fijas o frases hechas, llevada a cabo mediante consulta a un diccionario de expresiones—, e *identificación de nombres propios* —mediante reglas para el reconocimiento de nombres compuestos tales como *Ministerio de Educación, Cultura y Deportes.*

Se distinguen en la estructura general del sistema dos modos de trabajo del preprocesador. En el momento de la indexación se realiza un procesamiento *off-line* que a su vez consta de dos fases: una primera en la cual se realiza el entrenamiento de los nombres propios, y una segunda en la que los datos obtenidos son utilizados para realizar la preetiquetación de los documentos. En el momento de realizar una consulta, se activa el modo de procesamiento *on-line* con el fin de preprocesar la consulta.

3.2.2 Etiquetador

La etiquetación de las partes del discurso se realiza mediante un modelo oculto de Markov (HMM) de segundo orden. Los elementos del modelo y los procedimientos para estimar sus parámetros se basan en el trabajo de Brants [2], aunque se han realizado algunas extensiones con el fin de tratar segmentaciones ambiguas e integrar diccionarios externos [6], implementados mediante autómatas finitos acíclicos mínimos numerados por razones de eficiencia [4].

Características del modelo. Los estados del HMM representan pares de etiquetas, mientras que las salidas representan las palabras. La probabilidad del mejor camino se calcula mediante el algoritmo de Viterbi [15], utilizando las probabilidades de transición entre trigramas. Tanto las probabilidades de transición como las de salida se estiman a partir de un corpus etiquetado. Como primer paso, utilizamos las probabilidades de máxima verosimilitud derivadas de las frecuencias relativas. Como segundo paso, las frecuencias contextuales son suavizadas, al tiempo que se completan las frecuencias léxicas mediante el manejo de las palabras que no aparecen en el corpus de entrenamiento pero que están presentes en diccionarios externos.

Hemos comprobado que el paradigma de suavizado que proporciona mejores resultados es el de la interpolación de unigramas, bigramas y trigramas.

Manejo de palabras desconocidas. Dada una palabra desconocida, estableceremos su conjunto de etiquetas candidatas, así como sus probabilidades asociadas, mediante el estudio de los sufijos de la palabra. La distribución de probabilidad de un determinado sufijo se genera a partir de todas las palabras del corpus de entrenamiento que comparten el mismo sufijo, considerando una longitud máxima predefinida. Las probabilidades son entonces suavizadas mediante abstracción sucesiva [11].

Integración de diccionarios externos. Nuestro modelo de integración de diccionarios se basa en las fórmulas de Good-Turing [9], en las cuales cada par palabra-etiqueta presente únicamente en el diccionario externo es considerado como un evento con frecuencia de aparición nula en el corpus de entrenamiento. Estas fórmulas asignan probabilidades mayores que 0 a estos eventos, raros, pero que se dan en la práctica. Como ventaja adicional, esta técnica produce una menor distorsión en el modelo e incrementa el rendimiento en la etiquetación, especialmente con corpus de entrenamiento reducidos [6].

Manejo de segmentaciones ambiguas. Debido a las segmentaciones ambiguas detectadas por el preprocesador, el etiquetador debe ser capaz de tratar con secuencias de *tokens* de diferente longitud, lo cual implica no sólo decidir la etiqueta que les debe ser asignada, sino también determinar si algunos de ellos forman conjuntamente un sólo término. Para realizar este proceso, consideraremos la evaluación individual de cada flujo de *tokens* y su comparación posterior con el fin de seleccionar la secuencia más probable. Para ello es preciso definir un criterio objetivo de comparación. Cuando el paradigma de etiquetación utilizado es el de los modelos de Markov ocultos, como en nuestro caso, el criterio puede consistir en la comparación normalizada de las probabilidades logarítmicas acumuladas.

3.2.3 Generador de familias morfológicas

Una *familia morfológica* es el conjunto de palabras obtenidas a partir de una misma raíz mediante la aplicación de mecanismos de derivación. Es de esperar que exista una relación semántica entre las palabras de dicho conjunto, relación que normalmente es de tipo proceso-resultado (por ejemplo *fijación-fijado*), proceso-agente (por ejemplo *inhibi-*

ción-inhibidor), y similares.

Los mecanismos básicos de derivación en español y gallego son la *prefijación*, la *sufijación apreciativa*, la *sufijación no apreciativa*, la *parasíntesis* y la *derivación regresiva*. Un fenómeno importante a tener en cuenta es la existencia de *alomorfos*. Para obtener unos patrones regulares de formación de palabras podemos valernos de las llamadas *reglas de formación* [10], basadas en teorías tales como la Gramática Generativa Transformacional y el desarrollo de la denominada Fonología Derivativa. Aunque dicho paradigma no es completo, supone un avance considerable puesto que nos permite diseñar un sistema de generación automática de familias morfológicas con un grado aceptable de corrección y exhaustividad [13].

3.2.4 Analizador sintáctico superficial

En el ámbito de la recuperación de información se denomina *término multipalabra* a aquel término que contiene dos o más palabras con contenido (sustantivos, verbos y adjetivos). En la literatura se describen varios métodos para su obtención.

Uno de los más utilizados es el denominado *simplificación del texto*: en una primera fase, se realiza un *stemming* de las palabras individuales, y se procede a eliminar las *stopwords*; posteriormente se extraen y normalizan los términos empleando para ello patrones o técnicas estadísticas. Existe, pues, una clara falta de base lingüística en dichas operaciones[2], lo que redunda frecuentemente en simplificaciones erróneas. Sin embargo, es el método más sencillo y menos costoso. Existen otros métodos de base lingüística que realizan un *análisis sintáctico* del texto, el cual devuelve a su salida un conjunto de árboles que denotan relaciones de dependencia entre las palabras involucradas. De este modo, estructuras con relaciones de dependencia similares pueden ser normalizadas a una misma forma. A medio camino estaría la *correspondencia de patrones*, que se basa en la hipótesis de que las partes más informativas del texto siguen unas construcciones sintácticas bastante bien definidas que se pueden aproximar mediante patrones.

La aproximación que hemos incorporado en nuestro sistema conjuga los dos últimos métodos sobre la base de la indexación de los sintagmas nominales y de sus *variantes sintácticas y morfosintácticas* [8]. Desde el punto de vista morfológico, las variantes sintácticas hacen referencia a la morfología flexiva, mientras que las morfosintácticas entran además en el ámbito de la morfología derivativa. En lo referente a la sintaxis, las variantes sintácticas tienen un campo de actuación mucho más restringido, el sintagma nominal, mientras que las variantes morfosintácticas lo amplían a prácticamente toda la oración, incluyendo los verbos y sus objetos.

Las consultas realizadas a un sistema de recuperación de información suelen expresarse en forma de grupos nominales de complejidad diversa. Teniendo esto en cuenta tomaremos los grupos nominales como términos base a partir de los cuales, mediante la aplicación de las transformaciones correspondientes, se obtendrán sus variantes sintácticas y morfosintácticas, no necesariamente grupos nominales. Una vez definidos los árboles básicos de las frases nominales y de sus variantes, éstos pueden ser compilados en un conjunto de expresiones regulares que serán emparejadas con el texto para extraer los pares de dependencias que serán utilizados como términos índice, tal y como se describe en [12]. De esta forma, estaremos identificando los pares de dependencias mediante un simple emparejamiento de patrones sobre la salida del etiquetador/lematizador, trabajando siempre con técnicas de estado finito, lo que conlleva una importante reducción en lo que respecta al tiempo de ejecución.

4 Evaluación preliminar del sistema

4.1 Los corpus utilizados en la evaluación

En este apartado describimos el proceso que llevó a la obtención del corpus utilizado para evaluar el sistema de RI descrito en los apartados previos, así como el corpus. Se trata de un corpus de evaluación creado *ad hoc*, pensando en la evaluación del sistema de RI propuesto, para su aplicación, en primer lugar, en un entorno muy concreto, *Editorial Compostela*[3].

[2] Por ejemplo, algunas *stopwords* tales como artículos y preposiciones son componentes clave de la estructura sintáctica.

[3] A pesar de ello, creemos que puede ser de utilidad general para evaluar sistemas de RI independientes de

Con tal fin, *Editorial Compostela* facilitó, en el formato HTML propio de su edición en web, una base de datos documental de 61.806 textos de noticias en español o gallego, publicadas entre las fechas 10 de febrero de 1999 y 15 de marzo de 2001, y distribuidas en secciones tal como aparece en las tres primeras columnas de la Tabla 2.

Aunque la asignación de cada noticia a una sección del periódico estaba explícitamente recogida en el documento HTML que la contenía, su redacción en una u otra lengua, no, por lo que hubo que desarrollar herramientas para la asignación automática de lengua a cada uno de los documentos. Siendo gallego y español lenguas tan próximas desde todos los puntos de vista, pero especialmente para lo que aquí nos interesa, desde los puntos de vista fónico, gráfico y léxico, las técnicas estadísticas más habituales, basadas en la identificación de secuencias de caracteres (N-gramas, véase [3]) o palabras cortas (véase [7]), no resultaban efectivas por sí solas, así que hubo que enriquecerlas con el recurso a técnicas adicionales basadas en el uso de información lingüística extraída de diccionarios en ambas lenguas, en concreto mediante la identificación de palabras gramaticales frecuentes y exclusivas de una de las lenguas frente al conjunto del léxico de la otra.

Los textos fueron además tratados para convertirlos del formato HTML proporcionado por la edición web del periódico a un formato SGML especialmente diseñado para el almacenaje en una base de datos documental adecuada para RI de acuerdo simultáneamente con los requerimientos, por un lado, del propio sistema de RI propuesto y, por otro, de *Editorial Compostela* en tanto que usuario de la aplicación. En este formato, cada documento constituye un elemento <NOTICIA>, que comprende los elementos, de contenido evidente a partir de su denominación, <LENGUA>, <FECHA>, <SECCION>, <AUTOR> (opcional), <LUGAR> (opcional), <PRETITULAR> (opcional), <TITULAR>, <RESUMEN> (opcional), <CUERPO>, que recoge el texto principal de la noticia, y uno o más elementos <APARTADO> (opcionales), que recogen generalmente noticias en cierto modo unitarias en sí mismas, pero relacionadas con la que se considera noticia principal en el documento en cuestión. Estos elementos <APARTADO> constan a su vez de <TITULAR> y <CUERPO>, y éste último contiene siempre uno o una serie de elementos <P> (párrafo).

La última fase de elaboración del corpus de evaluación es ya estrictamente manual y consiste i) en la búsqueda de posibles interrogaciones a la base de datos documental y ii) en la identificación de la que sería la respuesta deseable del sistema a las mismas. Como el examen pormenorizado de 61.806 documentos resultaba imposible, optamos por una reducción del corpus a unas dimensiones manejables en el lapso de tiempo disponible: seleccionamos los primeros 10.000 que en español correspondían a miércoles, viernes o domingo a partir de la fecha que fijamos de inicio, el 23 de noviembre de 1999, y que, de este modo, resultaron comprendidos entre esa fecha y el 10 de octubre de 2000. Dado que entre esas fechas había más de 10.000 textos en gallego, seleccionamos aleatoriamente entre ellos los 10.000 que entrarían en el corpus, respetando estrictamente la que habría sido la proporción por sección en esa lengua si, como habíamos hecho para el español, hubiéramos tratado simplemente los 10.000 primeros textos correspondientes a miércoles, viernes y domingo a partir del 23 de noviembre de 1999. El corpus resultante responde, en cuanto al número de documentos, a los datos recogidos en las tres últimas columnas de la Tabla 2.

Todos los documentos de este corpus fueron leídos para identificar en ellos a) interrogaciones posibles formuladas en lenguaje natural, relevantes para series de documentos, y b) todos los documentos relevantes para las interrogaciones identificadas. El método de búsqueda fue estrictamente manual, en él participaron unas 10 personas que atendían a secciones concretas y distintas. Tiene dos problemas fundamentales: a) aunque no es de esperar que los usuarios finales, los reporteros de *Editorial Compostela*, sean mucho más originales al respecto, evidentemente la lectura de los documentos condiciona inevitablemente la formulación de la interrogación, y b) los límites que determinan la relevancia de un documento para una interrogación son

la lengua o que trabajen específicamente con español o gallego. Es, de hecho, nuestra intención distribuir libremente el corpus para evaluación desarrollado, para lo cual estamos actualmente en negociaciones con el grupo *Editorial Compostela*, a quien pertenecen los textos.

	Corpus original			Corpus reducido		
	Castellano	**Gallego**	**Total**	**Castellano**	**Gallego**	**Total**
Cultura-Sociedad	321	7182	7503	70	2404	2474
Deportes	7050	181	7203	1752	79	1831
Galicia	4339	11158	15497	1796	3388	5184
Internacional	458	3499	3957	169	985	1154
Local	13149	153	13302	4999	51	5050
Nacional	994	5511	6505	318	1606	1924
Opinión	2813	4998	7811	896	1487	2383
Total	29124	32682	61806	10000	10000	20000

Tabla 2: Datos del corpus de evaluación.

muy difusos, un documento suele tener como foco esencial un tema, pero estar relacionado, de modo más o menos estrecho, con más de uno. Se identificaron en total 230 interrogaciones posibles para evaluar el sistema.

4.2 Ejecución y valoración de los resultados

El rendimiento del sistema depende tanto de las características de las consultas como de las de los documentos sobre los que se va a realizar la búsqueda. Utilizaremos dos ejemplos para ilustrar el comportamiento de los diferentes métodos de búsqueda incorporados en LEIRA. En lugar de calcular unas medidas de precisión y cobertura sobre el total de documentos devueltos, aplicaremos una visión de usuario, indicando el número de documentos relevantes que aparecen en las dos primeras páginas, esto es, los primeros 20 documentos devueltos. Por razones prácticas hemos limitado a 100 el número máximo de documentos devueltos para cada consulta.

En primer lugar, consideraremos la consulta en gallego *Aumentos na afiliación á seguridade social*[4]. La búsqueda simple con *stemmer* devuelve el número máximo permitido de documentos, pero solamente se encuentran 8 relevantes entre los 20 primeros. La utilización de lematización incrementa el rendimiento, de tal forma que 14 de los 20 primeros documentos son relevantes. El número total de documentos devueltos sigue siendo 100. La utilización de familias morfológicas permite incrementar a 16 el número de relevantes entre los 20 primeros. La utilización de pares de dependencia sintáctica disminuye drásticamente el número de documentos devueltos, que pasa a ser de 17, con 15 relevantes. El método de búsqueda que comprueba que todos los lemas de las palabras con contenido aparezcan en los documentos devuelve tan sólo 6 documentos, pero todos ellos relevantes.

Consideraremos ahora la consulta en español *Retirada de la inmunidad parlamentaria a Pinochet*, para la que sólo hay 4 documentos relevantes. Las búsquedas basadas en la utilización de *stemmer*, lematización y familias morfológicas devuelven el número máximo de documentos permitido, situando los 4 relevantes entre los 4 primeros. La utilización de pares de dependencia sintáctica logra devolver únicamente los 4 documentos relevantes. El método de búsqueda que comprueba que todos los lemas de las palabras con contenido aparezcan en los documentos devuelve 2 de éstos, ambos relevantes. Como se puede observar, las búsquedas basadas en *stemmers*, lematización y familias morfológicas devuelven un gran número de documentos irrelevantes, aunque logran ubicar un buen número de los relevantes entre los 20 primeros. La característica más destacable de la búsqueda mediante pares de dependencia sintáctica consiste en que devuelve muy pocos documentos irrelevantes. Si exigimos que todas las palabras con contenido de la consulta aparezcan en el documento, el número de documentos irrelevantes devueltos disminuye incluso más, pero a costa de perder, en algunos casos, una porción importante de los relevantes.

5 Futuros desarrollos

En lo concerniente a la evolución futura, destacamos las siguientes líneas de trabajo:

- En lo que respecta al proceso de etiquetación, diseñar una extensión del algoritmo de Viterbi que permita realizar la selección de la secuencia de *tokens* más probable al mismo tiempo que se realiza la etiquetación de todas ellas, sin necesidad de evaluar por separado cada uno de los posibles flujos de *tokens*.

[4] Aumentos en la afiliación a la seguridad social.

- En lo que respecta a las familias morfológicas, estudiar el impacto de los diferentes mecanismos de derivación en el rendimiento de la indexación.
- Respecto al analizador sintáctico superficial, realizar una extracción más precisa de los pares de dependencias mediante la utilización de cascadas de autómatas finitos en lugar del emparejamiento de expresiones regulares.
- En relación al proceso de indexación, realizar una expansión de los términos de la consulta en función de una relación de sinonimia ponderada. También estamos estudiando la viabilidad de integrar un mecanismo de comparación de árboles, tal como el descrito en [14].

Referencias

[1] C. Álvarez Lebredo, P. Alvariño Alvariño, A. Gil Martínez, T. Romero Quintáns, M.ª P. Santalla del Río, S. Sotelo Docío. 1998. AVALON: Una Gramática Formal basada en Corpus. *Procesamiento del Lenguaje Natural*, 23, páginas 132–139.

[2] T. Brants. TnT - a statistical part-of-speech tagger. 2000. In *Proc. of ANLP'2000*, Seattle.

[3] W. Cavnar, J. Trenkle. 1994. N-gram based text categorization. Symposium on Document Analysis and IR, Universidad de Nevada, Las Vegas.

[4] J. Graña Gil, F. M. Barcala Rodríguez, M. A. Alonso Pardo. 2001. Compilation methods of minimal acyclic automata for large dictionaries. In B. W. Watson, D. Wood (eds.), *Proceedings of CIAA 2001*, páginas 116–129, Pretoria, Sudáfrica.

[5] J. Graña Gil, F. M. Barcala Rodríguez, J. Vilares Ferro. 2002. Formal methods of tokenization for part-of-speech tagging. In A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, volumen 2276 de *LNCS*, páginas 240–249. Springer-Verlag, Berlín-Heidelberg-Nueva York.

[6] J. Graña Gil, J.-C. Chappelier, M. Vilares Ferro. 2001. Integrating external dictionaries into stochastic part-of-speech taggers. In *Proc. of RANLP 2001*, páginas 122–128.

[7] G. Grefenstette. 1995. Comparing two language identification schemes. 3rd International Conference on Statistical Analysis of Textual Data, Roma.

[8] C. Jacquemin, E. Tzoukermann. 1999. NLP for term variant extraction: synergy between morphology, lexicon and syntax. In T. Strzalkowski (ed.), *Natural Language Information Retrieval*, volumen 7 de *Text, Speech and Language Technology*, páginas 25–74. Kluwer Academic Publishers, Dordrecht/Boston/Londres.

[9] F. Jelinek. 1998. *Statistical Methods for Speech Recognition.* MIT Press, Cambridge, MA.

[10] M. F. Lang. 1990. *Spanish Word Formation: Productive Derivational Morphology in the Modern Lexis.* Croom Helm. Routledge, Londres y Nueva York.

[11] C. Samuelsson. 1993. Morphological tagging based entirely on bayesian inference. In R. Eklund (ed.), *Proceedings of the 9th Nordic Conference on Computational Linguistics*, Estocolmo, Suecia.

[12] J. Vilares Ferro, F. M. Barcala Rodríguez, M. A. Alonso Pardo. 2002. Using syntactic dependency-pairs conflation to improve retrieval performance in Spanish. In A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, volumen 2276 de *LNCS*, páginas 381–390. Springer-Verlag, Berlín-Heidelberg-Nueva York, 2002.

[13] J. Vilares Ferro, D. Cabrero Souto, M. A. Alonso Pardo. 2001. Applying productive derivational morphology to term indexing of Spanish texts. In A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, volumen 2004 de *LNCS*, páginas 336–348. Springer-Verlag, Berlín-Heidelberg-Nueva York, 2001.

[14] M. Vilares Ferro, F. J. Ribadas Pena, J. Graña Gil. 2001. Approximately common patterns in shared-forests. In H. Paques, L. Liu, D. Grossman (eds.), *Proc. of ACM CIKM 2001*, páginas 73–80, Atlanta, Georgia, USA. ACM.

[15] A. J. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Trans. Inform. Theory*, 13:260–269.