RAQUEL GÓMEZ DÍAZ

ESTUDIO DE LA INCIDENCIA DEL CONOCIMIENTO LINGÜÍSTICO EN LOS SISTEMAS DE RECUPERACIÓN DE LA INFORMACIÓN PARA EL ESPAÑOL

EDICIONES UNIVERSIDAD DE SALAMANCA

COLECCIÓN VITOR
76
C
Ediciones Universidad de Salamanca
Raquel Gómez Díaz

1ª Edición: Enero, 2002
I.S.B.N.: 84-7800-831-4
Depósito Legal: S.1785-2001
Ediciones Universidad de Salamanca
Apartado postal 325
E-37080 Salamanca (España)

Edeltex, S.L.
C/ Valle Inclán 23, 4º B
37007 – Salamanca
Tfno: 923 238705

Impreso en España-Printed in Spain

CEP. Servicio de Bibliotecas

GÓMEZ DÍAZ, Raquel
Estudio de la indicidencia del conocimiento lingüístico en los sistemas de recuperación de la información para el español [Archivo de ordenador] / Raquel Gómez Díaz.—1ª ed.—Salamanca : Ediciones Universidad de Salamanca, 2001
1 disco compacto.—(Colección Vitor ; 76)
Tesis-Universidad de Salamanca, 2001
1 Universidad de Salamanca (España)-Tesis y disertaciones académicas.
2. Recuperación de la información.3.Búsqueda documental automatizada.
4. Español (Lengua)

Resumen

Hoy en día es necesario estar bien informado, por las características de la información necesitamos sistemas que trabajen con lenguaje natural o donde el control de los términos sea mínimo.

Para este trabajo hemos creado un lematizador mediante un máquina de estados finitos no determinista con el fin de aplicarlo a la recuperación de información en español. La función del lematizador es eliminar los sufijos de manera utomática y establecer su lema. A partir de los lemas se hace la indización y posteror recuperación. Para probar la eficacia del mismo, se realizan experimentos de lematización flexiva y derivativa, combinando esto con la supresión de palabras vacías.

Abstract

Nowaday it is very importan to be well informed, and because of the characteristic of the information we a need a system to work with natural languaje or with minimum ter control.

A stemmer was created by means of non-determnistic finite state machine to be applied to information retrievan in Spanish. The functtions of this stemmer is to remove the suffixes and to establish the stem of the words. This is done for the indexing and subsequient retrieval of the documents. The efficiency of the stemmer has been proved by test of flexinal and derivative stemming, together with the removal of stop words.

Índice general

Índice de dibujos

Índice de tablas

Índice de ecuaciones

Índice de gráficos

INTRODUCCIÓN

1. Introducción.

Hoy en día, nadie duda de la necesidad de estar bien informado. Debido al crecimiento exponencial de la producción científica, el volumen de datos que tenemos que manejar, crece sin parar. Por un lado, hay mucha más información de la que somos capaces de asimilar, lo que Blair y Maron denominan *sobrecarga informativa*[1]; y por otro, no toda la información que se genera es válida, lo que Pablo de la Fuente denomina *contaminación informativa*[2]. Esta situación hace que cada vez sea más difícil encontrar la información verdaderamente útil.

En los últimos años hemos ido asistiendo al cambio de los soportes que contienen la información, y de los mecanismos de difusión de la misma; a esto hay que añadir que cada vez tenemos ordenadores con una capacidad mayor, lo que hace posible crear grandes bases de datos donde se contiene mucha más información que en décadas pasadas[3]. A la capacidad individual de los ordenadores, hay que añadir el potencial que tienen cuando se conectan en red. No podemos hablar de información, ya sea de su tratamiento o de su difusión, sin mencionar la importancia de Internet, que está poniendo a disposición de los usuarios gran volumen de información a bajo coste.

Este gran volumen de información está provocando varios problemas. Por un lado la capacidad de digerir tanta información por parte de los usuarios no

---

[1] D.C. BLAIR and M. E. MARON An evaluation of retrieval effectiveness for a full-text document retrieval systems. *Comunication to AC*M March 1985 28 (3) p. 289-299

[2] P. DE LA FUENTE REDONDO. *Bibliotecas digitales*. [Conferencia pronunciada en Valladolid el 16 de marzo de 1998 en "*Nuevas tendencias en gestión de la Información*". Valladolid 12 al 18 de Marzo de 1998.]

[3] C. BELTRÁN. *Modelo informático de recuperación documental*. [en línea] <http://www.ucm.es/info/multidoc/revista/cuadern5/ceseda.htm [consultado el 17/06/99]

ha crecido de la misma manera que la producción de información[4]. Por lo tanto, cuanto mayor es el volumen de información disponible, los problemas de recuperación serán mayores[5], por lo que cada vez se hacen más necesarios sistemas que seleccionen bien, aquellos documentos que responden a las necesidades de los usuarios, descartando los que no lo hacen.

Por otro lado, los sistemas de tratamiento y recuperación que se vienen aplicando a Internet, que eran útiles hace años cuando las búsquedas se hacían con un volumen menor y la información variaba más lentamente, hoy en día ya no son tan útiles, por lo que es necesario buscar nuevos métodos que faciliten el tratamiento, y el acceso a esa gran cantidad de información que cada día se genera[6].

Partiendo de estas ideas, decidimos buscar un tema de investigación que pudiera contribuir a la mejora de los sistemas de tratamiento y recuperación de la información, teniendo muy presente que una de las cualidades del sistema fuera la facilidad de utilización para los usuarios finales de la información. Después de revisar trabajos sobre recuperación de información, y reflexionar sobre las ideas antes mencionadas, elegimos estudiar las aplicaciones del lenguaje natural a la recuperación de información, por tres motivos: la facilidad de uso que el lenguaje natural tiene para los usuarios[7], también porque éste, como veremos a continuación, implica un ahorro de tiempo y, finalmente, por la actualidad del tema.

En los sistemas tradicionales de recuperación de información, el usuario expresaba su demanda informativa al documentalista, que era el que la

---

[4] P. JACOB Text interpretation: Extracting Information En *Survey of the State of the Art in Human Language Technology*. Oregon: National Science Foundation, 1995 p 263-265

[5] Mª D. OLVERA LOBO Métodos y técnicas para la indización y la recuperación de recursos de la World Wide Web. *Boletín de la Asociación Andaluza de Bibliotecarios.* 1999 n. 57 p. 11-22

[6] D. HARMAN, P. SHAÜBLE, A. SMEATON. Document Retrieval En *Survey of the State of the Art in Human Language Technology*. Oregon: National Science Foundation, 1995. p. 259-262

[7] A. G. TAILOR. *The organization of information*. Englewood: Libraries Unlimited Inc, 1999.

interpretaba y la traducía al lenguaje en el que estaba la base de datos (lenguaje controlado), hacía las búsquedas pertinentes y le devolvía al usuario la respuesta obtenida. En estos pasos necesarios hay dos problemas: el primero es el tiempo, no sólo el que empleaba el usuario en comunicar lo que quería y el documentalista en interpretarlo y hacer la recuperación, sino que también el que el profesional de la información, tardaba en preparar la base de datos para el proceso. El otro problema de los sistemas tradicionales es la limitación que supone el tener que usar un lenguaje controlado para hacer la indización y las búsquedas, por dos motivos, por la propia característica de este tipo de lenguaje: la rigidez, y que no siempre es conocido por los usuarios. También porque como la manera de hacer la indización es manual, interviene la subjetividad de los indizadores y es muy común que un documento si es indizado por dos personas distintas se le asignen términos diferentes. La solución a este problema viene de la mano de la utilización del lenguaje natural en el proceso de recuperación, ya que así reducimos mucho el tiempo de preparación de la base de datos y obtenemos la gran ventaja de que es el propio usuario, sin la necesidad de especialistas que hagan de intermediarios, el que puede plasmar su demanda, en una estrategia de búsqueda que él mismo desarrollará, sin la necesidad del especialista, puesto que es el propio usuario quien mejor conoce su necesidad informativa.

En cuanto a la actualidad del tema, si analizamos tanto las publicaciones periódicas (*Journal of the American Society for Information Science, Journal of Documentation*, *Journal of Information Science*...) como los congresos internacionales (*TREC Conference, CLEF*...), más importantes referidos a los temas de recuperación de información, podemos ver, cómo las últimas tendencias en recuperación de la información están en la línea del procesamiento del lenguaje natural. En este sentido, hay que decir que la mayor parte de los trabajos que se han realizado y se están realizando proceden del área anglosajona. En cambio, los trabajos para el español son escasos como mostraremos más adelante. Por eso, una parte significativa de la novedad de nuestra investigación es la lengua elegida.

Dando vueltas a estas ideas, hemos ido concretando el tema de investigación hasta centrar el trabajo en la búsqueda de un sistema que aplica conocimiento lingüístico a la recuperación de información en español. Quizá la elección de por qué en español parece obvia, dado el contexto donde se desarrolla este trabajo, pero la revisión bibliográfica nos ha servido para darnos cuenta de que el problema que tienen algunos sistemas de recuperación que aplican conocimiento lingüístico, es precisamente que los que hacen la aplicación no son hablantes de la lengua que pretenden aplicar, por lo que se cometen errores que un hablante de la misma no cometería. La razón para elegir en concreto los sufijos, dentro del conocimiento lingüístico, es porque es el mecanismo de producción léxica del español más importante.

2. Objetivos.

El principal objetivo de este trabajo es mostrar cómo influye la aplicación del conocimiento lingüístico en los sistemas de R.I. para el español. Junto con este objetivo están los siguientes, que no son más que el desarrollo y complemento del mismo.

Respecto a los objetivos aquí marcados hay que indicar que no están puestos en orden jerárquico.

1. Ver cuál es el estado de la cuestión de la recuperación de la información: modelos más importantes, medidas de evaluación experimentos más significativos hechos con el español.

2. Mostrar si es eficaz un modelo de recuperación basado en información no estructurada en campos.

3. Hacer un estudio más detallado de la lematización y de los algoritmos de lematización, tanto de los elaborados para el inglés como los realizados para otros idiomas. Ver las distintas clasificaciones que hay al respecto.

4. Ver si es posible la creación para el español de un lematizador flexivo y otro derivativo mediante una máquina de estados finitos.

5. Si es posible la creación del lematizador, ver si se puede aplicar a la recuperación de información y si ello produce mejoras en términos de precisión y exhaustividad en las búsquedas. Establecer qué tipo de lematización es más ventajosa para la recuperación de información en español, si la flexiva o la derivativa.

6. Mostrar cómo incide la eliminación de palabras vacías en la recuperación, qué criterios se deben elegir a la hora de crear las dichas listas. Mostrar si hay diferencias significativas entre los distintos tipos de listas.

3. Antecedentes.

La idea del tema elegido, surgió del estudio de los trabajos realizados para las TREC[8] en el periodo 1994-1996, en concreto de los lematizadores para el español que allí se presentaron. Una vez analizados los problemas que dichos trabajos tenían, pensamos que haciendo un estudio más exhaustivo de las peculiaridades del español, el sistema podría tener un rendimiento mejor. Con esta idea, en junio de 1998 presentamos el trabajo de Grado de licenciatura[9] en esta misma Universidad. Hoy, años más tarde, analizados de nuevo aquellos

[8] http://trec.nist.gov

[9] R. GÓMEZ DÍAZ. La Recuperación de la Información en español: evaluación del efecto de sus peculiaridades lingüísticas. Universidad de Salamanca. Trabajo de Grado, 1998. [trabajo no publicado]

resultados, tratamos de seguir profundizando en la idea de que la aplicación de la información lingüística a la recuperación de información en español, puede aportar mejoras a los sistemas que lo apliquen.

4. Dificultades.

El trabajo de investigación no ha sido fácil, en concreto las dificultades que nos hemos ido encontrando se pueden sistematizar en tres grupos: derivadas de la investigación en recuperación de la información, de los sistemas que trabajan con lenguaje natural y de los sistemas que aplican el español.

En primer lugar, están las derivadas de la investigación en recuperación de la información. En este sentido, uno de los principales problemas es la dificultad de definir, tanto conceptual como operativamente, tal y como señala Olvera Lobo[10], muchos de los conceptos que aquí se van a manejar, por ejemplo el simple concepto de *necesidad informativa* se puede definir desde distintos enfoques, como veremos más adelante. Relacionado también con la disciplina de estudio, está el hecho de que en España hay muy poca investigación en el área de la documentación, y dentro de ésta, la recuperación de la información ocupa un nivel muy escaso[11]. Esto hace que no hayamos encontrado ningún trabajo parecido para el español, que nos sirva de referente, por lo que casi todos los referentes utilizados son anglosajones, con la dificultad añadida que supone, al tratarse de un trabajo donde la base lingüística es fundamental. Además, al no existir muchos trabajos en el español sobre recuperación de información, la

[10] Mª D. OLVERA LOBO. Evaluación de sistemas de recuperación de información: aproximaciones y nuevas tendencias. *El profesional de la información*. 1999 Vol. 8 (11) p. 4-14

[11] Para mayor información de los porcentajes de autores citados y de la representatividad dentro de cada área consultar Moya Anegón, Félix. La investigación española en Recuperación de Información (R.I.): análisis bibliométrico (1984-1999). EN *Revista de investigación Iberoamericana en Ciencia de la Información y documentación*. 2000 1 (1) 117-123

terminología no está suficientemente asentada, por lo que algunos términos se emplearán en inglés (cluster, browsing...), porque utilizar su traducción literal puede inducir a error, aunque siempre que exista el término establecido en español lo utilizaremos.

Otra de las dificultades encontradas, relacionada con el tema de investigación, como señalan Gil Leiva y Rodríguez Muñoz[12], es que al tratarse de un área interdisciplinar se han tenido que emplear conceptos y herramientas de lingüistas e informáticos, lo que dificulta el trabajo de investigación para un documentalista, aunque también hay que decir que la formación documental es muy importante en este tipo de investigaciones porque es necesario conocer bien el proceso documental, y sobre todo no perder la perspectiva de los usuarios. Ambas cosas, muy fáciles para un documentalista.

El segundo grupo de dificultades son las provenientes de los sistemas que trabajan con lenguaje natural: por un lado aunque el hablante conoce las reglas derivativas y el orden establecido en que se aplican, este conocimiento no es reflexivo por lo que resulta difícil establecer las reglas de manera que se puedan aplicar al lematizador. También hay que tener en cuenta que el hablante nativo de una lengua tiene la capacidad para reconocer palabras posibles y no posibles, pero es difícil a la hora de elaborar un sistema lingüístico dotarle de esta capacidad.

El tercer grupo de dificultades son propias de un sistema que trata de aplicar conocimiento lingüístico español. Por un lado, hay que tener en cuenta que tiene rasgos tipológicos de varias lenguas y la complejidad morfológica del español, como mostraremos más adelante, lo que hace necesario tener en consideración un mayor número de aspectos que si se tratara de una lengua más

[12] I. GIL LEIVA, J. V. RODRÍGUEZ MUÑOZ El procesamiento del lenguaje natural aplicado al análisis de contenido de los documentos. *Revista General de Información y Documentación*. 1996 Vol. 6 (2) 2 p. 205-218

"pura". El otro problema también relacionado con el idioma, es específico del conocimiento que pretendemos aplicar: los sufijos. En el caso de la derivación en español, hay que tener presente una amplia lista de sufijos con una considerable lista de variantes alomórficas[13].

5. El español.

Antes de comenzar con el desarrollo del trabajo, creemos que es necesario explicar brevemente los rasgos tipológicos y en qué consiste la complejidad morfológica del español, para que en el momento de desarrollar herramientas lingüísticas, se reduzcan los errores.

Tradicionalmente se han establecido dos criterios para la clasificación de las lenguas, el genealógico y el tipológico. El primero de ellos se basa en el supuesto de que las lenguas se han separado de un antecesor común. El tipológico, se basa en la comparación de las similitudes formales existentes en las distintas lenguas e intenta agruparlas en tipos estructurales basándose en su fonología, gramática o vocabulario, en lugar de en sus relaciones históricas. Este segundo criterio fue el que eligió Schelicher para hablar de lenguas **aislantes**, **aglutinantes** y **flexivas**[14], aunque en la realidad no suelen presentarse los tipos puros.

En las lenguas aislantes, analíticas o de raíces, las palabras son invariables, no hay terminaciones. Suelen estar formadas por monosílabos que adquieren un sentido concreto y preciso en la frase. Las relaciones gramaticales se manifiestan en el orden la las palabras. Ejemplos de estas lenguas son el chino y el vietnamita.

---

[13] Cf. I. BOSQUE, V. DEMONTE (dir) *Gramática descriptiva de la lengua española*. Madrid: Espasa, vol III p 4305-5096

[14] F. LÁZARO CARRETER *Diccionario de términos filológicos* 3ª ed. Madrid: Gredos, 1987 p 32, 189, 248.

Las flexivas, sintéticas o fusionales son en las relaciones gramaticales se expresan combinando la estructura interna de las palabras, generalmente cambiando el uso de las terminaciones flexivas que reflejan simultáneamente varios significados gramaticales. Ejemplos de estas lenguas son el latín, el griego y el árabe.

Las aglutinantes o aglutinativas son en las que las palabras se forman por una secuencia de unidades, cada una de las cuales expresa un significado gramatical particular. Pertenecen a este grupo de lenguas las que usan prefijos y sufijos, como son el turco, el finés y el japonés. Dentro de este grupo están también lo que algunos expertos denominan lenguas polisintéticas o incorporantes, que son aquellas que están formadas por palabras muy largas y complejas y tiene una mezcla de rasgos aglutinantes y flexivos, como es el esquimal.

A la hora de adscribir una lengua a uno de estos grupos no podemos olvidar las relaciones culturales que se dan entre las lenguas, sobre todo a través de los préstamos lingüísticos. Por esta razón, al tratar de clasificar el español, nos encontramos con que genealógicamente es una lengua romance pero desde el punto de vista cultural no solo está relacionada con otras lenguas con un origen común, como puede ser el francés, sino que se relaciona también con lenguas como el árabe o el inglés al incorporar términos procedentes de ellas. Desde el punto de vista tipológico, el español se parece más a una lengua flexiva como el latín que una aislante como el chino (las desinencias de las palabras informan más de la función gramatical que el orden en que aparezcan) y sin embargo podemos encontrar rasgos de varios tipos de lenguas, así tomando el siguiente ejemplo podemos ver como tiene características de los tres grupos:

- Aislante: El rey da pan al can.

- Flexiva: Los reyes dieron buenísimos panes.

- Aglutinante: Anti-inflaccion-ista.[15]

Esto nos muestra la complejidad de nuestro idioma lo que dificulta el desarrollo de herramientas lingüísticas, y si queremos construir herramientas eficaces no podemos perder esto de vista.

A esta lengua también se la denomina castellano, pero en este trabajo utilizaremos el término *español*, ya que según Lapesa[16], desde el siglo XVI tiene absoluta justificación y se sobrepone al de lengua castellana o castellano. Según explica Menéndez Pidal, sus orígenes están en el latín vulgar, propagado en España desde finales del siglo III a.C. No hay que olvidar que el español es una lengua que a lo largo del tiempo ha ido incorporando en distintos momentos a su léxico, términos de otras lenguas con raíces distintas como es el caso del griego, de los pueblos germánicos, del árabe...

El español, como todas las lenguas románicas, es flexivo, aunque en menor medida de lo que fue el latín. Conserva desinencias para el género, pero perdió el neutro en los nombres y los adjetivos aunque lo conservó en los pronombres como *eso, lo vuestro,* y en el artículo determinado *lo*[17].

Ya en el siglo VI, las desinencias de los casos de los nombres, habían sido sustituidas por el empleo de las preposiciones, al igual que en el resto de las lenguas románicas.

Los verbos redujeron de cuatro a tres las conjugaciones del latín. El verbo español posee desinencias para las personas, el número, el tiempo, el modo

---

[15] Ejemplo tomado de D. Crystal. *Enciclopedia del lenguaje de la Universidad de Cambridge*. Madrid: Taurus, 1994 p. 106

[16] R. LAPESA *Historia de la lengua española*. 9ª ed. Cor y aum. Madrid: Gredos, 1988 p. 299

[17] R. MENÉNDEZ PIDAL Manual de gramática histórica española. 20ª ed. Madrid: Espasa-Calpe, 1989 p 213-217

y la voz. Por su conjugación podemos hacer una clasificación en los verbos regulares e irregulares, aunque éstos se pueden agrupar en distinto número de modelos según los autores consultados[18].

Toda esta complejidad tendrá que ser tenida en cuenta a la hora de diseñar cualquier herramienta lingüística.

6. Estructura del trabajo.

El trabajo está estructurado en cinco capítulos, más una introducción, un glosario de términos y un apéndice.

En el primero, **La Recuperación de información**, hacemos una revisión bibliográfica del concepto de recuperación de información, según la visión de distintos especialistas; y se hace un recorrido por los diferentes modelos de recuperación de información; la revisión de los principales conceptos y medidas de evaluación más utilizadas en recuperación de la información. Finalmente hemos incluido aquí los principales los experimentos de recuperación realizados para el español

En el segundo capítulo, **La lematización,** se hace un estudio de qué es la lematización, comenzando por el propio término. También se hace una revisión bibliográfica de los distintos algoritmos de lematización, tanto los realizados para el inglés, como los de otros idiomas. Finalmente, se estudian los distintos enfoques de la evaluación en lematización.

El capítulo tercero, **El lematizador**, es el más importante de este trabajo. Es el fruto del estudio de todo lo anterior y se trata de poner en práctica lo aprendido de los experimentos realizados tanto para el inglés como para otros

[18] Vid S. ALCOBA La flexión verbal EN I. BOSQUE, V. DEMONTE (1999) op. cit p. 4917-4991

idiomas. Aquí aplicamos el conocimiento lingüístico específico, en nuestro caso el de los sufijos; para ello ha sido necesario estudiar los distintos autores especialistas en la materia. En este capítulo también se explican todos los pasos que han sido necesarios en la creación del lematizador y su posterior aplicación a la recuperación de información. Finalmente se miden los resultados de los experimentos y se evalúan para extraer las conclusiones.

En el capítulo cuarto, **revisión de objetivos y conclusiones**, se analizan las conclusiones y se revisa en qué medida se han conseguido los objetivos marcados en la introducción.

En el capítulo quinto se da la **bibliografía** que hemos utilizado en la realización de este trabajo.

Hemos querido incluir un **glosario de términos** para contribuir al asentamiento de la terminología, ya que como indicábamos al principio es una de las carencias que encontramos al inicio de la investigación en recuperación de información en español. Así mismo consideramos que es muy útil para ayudar a clarificar conceptos, que no siempre es procedente aclarar entre el texto.

Los **índices** ayudarán a la localización de las partes del texto así como las ecuaciones, gráficos, dibujos y tablas que se encuentran repartidas a lo largo del texto.

El apéndice final tiene como fin dar información complementaria de los resultados por cada una de las preguntas de cada experimento.

I LA RECUPERACIÓN DE INFORMACIÓN

1. Concepto de recuperación de información.

El concepto de Recuperación de Información (en adelante R.I.) es relativamente reciente. En 1951, Calvins Mooers[19] lo utilizaba con el sentido de "*el proceso o el método por el cual un usuario es capaz de convertir su necesidad informativa, en una lista de citas de documentos almacenados, que contienen la información útil para él*". De este modo la R.I. abarca el aspecto intelectual de la descripción de información y su especificación para la búsqueda, y también todo lo que los sistemas, tanto técnicas como máquinas empleadas, conllevan. Según esta definición la R.I. abarca todo el proceso documental.

Esta misma línea siguen una serie de autores, como veremos a continuación: Lancaster[20], al definirla como *"el proceso de búsqueda en una colección de documentos con el objetivo de identificar documentos relativos a un tema particular*"; Tagge-Stucliffe[21], "*la R.I. es un proceso por el cual se busca un conjunto de documentos para satisfacer las necesidades de información o el interés de grupos o individuos*"; para Codina[22] es "*una operación que consiste en la interpretación de una necesidad de información con el fin de seleccionar los documentos más relevantes capaces de solucionarla*"; como vemos, en esta definición no se incluye la fase documental de preparación del documento, al

[19] A. SPINK and R. M. LOSEE Feedback in Information Retrieval*. Annual Review of Information Science and Technology* 1996, vol 13. p. 31-81.

[20] F. W. LANCASTER. *Information Retrieval Systems: Characteristic, Testing and Evaluation*, 2nd ed. New York: Wiley, 1979.

[21] J.M. TAGUE-STUCLIFFE Some Perspectives on the Evaluation of Information Retrieval Systems *Journal of the American Society for Information Science* 1996, 47 (1) p. 1-3

[22] L. CODINA, Teoría de recuperación de información: modelos fundamentales y aplicaciones a la gestión documental. *Information World en español*. 1995, n 38 p. 18-22

igual que ocurre en la definición de Álvarez Pérez-Ossorio[23]: "*extraer de una colección de documentos aquéllos que se ajustan a las especificaciones determinadas*". Este autor cuando explica las fases, señala que la primera es la traducción a un lenguaje de indización, lo cual nos muestra un concepto un tanto anticuado; según Rijsbergen[24] "*un sistema de recuperación de información no informa, no cambia el estado del conocimiento del usuario en la materia que está preguntando, sólo informa de la existencia o no existencia y del paradero de los documentos relativos a una pregunt*a" ; el concepto de Guerrie[25], es muy similar al Rijsbergen. Para Guerrie, los sistemas proporcionan documentos o citas de ellos, distinguiendo de este modo los sistemas de pregunta-respuesta.

Por otro lado, están las definiciones que engloban, dentro de este concepto, las fases correspondientes a la preparación del documento para la búsqueda, es decir la preparación del almacenamiento y el propio almacenamiento. Estas definiciones son las de Cleverdon, para el que la R.I. es "*toda organización para obtener, almacenar y hacer disponible la información"*[26] y Kowalski, que dice que *"un sistema de R.I. es aquel que es capaz de almacenar, recuperar y mantener información"* [27].

Para nosotros la R.I. es el proceso por el cual, una vez preparado el documento (por lo tanto la fase de preparación del documento está incluida en la R.I.), e identificada la necesidad informativa, se produce una comparación entre

---

[23] J.R. PÉREZ ÁLVAREZ -OSSORIO *Introducción a la información y documentación científica*. Madrid: Alhambra, 1990 p. 59

[24] K. V RIJSBERGEN. Information Retrieval. 2nd prin. London: Butterworths, 1979. [también en línea] http://www.dcs.gla.ac.uk/Keith/Chapter.1 [consultado el 12/03/1999]

[25] B. GUERRIE. Online Information System: Use and operating Characteristics, Limitations and Desing Alternatives. *Information Resources Pres*, 1983.

[26] C. W. CLEVERDON Design and Evaluation of Information System. *Annual review of information Science and Tecnology*. 1971, n 6, p. 42-73

[27] G. KOWALSKI *Information Retrieval System: theory and implementation*. 2nd prin. Boston: Kluwer Academic Publisher, 1998.

ambas para producir unos resultados satisfactorios para el usuario. Pensamos que si perdemos de vista el aspecto de la evaluación los sistemas estarán incompletos, por lo que para nosotros, la R.I. tiene las siguiente cinco fases:

1. Preparación de la información: este tratamiento puede ser mínimo, consistiendo simplemente en un cambio de soporte, o más complejo, como puede ser un sistema de indización por las raíces de las palabras.

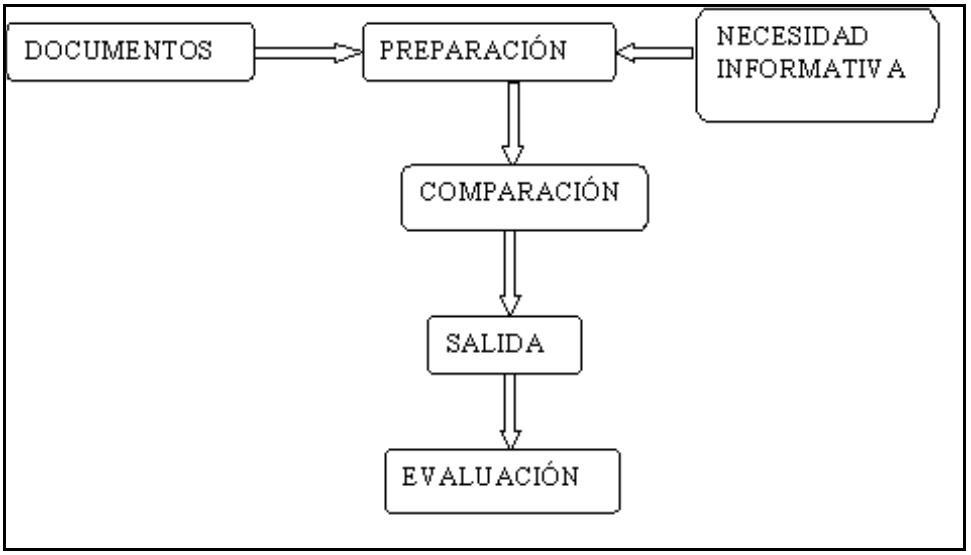
2. Identificación de la necesidad informativa, preparándola para que pueda interrogar a la base de datos. Este proceso será más o menos complejo en función del lenguaje de interrogación que empleemos.

3. Comparación de la pregunta que expresa la necesidad del usuario, con el contenido de la base de datos. Los métodos de comparación varían en función del sistema con el que trabajemos.

4. Salida del resultado de la fase anterior.

5. Evaluación de los resultados.

Estos pasos los podemos ver en el siguiente gráfico:



Dibujo 1 Flujo de pregunta respuesta

2. Distinción entre recuperación de información y recuperación de datos.

Al hablar de R.I. es necesario tener clara la distinción entre recuperación de información y recuperación de datos. Mientras que en el primer caso la información no está estructurada en campos, en el segundo sí lo está, y además se incluye una descripción asociada con cada atributo; por lo tanto, los mecanismos y los resultados de la interrogación entre uno y otro son distintos. La información puede satisfacer la demanda en sí del usuario, o simplemente indicar donde la puede encontrar. La información sería por ejemplo una lista de artículos donde se contiene la información que necesita el usuario. Los datos son lo contenido en esos artículos.

3. Historia de la recuperación de la información.

En los años 40, se comenzó a plantear el problema del almacenamiento y la recuperación de documentos. A finales de los años 50 y principios de los 60, con el incremento exponencial de la producción científica, los métodos tradicionales de almacenamiento y recuperación fueron disminuyendo su efectividad. Al mismo tiempo, se fueron identificando sistemas de información cada vez más operativos. También fue aumentando el número y las áreas de procedencia de los investigadores en el tratamiento de la información. En este proceso, los ordenadores han ido adquiriendo cada vez mayor importancia hasta convertirse hoy en día en herramientas imprescindibles para el almacenamiento, tratamiento y difusión de la información contenida en los diferentes soportes.

Podemos situar los comienzos de la R.I. en los años 50. Es en esta época, y debido a los motivos antes enunciados, cuando se empiezan a dar los experimentos de este campo. Fue Luhn, en estos años, quien sugirió que los

sistemas de recuperación de textos se debían diseñar basándose en la comparación entre los identificadores de contenido del texto y las peticiones de las preguntas[28]. Los primeros sistemas de recuperación, además de la comparación, introdujeron el álgebra de Boole para expandir y limitar las búsquedas. Hoy en día esto sigue estando presente en muchos sistemas de recuperación.

Telfo Saracevic[29], señala que había sido Bradford entre los años 30 y 40 el primero en usar el término relevancia en el contexto de las ciencias de la información, pero será en 1953 en un experimento realizado por la Agencia de Información Técnica de los Servicios de la Armada (*Armed Services Technical Information Agency*: ASTIA) de U.S.A., y El Colegio de Aeronáuticos (*College of Aeronautics)* de Gran Bretaña, sobre recuperación de documentos representados con unitérminos extraídos del título o del resumen, cuando se aplique por primera vez como criterio de evaluación para los sistemas de R.I[30].

En la década siguiente, empiezan a aparecer los experimentos con procesamiento del lenguaje natural[31], y con métodos estadísticos. Dentro de esta línea de investigación destaca Luhn[32], quien usa la frecuencia de aparición de palabras en un documento, para determinar si son suficientemente significativas como para representar el contenido de un documento. Es en esta época también, cuando se empieza a estudiar la frecuencia de coocurrencia de los términos, es decir, determinando el número de veces que aparecen juntos, se establece el grado de relación que hay entre ellos. Sobre este mismo tema investigaron en la década de los 60 y los 70 autores como Maron y Kuhns, Slites, Spark Jones y Robertson

---

[28] H. P. LUNH,. A Statistical approach to mechanized encoding and searching of litery information *IBM Journal of Research and Development* 1957, 1 (4) p. 309-313

[29] T. SARACEVIC, Relevance: A review of a framework for the thinking on the notion in information Science. *Journal of the American Society for Information Science* 1975, 26 (6) p. 321-343

[30] D. ELLIS *New Horizons in Information Retrieval*. London: Library Association, 1990.

[31] Esta parte la desarrollaremos más adelante

[32] H. P. LUNH. (1957) op. cit.

entre otros, como apuntan en sus trabajos Rijsbergen[33] y Hsinchun Chen[34]. En estos años además se comenzó a experimentar en la línea de la estructura de la información, dejando un poco de lado los términos, que era lo que se había estado haciendo hasta el momento. Será Salton uno de los primeros autores que comience a abordar este tema, formulando el sistema de espacio vectorial, y posteriormente el de clustering, como explicaremos más adelante.

A finales de los años 80, se comienzan a usar técnicas basadas en el conocimiento, aquí destacan los esfuerzos realizados en la línea de la creación de los sistemas expertos y del mantenimiento y actualización de la base del conocimiento. Otra de las líneas que adquiere importancia en esta época es la del procesamiento del lenguaje natural, que como apuntó Salton[35], tiene cinco niveles (fonológico, morfológico, léxico, semántico, y pragmático), todos ellos de gran interés para la R.I.

Las últimas tendencias en R.I., combinan el procesamiento del lenguaje natural, con métodos de análisis sintáctico, sistemas de supresión de sufijos (lematización)[36], n-gramas y la inteligencia artificial[37], aplicándose a los sistemas expertos y a redes neuronales. Otra de las investigaciones más novedosas es la de los algoritmos genéticos.

Es importante tener en cuenta que la R.I., desde el punto de vista de las ciencias experimentales, actualmente está en pleno crecimiento y expansión, por

---

[33] K. V. RIJSBERGEN (1979) op.cit

[34] H. CHEN Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms. *Journal of the American Society for Information Science* 1995, 46 (3). p 194-216.

[35] G. SALTON and M. McGILL. *Introduction to Moderm Information Retrieval*. New York: McgrwHill, 1983.

[36] Esta parte la desarrollaremos en el siguiente capítulo de este trabajo.

[37] "Sistemas que muestran las características que pueden asociarse a la inteligencia en lo que se refiere al comportamiento humano: comprensión del lenguaje, aprendizaje, razonamiento, resolución de problemas..." N. AMAT I NOGUERA. *Documentación científica y nuevas tecnologías de la información.* Madrid: Pirámide, 1989.

lo que es de esperar que tanto los temas, como la manera de investigar continúe variando, tan deprisa o más que hasta ahora.

No podemos terminar este breve repaso por la historia de la R.I., sin hacer referencia a las TREC (Text Retrieval Conference), ya que se trata, casi con toda seguridad, de los experimentos más importantes en lo que a este campo de investigación se refiere.

Las TREC son unas conferencias anuales, de origen norteamericano, que tienen como misión el estudio de la evolución, la comparación y evaluación de los sistemas de búsqueda y recuperación de información, trabajando con grandes volúmenes de información. Los participantes, son en su mayoría del área anglosajona por lo que muchas de las operaciones de búsqueda se ajustan a documentos en inglés, aunque también se hacen experimentos en otros idiomas como el francés, el chino o el español. Como indica Korfhage[38], *representan los primeros esfuerzos en experimentos con bases de datos de texto completo, en las que participan distintos grupos con diversas técnicas pero con los mismos documentos y mismos juicios de relevancia*.

Estas conferencias nacieron en 1992 con la iniciativa de las agencias americanas NIST (*National Institute of Stantads Technology*) y el ARPA (antiguo *DARPA, Defense Advanced Research Proyect Agency*).

Las TREC tienen tres objetivos principales[39]:

- Desarrollar métodos de R.I. y distribuir las metodologías de evaluación.

[38] R. R. KORFAGE *Information Storage and Retrieval*. New York: John Wiley and Sons, 1997. p 233

[39] K. LESPINASSE TREC: une conférence pour l´évaluation des systèmes de recherche d´information. *Documentaliste Sciences de l´information*. 1997, 34 (2) p. 74-81

- Ser un foro abierto de discusión entre la industria, los centros de investigación universitarios y los gobiernos.
- Permitir la trasferencia de los equipos de investigación universitaria a los sectores comerciales.

Los métodos de evaluación puestos en práctica, precisión y exhaustividad fundamentalmente, se basan principalmente en la obra de Gerald Salton y Michael McGill[40].

La metodología que se sigue es la de proporcionar una serie de tareas a realizar sobre un conjunto de documentos; las recuperaciones se miden según unos patrones establecidos y posteriormente se comparan los resultados. Las preguntas las elabora personal especializado. La colección de datos es heterogénea: suelen ser de periódicos y publicaciones susceptibles de presentar dificultades (por ejemplo, se mantienen los errores tipográficos). En la actualidad la pertinencia se establece de mediante un *polling*[41], pero los dos primeros años se hizo de manera manual. Los sistemas se comparan mediante la curva de precisión y exhaustividad.

4. Modelos de recuperación de la información.

Resulta difícil establecer una clasificación de los distintos modelos de R.I.; la más conocida es la que estableció Belkin en 1987[42]. Esta clasificación

[40] G. SALTON (1986) op. Cit.

[41] Ver cálculos de la relevancia.

[42] N. J. BELKIN , C. W. BRUCE. Retrieval Tecniques *Annual of Information Science and Tecnology*. 1987, vol 22. p 109-145.

presenta dos problemas: el primero, como señala Frakes[43], es que define las categorías como instrumentos excluyentes y hay que tener en cuenta que los modelos sólo son puros en la teoría; y la segunda es que el artículo únicamente es válido para los modelos desarrollados hasta mediados de los años 80. A partir de esta fecha aproximadamente, empiezan a desarrollarse aplicaciones del Procesamiento del Lenguaje Natural (a partir de aquí P.L.N.) a los sistemas de R.I. En la década de los 90 comienzan las aplicaciones de los sistemas expertos, redes neuronales y algoritmos genéticos como explicaremos a continuación.

4.1. MODELOS TEÓRICOS SEGÚN LA CLASIFICACIÓN DE BELKIN

4.1.1. Coincidencia exacta

4.1.2. Coincidencia parcial

4.1.2.1 Individual

4.1.2.1.1 Basado en estructura

A) Lógica

B) Gráfica

4.1.2.1.2 Basado en características

A) Espacio vectorial

B) Probabilístico

C) Conjuntos borrosos (lógica difusa)

[43] W.B. FRAKES Introduction to Information Storage and Retrieval Systems. En FRAKES, W. B. and BAEZA YATES *Information Retrieval and data Structures and Algorithms*. Mexico: Prentice-Hall Hispanoamericana, 1992

4.1. MODELOS TEÓRICOS SEGÚN LA CLASIFICACIÓN DE BELKIN

Belkin hace la primera distinción de las técnicas de recuperación, en función del conjunto de documentos recuperados. La coincidencia podrá ser total o parcial; en este caso se incluirán también aquellos documentos que tengan coincidencia exacta con los términos que aparecen en la pregunta.

Dentro de los de coincidencia parcial, Belkin distingue entre los que comparan la pregunta con documentos individuales representativos y los que usan

una representación del documento estableciendo conexiones a otros documentos en una red. La recuperación, en estos casos, se basará en las conexiones y en el contenido. Dentro de la categoría de red identificamos las subcategorías, basadas en buscadores: *"cluster", "browsing", "spreading disemination"* [44]. La categoría individual se subdivide en la representación de preguntas, documentos y estructuras. La de preguntas comprende el sistema de índices y el pesado de términos; puede representar entidades más complejas de texto que palabras simples. Las de estructuras, se dividen en representaciones lógicas, donde la pregunta y el documento se representan mediante la lógica formal y las gráficas, donde la pregunta y el documento se representan por grafos, es decir, estructuras compuestas de nodos y arcos conectando esos nodos. Los grafos se pueden crear mediante el procesamiento del lenguaje natural o técnicas estadísticas.

La categoría basada en las características incluye las técnicas basadas en modelos formales. Incluyen el de espacio vectorial, el probabilístico, la teoría de conjuntos borrosos.

4.1.1. Modelos de coincidencia exacta

En este modelo se establece una comparación entre el contenido de un campo y el concepto concreto. Los registros que se recuperan, son aquellos que cumplen las condiciones fijadas con anterioridad. Dentro de estos sistemas se encuentran los booleanos, los de texto completo o las cadenas de búsqueda. Esta técnica de R.I. ha sido muy utilizada debido a su simplicidad.

[44] Optamos por mantener estos tres términos en inglés ya que normalmente aparecen en esta forma en la escasa documentación existente en español. Introducir las traducciones podría inducir a error. En algunas ocasiones *cluster* se ha traducido por "centroide", *browsing* por *"ojeo"*, u *"hojeo"* ya que el término es susceptible de ser traducido de ambas maneras. "*Spreading disemination*" no lo hemos encontrado traducido en ningún trabajo en español.

En los métodos booleanos, los términos se combinan mediante los operadores *AND*, *OR*, *NOT* y paréntesis. En estos sistemas influye el orden de los términos en la pregunta.

Los principales inconvenientes que presenta este modelo son:

- Pérdida de los documentos relevantes cuya representación coincida sólo parcialmente con la pregunta.
- Los documentos no se recuperan en orden de relevancia.
- No se tiene en cuenta la importancia del término dentro del contexto.
- Implica el uso de la formulación lógica, con las complicaciones que ello conlleva.
- Necesita que se empleen los mismos términos en la indización de la pregunta y en la del documento.
- La falta de normalización en la indización induce a error.
- No funciona bien en documentos de texto completo.

Hoy en día estos sistemas están en desuso debido a los inconvenientes mencionados, por lo que se empezaron a crear sistemas híbridos que buscan que la coincidencia no sea del todo exacta, dentro de aquí entran los que hacen truncamientos a la derecha, y algunos modelos de ponderación, donde se solventan algunas de los inconvenientes antes especificados.

4.1.2. Modelos de coincidencia parcial

Hay otros autores que a estas técnicas las denominan *"Best match"* o comparación mejor. Con estos métodos, lo que se hace es buscar aquellos documentos que se ajustan mejor a las condiciones especificadas en la pregunta. Éstas se comparan con documentos o términos de índice. Los documentos y las preguntas pueden ser indizados de manera manual o automatizada, con palabras simples, raíces, o conceptos que pueden llevar asociado o no un peso. En función de estos matices tendremos distintas técnicas de recuperación. Supone un avance respecto a los sistemas de coincidencia exacta.

4.1.2.1 Técnicas de coincidencia parcial individual

En esta categoría, tanto la pregunta como el documento se representan mediante estructuras más complicadas que un simple conjunto de términos.

4.1.2.1.1 Técnicas basadas en la estructura

A) El modelo lógico**:** teóricamente es posible representar información contenida en el texto de los documentos como frases en lógica formal. A medida que las frases sean más complejas, la representación será más complicada. Dando una representación lógica al contenido del documento y la misma lógica a la pregunta, por inferencia, y usando las normas asociadas a la lógica, se interroga a la base de datos. Esto ha sido estudiado por Charman y Mcdermott, Walker y Hobs y Simmons[45], entre otros. El principal problema es la traducción del texto a la lógica. En los experimentos realizados hasta 1985, se hacía de manera manual.

[45] C.F N. J. BELKIN (1987) op.cit.

Rijsbergen[46] ha propuesto un sistema para la R.I. basado en la lógica. Describe la recuperación como un proceso de determinación de una pregunta (expresada en lógica). En la mayoría de los casos esta inferencia no puede hacerse directamente porque se perdería información del documento, por lo que la deducción es incierta.

B) Gráfica**:** La principal característica es una representación con grafos, un conjunto de nodos y arcos que conectan estos nodos. Un ejemplo específico de esto son las redes semánticas y estructuras estudiadas en el procesamiento del lenguaje natural. Las estructuras más simples pueden ser producidas por métodos estadísticos. Las técnicas de recuperación deben buscar similaridad, es decir, la mayor coincidencia, en las estructuras de grafos de preguntas y documentos. Esta similaridad, se puede usar directamente para determinar si el documento debe ser recuperado y establecer la posición de la recuperación del documento.

4.1.2.2.2Técnicas basadas en las características

A) El modelo de espacio vectorial: Fue estudiado a comienzos de los años 70 por Gerald Salton, y posteriormente investigado por Worn y Raghava[47]. En este sistema, los documentos y las preguntas son vectores de una dimensión, con n elementos en el espacio. Cada elemento corresponde a un término de índice. Los documentos se representan gracias a un conjunto de términos, donde d_i indica la presencia (mediante el valor 1) o la ausencia (valor 0) del término i en el

[46] K. V. RIJSBERGEN (1979) op. cit.

[47] S. K WORNG, M. RAGHAVA *Vector Space model of information retrieval*. Research & Development in Information Retrieval. Cambridge: University Pres, 1984. Citado por ARENAS ALEGRÍA, L.. *Efectividad y dinamismo en la Recuperación Documental mediante Análisis Cluster*. [Microforma] Tesis Doctoral. Bilbao: Departamento de publicaciones de la Universidad de Deusto, 1991.

documento d. Este modelo de recuperación puede hacerse de manera binaria (indicando su presencia o no), o de manera ponderada, calculando en este caso pesos en función de la importancia que tenga el término en el documento. Con las preguntas se hace la misma operación, q se refiere a la presencia de i en los términos de la pregunta.

Veamos esto en un ejemplo:

Supongamos que tenemos la siguiente demanda informativa: *La evaluación del impacto de la investigación en biblioteconomía y documentación*. Esta información la representamos mediante un vector de n elementos. Para simplificar el ejemplo vamos considerar que el tamaño de n es igual a siete (1 1 1 1 1 0 0), pero en la práctica este número es mucho mayor. Los 1 mancan la presencia de esos términos, los 0 la ausencia.

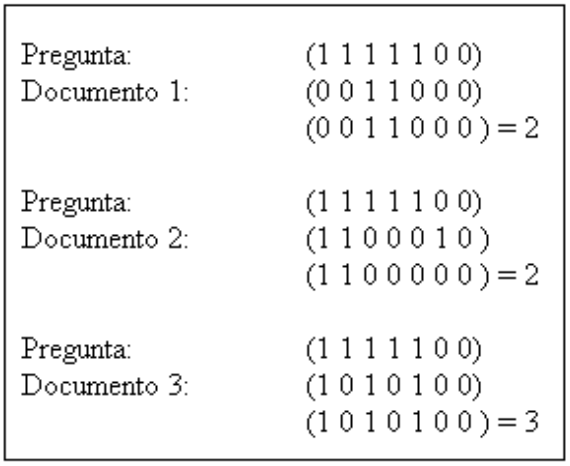
Tenemos un conjunto de documentos, en los cuales aparecen algunos de los términos que tienen la pregunta:

Documento 1: *la investigación en biblioteconomía*

Documento 2: *el impacto de la recuperación en Internet*

Documento 3: *la evaluación de la investigación en documentación*

El sistema compara el documento con la pregunta y ofrece una salida de documentos ordenados en función de la similaridad. Podemos establecer, un umbral por debajo del cual no queremos que se recuperen los documentos. Uno de los sistemas más sencillos consiste en aplicar el sumatorio de los productos, es decir, los números que indican la presencia o ausencia del término en el documento y en la pregunta se multiplican entre sí, y los productos se suman. El resultado de la suma es la similaridad. Veamos esto en un el ejemplo.



Dibujo 2 Modelo vectorial.

En este caso el documento 3 sería recuperado en primer lugar ya que es que tiene una mayor coincidencia entre los términos de la pregunta y el documento.

En el caso de que fuera un sistema ponderado, en lugar de ceros y unos, se pondría el peso del término en el documento, la manera de hacer el cálculo es igual a la anterior.

Pregunta	0,6	0,3	0,9	0,2	0	0	0
Documento 1	0	0,4	0,8	0,1	0,9	0	0
	0	0,12	0,72	0,02	0	0	0=0,86
Pregunta	0,6	0,3	0,9	0,2	0	0	0
Documento 2	0,5	0,5	0,3	0	0,6	0,1	0
	0,3	0,15	0,27	0	0	0	0=0,72
Pregunta	0,6	0,3	0,9	0,2	0	0	0
Documento 3	0,5	0,9	0,9	0	0	0	0
	0,3	0,27	0,81	0	0	0	0=1,38

Dibujo 3 Modelo vectorial ponderado.

Con este sistema lo que hace es poner en relación los objetos del texto. Cuando los vectores de varios documentos son similares, se entiende que los documentos están semánticamente relacionados. Dos vectores tienen algún grado de similaridad, siempre y cuando tengan algún elemento común. A esta relación entre términos se la denomina coocurrencia, y mediante ésta, valoramos la relación de aparición conjunta de entre términos. La coocurrencia sirve para expandir consultas y garantizar que el resultado de la misma es correcto mediante su aparición en los documentos resultantes de la consulta. Esta medida ha de utilizarse con precaución, porque a medida que descendemos perdemos precisión en la recuperación.

Estas técnicas derivan puramente de las aproximaciones basadas en representaciones. El modelo tiene un llamamiento intuitivo y ha formado la base

de gran parte de los sistemas de R.I, incluido el SMART[48] de Salton. Este autor hace una serie de recomendaciones para el proceso de recuperación:

El peso de los términos se calcula usando una combinación normalizada de la frecuencia de aparición de los términos en el documento (tf) y el inverso de la frecuencia de aparición (idf)[49]. Este pesado (tf * idf) se puede calcular para los términos del documento, para cada parte del proceso de recuperación o al indizar el documento. Hay distintas ecuaciones para calcularlo, como veremos a continuación.

- El poder discriminatorio de un término es inversamente proporcional a su frecuencia de aparición en la colección de documentos y es directamente proporcional a su frecuencia de aparición en un documento. El peso de un término depende de:

- El inverso del número de veces que aparece el término en toda la colección (idf)

- El número de veces que aparece el término en ese documento (tf)

- El cálculo de los pesos tf * idf puede calcularse mediante distintas ecuaciones, por ejemplo Belkin[50] dice que es común calcular el peso idf al normalizar la frecuencia de aparición del término en la colección con la frecuencia máxima. Harman, propone dos modos para calcularlo, uno de ellos es calculando el inverso de la frecuencia de aparición del término K en la base de datos, la otra manera es mediante la siguiente ecuación[51]:

---

[48] G. SALTON, Automatic Information Organization and Retrieval. New York: McGraw-Hill, 1968.

[49] *Inverse document frecuency*. Optamos por dejar la abreviatura en inglés ya que en la literatura consultada en español está muy extendido, e introducir la abreviatura traducida podría crear confusión.

[50] N. J. BELKIN (1987) op. Cit.

[51] D. HARMAN How effective is sufixing*?* J*ournal of the American Society for Information Science* 42 (1) 1991 p. 7-15

$$\text{idf} = \text{Log}_2 \frac{N}{NumD_K} + 1$$

Ecuación 1 Cálculo idf. Harman

Donde N es el número de documentos en la base de datos

$NumD_K$: es el número de documentos en la colección que contiene al menos una vez el término K.

Salton[52] calcula el idf mediante otra ecuación, aunque muy parecida a la de Harman.

$$\text{idf} = \log \frac{n}{df_i}$$

Ecuación 2 Cálculo del idf. Salton

Donde df_i: el número de documentos en una colección de n documentos en la que término t aparece.

Sparck Jones[53] para el cálculo del idf propone las siguientes ecuaciones:

$$\text{idf} = \log_2 \frac{N}{n_i} + 1$$

Ecuación 3 Cálculo del idf. Spark Jones (1)

[52] SALTON G. (1989) op. cit 280

[53] K. SPARK JONES. A Statistical interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation* 1972 28 (1) 11-20

Donde N es el número total de documentos en la colección

n_i = el número total de aparición del término i en la colección

La otra ecuación propuesta por esta autora es:

$$\text{idf} = \log_2 \frac{max}{n_i} + 1$$

Ecuación 4 Cálculo del idf. Spark Jones (2)

Donde max_n es la frecuencia máxima de un término en una colección[54]

Los documentos se ordenan en orden decreciente respecto a la similaridad de la pregunta como medida del coeficiente de correlación (intuitivamente, la recuperación de aquellos documentos contenidos en el espacio vectorial de la pregunta). Para calcular la similaridad hay diversas ecuaciones, como vemos la siguiente tabla[55].

[54] K. SPARK JONES. Experiment in Relevance weighting of Search Term. *Information Processing and Management* 1972 15 (3) 133-144.

[55] G. SALTON Automatic text Procesing: the transformation, analysis and retrieval of information by computer. Massachusset: Addison-Wesley, 1989 p.318

Medidas de similaridad	Evaluación de Vectores Binarios	Evaluación de vectores ponderados
Producto entre vectores	$\lvert X \cap Y \rvert$	$\sum_{i=1}^{t} x_i \cdot y_i$
Coeficiénte de Dice	$2\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$	$\frac{2\sum_{i=1}^{t} x_i \cdot y_i}{\sum_{i=1}^{t} x_{i\,i}^2 + \sum_{i=1}^{t} y_i^2}$
Coeficiente del coseno	$2\frac{\lvert X \cap Y \rvert}{\lvert X \rvert^{1/2} \bullet \lvert Y \rvert^{1/2}}$	$\frac{\sum_{i=1}^{t} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{t} x_i^2 \bullet \sum_{i=1}^{t} y_i^2}}$
Coeficiente de Jaccard	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert - \lvert X \cap Y \rvert}$	$\frac{\sum_{i=1}^{t} x_i y_i}{\sum_{i=1}^{t} x_i^2 + \sum_{i=1}^{t} y_i^2 - \sum_{i=1}^{t} x_i y_i}$

Ecuación 5 Similaridad Salton

Donde X=(x1 , x2, x3,... xn)

|X|= número de términos en X

|X∩ Y|= número de términos que aparecen juntos en X e Y

Las preguntas no se comparan directamente con cada documento de la colección para producir una posición, sino que se usa generalmente un fichero invertido, que excluye los documentos que no tienen en común los términos con la pregunta. Finalmente, aunque el pesado de términos se hace en los pasos 1 y 2 con frecuencia, también puede hacerse durante la recuperación.

Una importante extensión de las técnicas de recuperación basadas en el modelo de espacio vectorial coincide, en parte, con la estructura basada en categorías, porque usa preguntas estructuradas (al igual que en la recuperación booleana), es decir, que la pregunta se formula con términos de índice y operadores boolenos (AND, OR, NOT). Cuando expandimos una consulta, se obtiene un nuevo vector de consulta que contendrá términos de la consulta original, más algunos otros procedentes de los primeros documentos relevantes recuperados, de manera que al hacer la consulta obtendremos más documentos que se ajusten a lo especificado. Esto se conoce como "realimentación de consultas" o "realimentación de la relevancia"[56].

Las ventajas del modelo de espacio vectorial son :

- Permite la incorporación de pesos tanto en la pregunta como en el documento.
- Ofrece una salida ordenada en función de la similaridad.
- El tamaño del conjunto recuperado se puede adaptar a las necesidades del usuario.

B) **El modelo probabilístico**: Este modelo ha sido estudiado por Salton[57], entre otros autores. El objetivo de este método es estimar la probabilidad de la relevancia en un documento para un usuario con respecto a una pregunta dada, de ahí surge la dificultad del modelo, puesto que la relevancia, como explicaremos más adelante, aunque trata de ser un valor objetivo, tiene un fuerte

[56] Algunos autores hablan de *retroalimentación de la relevancia*, con este mismo sentido

[57] G. SALTON. On the relationship between theoretical Retrieval Models. EGGHE, L. ROUSSEAU, R. (ed) *Infometrics* 87/88. Amsterdam: Elsevier, 1988 p. 263-270

componente subjetivo. Fue propuesto por Maron y Kuhn[58] en los años 60, desarrollado por Robertson y Spark Jones en la década siguiente[59] y estudiado por diversos autores en las décadas de los 80 y 90[60]. Los antecedentes de este modelo están en los estudios que Luhn[61], y Zipf[62] realizaron en torno a los años 50. El primero usaba la frecuencia de aparición de palabras en un documento para determinar si eran suficientemente significativas para representar el contenido o las características del mismo, por lo tanto la frecuencia de aparición de esa palabra en el cuerpo del documento podía ser también utilizada para indicar el grado de significación. Esto proporcionó un simple esquema de pesado de palabras clave, en cada lista y hacía posible la representación de un documento en la forma de *"pesado de palabras clave-descripción"*

Este modelo es parecido al de espacio vectorial, aunque se diferencia de éste en que aquí estimamos la relevancia de un término en un documento en función de su frecuencia de aparición.

Cada combinación documento-término tiene un peso entre 0 y 1 que nos aporta la información del valor que el término tiene en el documento. El valor de este peso se calcula multiplicando la frecuencia del término en el documento (tf) por el número de veces que ese término aparece en la base de datos:

---

[58] M.E. MARONS and J.L. KUHNS On Relevance, Probabilistic Indexing and Information Retrieval. Journal Association for Computing Machinery, 7 (3) p. 216-44 citado por D. HARMAN. Ranking algorithms...

[59] ROBERTSON and K. SPARK JONES. Relevance Weighting on Search Term. *Journal of the American Society for Information Science* 1976 27 (3) 129-146

[60] R. R. KORFAGE (1997) op. cit.

[61] H. P LUNH. (1957) op. cit.

[62] H. P. ZIPF. *Human behaviour and the Principle or least effort* Addison Wesley. Massachusett: Cambridge, 1949

La ecuación que propone Belkin[63] es:

$$\boxed{\mathrm{tf} = \sum d_i q_i}$$

Ecuación 6 Modelo probabilístico. Belkin

Donde q_i es:

$$\boxed{q_i = \log pr_1 \frac{(1 - pnr_i)}{1 - pr_i}}$$

Ecuación 7 Modelos probabilístico (q_i). Belkin

pr_i es la probabilidad de que el término i aparezca en documentos relevantes

pnr_i es la probabilidad de que término i aparezca en el conjunto de documentos no relevantes.

La principal atracción de este modelo es la información que podemos extraer de los términos y las características de la aparición conjunta de varios de ellos, como por ejemplo la coocurrencia entre términos, la relación de términos que indican que son derivados, la existencia de redes semánticas; también podemos estudiar la aproximación a la inteligencia artificial, el conocimiento histórico sobre cómo ciertos términos han sido recuperados anteriormente como información relevante en respuesta a necesidades similares de información; información sobre el significado de los términos y la relación de los términos derivado de diccionarios y tesauros, la aparición de la distribución de términos en ciertas partes de la colección...

La dificultad de la aproximación probabilística es que la mayor parte de la información concreta sobre la dependencia de los términos y la caracterización de los mismos no suele estar disponible y la distribución entre información relevante y no relevante no se conoce, por lo que la información correcta de

[63] N.J. BELKIN (1987) p. 117

relevancia de hecho no es posible conocerla. En la práctica, esto hace necesario contar con un modelo probabilístico más simplificado que en realidad, no da más información que el modelo de espacio vectorial

El más conocido de los modelos probabilísticos simples, se basa en asumir que los términos son asignados de manera independiente, tanto a la relevancia y no relevancia de los documentos en valores binarios de indización. En este caso, por tanto, el término de información dependiente y el pesado de los términos son independientes. Por debajo de la independencia del término, se asume la probabilidad de relevancia o no relevancia de un documento. Convierte el producto de las probabilidades de relevancia/no relevancia de los términos individuales.

La simple aproximación probabilística, permite por tanto, la mejora de la asignación del peso al término de la pregunta al excluir la frecuencia del término y la normalización de la longitud en el documento, factores comúnmente usados en el sistema de procesado de vectores. Esto puede explicar el hecho de que el simple sistema probabilístico no se haya encontrado especialmente efectivo en la práctica.

Se han hecho algunos intentos para incluir el factor frecuencia de término (tf) en el modelo probabilístico para interpretar la importancia del factor de un término dado en un documento como una probabilidad de que sea estimado por la frecuencia de aparición de un término en un documento individual. En estas circunstancias el modelo se aproxima al modelo de espacio vectorial usando (tf * idf) peso, que sin embargo, ofrece la posibilidad de tener en cuenta alguna información de términos dependientes, que podrían convertirse en posibles. En la práctica, no se ha encontrado todavía un método útil para la estimación de características de grupos de términos dependientes en porciones de documentos relevantes-no relevantes.

C) **El sistema de conjuntos borrosos**: Su principal contribución ha sido la integración de las preguntas booleanas con las técnicas de orden de recuperación.

4.1.2.2 Técnicas de búsqueda en red

4.1.2.2.1 Cluster

Es un grupo de documentos con contenido similar. El análisis del cluster permite la identificación de grupos o clases, es decir, de objetos similares en un espacio multidimensional[64]. La clasificación de los objetos se hace según la descripción de los mismos, y éstos, se pueden describir por una o varias características, o por sus relaciones con los demás objetos. La jerarquía de los cluster, se forma dividiendo los cluster principales en otros más pequeños. El objetivo que persigue, como señalan Jain y Dubes[65], es simplemente encontrar una organización convenientemente válida de los datos. Esta técnica se desarrolló inicialmente para aplicarla a las ciencias de la vida, pero posteriormente ha tenido diversas aplicaciones en muchos otros campos de la ciencia.

Salton, en el SMART, aplica esta técnica[66], estableciendo una jerarquía que se forma dividiendo el documento en unos cuantos cluster, que a su vez se dividen en otros más pequeños y así sucesivamente. La búsqueda se desarrolla de arriba hacia abajo, comparando (usando la medida de similaridad) la pregunta al cluster, de lo genérico a lo específico buscando el mejor. Jardine y Rijsbergen[67]

---

[64] EL-HAMDOUCHI, P. WILLETT. Comparison of hierachic Agglomerative Clustering Methods for document retrieval. *The Computer Journal*. 1989 32, (3) p. 220-227.

[65] A. K JAIN, R. C DUBES *Algorithms for clustering Data*. New Jersey: Prentice Hall, 1988 citado por ARENAS ALEGRÍA, L. op. cit

[66] G. SALTON (1968) op cit.

[67] N. JARDINE, K. V. RIJSBERGEN The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 1971, 7 p. 217-240. Citado por ARENAS ALEGRÍA, L. (1992) op. cit.

también usan la jerarquía en la búsqueda a través de los cluster, pero elaborando los cluster de distinta manera, lo cual hace que se recuperen los documentos en su totalidad sin un orden de salida

Las ventajas que tiene el análisis del cluster frente a otros sistemas son que a la hora de la clasificación, cuando un objeto está descrito por varias características, es difícil de clasificar manualmente; en cambio cuando la lista de características asociadas a cada objeto es larga, el algoritmo de clustering, puede obtener las clases en un tiempo inferior. Otra de las ventajas en este sentido es que normaliza la clasificación, evitando que personas diferentes realicen clasificaciones distintas. Además el cluster proporciona la optimización de almacenamiento, reducción del tiempo de respuesta, y del número de comparaciones o búsquedas que se han de efectuar, para responder a una petición concreta del usuario, ya que sólo los documentos similares serán relevantes para una pregunta.

4.1.2.2.2 Browsing

Algunos investigadores han definido el browsing dentro del contexto del hipertexto. El objetivo de esta técnica es navegar a través de una colección de documentos o de información, buscando la información relevante. Es una estrategia de búsqueda exploratoria[68], especialmente apropiado para problemas mal definidos ya que la exploración va guiando en la nueva formulación de los planteamientos del usuario. Su planteamiento es que si los documentos, términos y la información bibliográfica se representa en un sistema como una red de nodos y conexiones, el usuario puede mirar a través de esta red, y orientándose a través de la información que va encontrando hasta llegar a la información que necesita.

[68] G. MARCHIONINI and B. SHNEIDERMAN Finding fatcs vs. Browsing Knowledge in hypertext systems. IEEE Computer, 1988 21 (3) p. 70-79 Citado por CHEN, H. And A. L. HOUSTON Internet Browsing and Searching: User Evaluation of Category Map Cocep Space Techniques. *Journal of the American Society for Information Science* 1988 49 (7) p. 582-603

En esta técnica se pone menor interés en la formulación de la pregunta inicial, que en otras técnicas y se confía más en la realimentación inmediata, proporcionada por las decisiones del usuario del sistema. Éste es el sistema que se utiliza para la exploración de páginas web.

4.1.2.2.3 Spreading disemination

Esta técnica es similar a la del browsing. Se usa una pregunta para activar partes de una red que describen el contenido del documento y cómo se relacionan unos con otros. En el caso más simple, la pregunta podría activar los nodos de los términos de índice que son conectados a nodos de documentos y a otros términos. En redes de mayor conocimiento, los enlaces y los nodos representan conceptos de materias de dominio y cómo ellos, se relacionan unos con otros, tanto como el documento que contiene esos conceptos. El nodo principal, se desarrolla siempre que por la pregunta, unos nodos se conecten a otros que a su vez son activados (de ahí el nombre de "*activación de la diseminación*"). Tanto por el valor que decrece, como por la activación propagada a través de la red, o las normas, sobre lo razonable de la inferencia que implica el uso de un enlace particular, se usan para controlar la diseminación de la activación. La activación puede converger en nodos de un documento particular de un número de enlaces.

En una red sencilla de documentos y términos, los documentos que tienen un alto nivel de activación, después del primer enlace de los nodos de la pregunta, estarán seguidos de esos documentos que tienen un alto número de términos en común con la pregunta. Si la activación de la diseminación de otros términos conectados a aquellos documentos y después a otros documentos, los documentos recuperados en la segunda fase serán similares a los encontrados por una búsqueda por cluster en redes neuronales. Cuando la activación es redefinida

usando normas de inferencia y más tipos de enlaces, es difícil encontrar los documentos recuperados a aquellos encontrados con técnicas convencionales. La técnica de recuperación en este caso es más parecida a la basada en estructuras.

4.2. MODELOS RELACIONADOS CON EL P. L. N.

Antes de comenzar a hablar del Procesamiento del Lenguaje Natural (en adelante P.L.N.) es importante distinguir entre el lenguaje natural y el controlado o también llamado lenguaje documental. Según M. Pinto Molina el lenguaje natural es *"aquel en el que está escrito el documento [...] tiene la ventaja de ser simple en su utilización"*, en cambio el controlado *"es artificial y se compone de una lista de términos con sus respectivas relaciones"*[69]Mientras que el natural es flexible, el mismo contenido se puede expresar de distintas formas, el controlado es mucho más rígido y tiende a que la misma información sólo se pueda representar de una única manera; de este modo, frente a la ambigüedad del natural, el controlado trata de ser unívoco. Unido a la flexibilidad, está la riqueza de su vocabulario, lo que permite expresar una serie de matices que con el controlado se pierden. Por último, el lenguaje natural, hasta hace poco más de dos décadas ha estado restringido al campo de la comunicación, mientras que el controlado era utilizado para la representación de contenido, esta situación está variando y ahora el lenguaje natural se está aplicando a tareas de representación y recuperación de la información, como veremos más adelante.

[69] M. PINTO MOLINA. *Análisis documental. Fundamentos y procedimientos.* 2ª ed. rev y aum. Madrid: EUDEMA, 1993. p. 214

4.2.1. Definición de P.L.N.

El P.L.N. es la parte de la Inteligencia Artificial que se encarga del estudio y análisis de los aspectos lingüísticos de un texto, a través de programas informáticos, investigando mecanismos efectivos para facilitar la comunicación hombre-máquina, de manera que ésta sea más fluida y menos rígida que con los lenguajes controlados.

En cualquier sistema de P.L.N., se trata de simular el comportamiento lingüístico humano. Para lograr esto, es preciso conocer las estructuras del lenguaje, cómo se forman y combinan las palabras, qué significan, tanto aisladas como en un contexto determinado...etc.

4.2.2. Niveles del P.L.N.

Salton[70] y E. Rich[71] hablan de los cinco niveles del P.L.N. , que se corresponden con la clasificación tradicional que los lingüistas han hecho de las formas de conocimiento de la lengua (fonética, morfología, semántica, sintaxis y pragmática o gramática del contexto):

- Fónico: está relacionado con la velocidad de entendimiento y generación de sistemas. Este nivel aún está por aplicar a la R.I.

- Morfológico: incluye tanto el procesamiento de palabras individuales, como el reconocimiento de porciones de palabras. Dentro de este plano se sitúan aquellos sistemas en los que se pretende indizar en función de la categoría gramatical de las palabras, así como aquellos sistemas basados en la supresión de sufijos (lematización).

---

[70] G. SALTON (1983) op. cit.

[71] E. RICH and K. KNIGHT Natural Language Processing En *Artificial Intelligence*. 2nd ed. New York: Mc Graw-Hill, 1991

- Léxico: a diferencia del anterior, las operaciones se hacen con la totalidad de la palabra. Este plano es de gran importancia para la R.I., ya que lo que los sistemas de recuperación pretenden es el casado entre el significado del documento y la pregunta. Las expresiones idiomáticas, con frecuencia, suponen un problema. El éxito de los sistemas de recuperación basados en la información semántica es bastante limitado a pesar del gran número de técnicas y sistemas que se están desarrollando.

- Sintáctico: Aplicado a la R.I. son los sistemas que indizan los términos teniendo en cuenta la función que desempeñan en la oración y ponderando en función de ella. La sintaxis se centra más en la estructura que en su significado; de este modo, las frases *"Pablo conduce un coche"* y *"un coche es conducido por Pablo"*, significan lo mismo aunque su estructura es distinta; lo que interesa a un sistema de recuperación es que si significan lo mismo, la estructura sintáctica no afecte a la hora de la recuperación. Hay numerosos sistemas que adaptan y aplican este conocimiento sintáctico, a la R.I., tanto al análisis de preguntas como de documentos. El principal problema de los sistemas de este nivel del lenguaje, es la ambigüedad, sobre todo porque en el lenguaje natural se utilizan los referentes y al aislar las frases del contexto, se da la ambigüedad[72].

- Pragmático: la información que aporta este nivel es sobre el contexto donde se desarrollan las palabras.

4.2.3. Historia del P.L.N. aplicado a la R.I.

El P.L.N. tiene diversas aplicaciones, aquí simplemente nos centraremos en la historia referente a la R.I.

[72] L. MORENO BORONAT [et al.] *Introducción al procesamiento del lenguaje natural*. Alicante: Universidad de Alicante, 1999 p. 28

Los primeros experimentos con P.L.N. se dieron a partir de los años 50, en el ámbito de la traducción automática para el Ruso-Inglés. Aquellos trabajos se basaban en las equivalencias de palabras. En esta época también se iniciaron experimentos para comprender el lenguaje en ámbitos muy específicos. En los 60 el P.L.N. consistió en métodos de análisis de palabras clave[73], como el sistema *LUNAR* de Woods, que permitía interrogar en inglés una base de datos de temas espaciales[74]. También se dieron en esta época los trabajos de lingüística de Chomsky en gramática transformacional y la teoría de los lenguajes formales, que supusieron un impulso para el trabajo en esta materia en las décadas siguientes. Así, desde mediados de los 60 hasta comienzos de los 70, el tratamiento de las estructuras sintácticas fue mejorando. Es también en estos años cuando se comienzan a dar las primeras aplicaciones del P.L.N. a la R.I. En este sentido los primeros trabajos buscaban conseguir índices como los elaborados de manera manual, como son los trabajos de Salton[75] y Bely[76]. Ambos estudios usaban tesauros para mostrar las relaciones entre los términos. Salton comparó su sistema con los de análisis estadístico, concluyendo que éstos eran mejores A finales de los 60, con el test de Crandfiel se muestra como las descripciones usando lenguajes controlados, no eran mejores que las que usaban términos simples o incluso lemas. En este sentido se estaba empezando a mostrar el gran potencial del P.L.N. desde prácticamente los comienzos de la investigación en el mismo. A

---

[73] Cf. B.F. GREEN [et al.] Baseball: An Automatic Question Answeree. *Computer and Thought*, 1963 p. 207-216.

Cf. B. RAPHAEL. SIR: A computer program form semantic information retrieval. *Semantic Information Procesing*, 1968 p. 33-145.

[74] L. MORENO BORONAT (1999) op. cit. p. 20

[75] Cf G. SALTON (1968) op. cit

[76] N. BELY [et al.] Procedures d' analyse sémantique appliquées a la documentation scientifique. Paris: Gauthier Villar, 1970. Citado por K. SPARK JONES The role of P.L.N. in Text Retrieval En T. STRZALKOWSKI *Natural language Information Retrieval*. Dordrecht: Kluwer Academic Publisher, 1999

finales de los 60 Lovins[77] crea el primer lematizador para aplicarlo a la R.I., inaugurando así una línea de investigación que hoy en día está teniendo gran repercusión, como mostraremos más adelante.

A partir de los 70 es cuando comienzan las primeras interfaces para bases de datos, en lenguaje natural. En esta misma época Spark Jones y Tait[78] usan el P.L.N. para determinar la estructura de las frases, de la cual se deben extraer los términos compuestos.

En los años 80 se produce, gracias a las teorías de Chomsky sobre "*rección y ligadura*" ("*Governement and Binding*") la unión entre las teorías lingüísticas y los mecanismos de análisis (parsing). En esta línea surgen las gramáticas de estructura sintagmática, gramáticas de léxico, de cláusulas... También continúan las investigaciones en el campo de la traducción automática. En esta década Porter[79] crea su lematizador para el inglés; este trabajo ha sido de gran importancia por la influencia que ha tenido en la década de los 90 en la creación de algoritmos de lematización de idiomas distintos del inglés, como son el francés, árabe, latín, esloveno...[80] Es también a mediados de los 80 cuando Smeaton empieza a investigar en P.L.N., intentando aplicar técnicas de análisis del lenguaje natural de las preguntas de los usuarios para mejorar la recuperación.[81]

---

[77] J.B LOVINS. Development of a Stemming Algorithms *Mechanichal Translations and Computational Linguistics*. 11 (1-2) 1968 p. 22-31

[78] K. SPARK JONES, and J. L. TAIT Automatic Search Term Variant Generation *Journal of Documentation* 1984 40 (1) p 50-66.

[79] M. F. PORTER. An algorithm for Suffix Stripping *Program*, 1980 14 (3) p. 130-137.

[80] Ver capítulo II La lematización.

[81] Cf. F. SMEATON and C.J. VAN RIJSBERGEN. Experiments on Incorporating Syntactic Processing of User Queries into a Document Retrieval Strategy In Proceedings of the 11 th International ACM-SIGIR Conference on Research and Development in Information Retrieval, Grenoble, France June 1988 p. 31-54. Citado por SMEATON, A. F. Using NLP or NLP resources for Information Retrieval En T. STRZALKOWSKI (1999) op. cit.

Finalmente, desde la década de los 90, y gracias a los avances técnicos se están utilizando aprendizajes mediante métodos de inferencia y de aprendizaje automático.

4.2.4. Líneas de investigación aplicadas a la R.I.

El P.L.N., se puede aplicar para mejorar cada fase del ciclo documental[82]. Así en la creación, ayudando en la preparación del conocimiento lingüístico. Dentro de aquí nos encontramos con los sistemas de lenguaje hablado como son los de dictado automático, también con los procesadores de texto, que hoy en día juegan un papel casi imprescindible en la fase de creación de documentos. Los procesadores de texto, comenzaron incluyendo correctores ortográficos simples, y han ido incorporando analizadores básicos que señalan problemas de concordancia o de estilo, diccionarios de sinónimos... en el caso del español queda aún mucho camino por recorrer, ya que los productos comerciales para nuestro idioma son pocos y de no mucha calidad.

En la fase del tratamiento documental, las distintas técnicas se pueden aplicar por ejemplo, para la traducción automática de los documentos. Los sistemas de traducción automática, como señalamos anteriormente, están los orígenes de las aplicaciones de técnicas del P.L.N. Estos sistemas se suelen basar en la transferencia, es decir, se representan las frases de la lengua original y a partir de esa representación se genera el texto en la lengua correspondiente. Dentro de esta fase de tratamiento, podemos encontrar también los sistemas de extracción de información, por ejemplo, los que utilizan los autómatas de estados finitos para obtener información semántica y/o sintáctica que nos ayudará en la interpretación de la estructura de un texto, o en la R.I. En esta fase también nos

---

[82] P. K. HALVORSEN, Overview EN ZAENEN (ed) Document Processing EN *Survey of the State of the Art in Human Language technology*. Oregon: National Science Foundation, 1995 pp 255-258

podemos encontrar los sistemas de etiquetado léxico de textos (*part of speech tagging*) que asignan a cada palabra, la unidad o categoría léxica correspondiente. Esto nos permitirá, por ejemplo, ponderar los términos en función de la categoría gramatical, también servirá para usar la información semántica para mejorar los sistemas de indización, esto se traduce en una mejora de la precisión, reduciendo en la recuperación el número de documentos irrelevantes. La información semántica se extrae del análisis del texto en lugar de los análisis de palabra a palabra. También podemos encontrar en esta fase de tratamiento, los sistemas que crean resúmenes automáticos[83], que antes estaban basados en la información estadística de los textos, ahora se basan en la información lingüística.

En la fase del almacenamiento, las distintas técnicas de P.L.N. se pueden aplicar para reducir el espacio[84].

Finalmente en la fase de recuperación, según indicó Salton[85]en 1983, la utilización del P.L.N. aplicado a la R.I. es beneficioso porque permite usar un lenguaje libre en las formulaciones de la demanda por parte de los usuarios; aporta ventajas de eficiencia y eficacia en las tasas de recuperación, haciendo posible formulaciones más precisas; es fiel reflejo de las necesidades de los usuarios; además el poder emplear expresiones largas, puede darle mayor precisión a las búsquedas, ya que los sistemas convencionales lo que hacen es indizar por términos y luego combinan términos. Esto simplificaría en gran medida las consultas de los usuarios que no necesitarían traducir sus preguntas a lenguajes controlados[86]. En este sentido Dona Harman,[87], afirma que solo una

---

[83] JACOB, P (1995) op. cit.

[84] Ibidem.

[85] G. SALTON (1983) op. cit.

[86] E. D. LIDDY *Enhanced Text Retrieal Using Natural Language Processing*. [en línea] http://www.asis.org/Bulletin/Apr-98/liddy.html [consultado el 19/09/00]

[87] D. HARMAN, P. SCHAÜBLE, and A. SMEATON Document Retrieval. EN *Document Processing* EN *Survey of the State of the Art in Human Language technology*. Oregon: National Science Foundation, 1995 p. 259-265

fracción pequeña de la investigación en R.I. se basa en técnicas de P.L.N. y éstas no son tan importantes como otras técnicas que no aplican el lenguaje natural. Ellen Riloff[88] afirma que las técnicas de P.L.N. y las de R.I. han ido por separado y juntarlas supondrá grandes beneficios. En este sentido en uno de sus trabajos al unirlas han conseguido altos niveles de precisión en la clasificación de textos[89]. Actualmente la aplicación P.L.N. ayuda en la creación de herramientas para gestionar grandes bases de datos y facilita la interrogación al usuario obteniendo de una manera más fácil para él resultados más satisfactorios

4.2.5. Algunas aplicaciones de P.L.N. a la R.I.

Como hemos indicado anteriormente, el P.L.N. tiene diversas aplicaciones dentro de la R.I., aquí simplemente nos centraremos en los n-gramas, en el siguiente capítulo desarrollaremos la lematización.

4.2.5.1. El sistema de n-gramas

Un n-grama es una ventana de n caracteres que se va desplazando a lo largo del texto. En un sistema de n-gramas se define el tamaño de n y se agrupan las secuencias de caracteres en función de este tamaño. Normalmente se suele tener en cuenta el espacio en blanco que precede y el que sigue a las palabras[90]. Así el término "biblioteca", si consideramos que n tiene un tamaño de 3, los trigramas serán: *"_bi", "bib", "ibl", "bli", "lio" "iot", "ote", "tec", "eca" "ca_"*, y los trigramas de una palabra semántica relacionada con ella, tendrán un

---

[88] E. RILOFF and J. LORENZEN Extraction-based text Categorization: Generating domain-Specific Role relationship automatically. En T. S T. STRZALKOWSKI (1999) op. cit p. 167-196

[89] E. RILOFF and W. LEHNER Information Extraction as a Basis for High Precision Text Clasification. ACM Transactions on Information Systems 1996 12 (3) p. 296-333

[90] A.M. ROBERTSON and P. WILLET. Applications of n-grams in textual information system. *Journal of Documentation* 1998 54 (1) p. 48-69

número común de n-gramas, por ejemplo si consideramos la palabra "bibliotecario", los gramas iguales serán: *"_bi"*, *"bib"*, *"ibl"*, *"bli"*, *"lio"* *"iot"*, *"ote"*, *"tec"*, *"eca"* y los distintos: *"car"*, *"ari"*, *"rio"*, *"io_"* .

Los n-gramas se pueden estudiar como una técnica especial de conflación[91] como señala Frakes[92], y por tanto dentro de los sistemas que aplican conocimiento lingüístico, o como una estructura única de un sistema de información. A diferencia de la lematización dónde los árboles determinan el lema de la palabra que representa semánticamente el significado de la palabra, los n-gramas no tienen ningún valor semántico, sino que cada n-grama es tratado como un elemento del vector en el modelo de espacio vectorial, de manera que cuando se ejecuta una consulta se calculan también los n-gramas de dicha consulta y se establece la comparación.

El primer uso que tuvieron fue en criptografía durante la II Guerra Mundial, posteriormente se han usado para la detección y corrección de errores ortográficos y finalmente se está investigando su aplicación a la R.I[93].

A priori las ventajas que presenta este sistema son las siguientes:

- Puede obviar los errores ortográficos y tipográficos[94]. De hecho la incidencia de éstos es menor que en otros sistemas ya que el error afectará a algunos de los n-gramas de la palabra, pero no a la totalidad de la representación.

---

[91] A pesar de que el Diccionario de la Real Academia indica que este término está en desuso, en el lo que se refiere a la R.I. es un término ampliamente utilizado para expresar el acto de agrupar varias palabras bajo una única forma canónica.

[92] W. B. FRAKES (1992) (b) op. cit.

[93] G. KOWALSKI. *Information Retrieval Sytstems: Theory and Implementarion*. 2nd prin. Boston: Kluwer Academic Publisher, 1997.

[94] J. J. POLLOCK, and A. ZAMORA. System Design for Detection and Correction of Spelling error in scientific Scholary Text. *Journal of the American Society for Information Science* 1984 35 (2) p. 104-109

- Es independiente del idioma que utilicemos, la única aplicación que habría que hacer es la del tamaño de los n-gramas porque aunque en un principio se pensaba que los n-gramas eran independientes de la lengua de recuperación, esto no es cierto, ya que mientras que en los trabajos para el inglés, el tamaño utilizado es el de 3 ó 4[95] para el español se obtienen mejores resultados con un tamaño mayor[96] (6 y 7), por lo que previsiblemente la única adaptación de un idioma a otro sea el tamaño.

- No es necesario tener un diccionario de palabras vacías, aunque muchos de los sistemas que trabajan con n-gramas los tienen[97].

- El principal inconveniente que tiene este sistema es que requiere un gran espacio de almacenamiento en disco, aunque dadas las capacidades de los ordenadores, hoy en día esto no supone un gran inconveniente.

4.3. MODELOS RELACIONADOS CON LA INTELIGENCIA ARTIFICIAL

4.3.1. Los sistemas expertos

Los sistemas expertos, también llamados "*sistemas basados en el conocimiento*" o sistemas inteligentes, son programas que ofrecen soluciones a

---

[94] J. J. POLLOCK, and A. ZAMORA. System Design for Detection and Correction of Spelling error in scientific Scholary Text. *Journal of the American Society for Information Science* 1984 35 (2) p. 104-109

[95] W. CAVNAR N-Gram-Based Text Filtering For TREC-2 1993 In HARMAN, D. In HARMAN, D. (Ed) Proceedings of TREC-2: Text Retrieval Pearce & Miller 25 Conference 2, Gaithersburg, MD, 1993. National Institute of Standards and Technology. [También en línea] http://trec.nist.gov/pubs/trec2/t2_proceedings.html [Consultado el 24/07/00].

[96] C. G. FIGUEROLA, R. GÓMEZ, E. LOPEZ DE SAN ROMÁN. Stemming and n-grams in Spanish: an Evaluation of their impact on Information Retrieval. *Journal of Information Science* 2000 26 (6) p. 461-467

[97] W. CAVNAR op. cit.

problemas, simulando el proceso de razonamiento humano mediante la aplicación específica del conocimiento y la inferencia. Se diferencian de los programas convencionales en su evolución asociada a las técnicas de desarrollo y a su estructura interna. Representan el conocimiento y aplican de manera experta para manipular el conocimiento y archivar óptimamente soluciones.

Empezaron a aplicarse en la década de los 50. Entre mediados de los 60 y los 70 tuvieron un alcance muy limitado. En la actualidad están teniendo gran desarrollo.

Un sistema experto tiene las siguientes fases[98]:

- Hechos: declaraciones que relacionan alguno elementos de la realidad con referencias del área específica.
- Reglas de procedimiento: reglas bien definidas e invariables que describen secuencias fundamentales de eventos y relaciones relativas al área.
- Reglas heurísticas: reglas en forma de opinión o reglas empíricas que sugieren procedimientos que se pueden seguir cuando no existen disponibles reglas de procedimiento invariable. Estas reglas son aproximadas y normalmente forman parte de la reflexión humana. Precisamente estas reglas son las que los distinguen de los programas tradicionales.

Su aplicación a los sistemas de R.I., tiene como objetivos:

- Crear modelos de búsqueda de acuerdo con unos parámetros definidos y manipular protocolos, tanto de conexión como de consulta.

[98] D. W. ROLSTON Principios de inteligencia artificial y sistemas expertos. Bogotá: McGraw-Hill, 1990

- Conocer cuántas y cuáles son las bases de datos a consultar en función del tema, y la profundidad de la búsqueda.
- Optimizar los resultados de las búsquedas mediante la utilización de reglas.
- Utilizar el lenguaje natural para consultas (con las ventajas que esto conlleva para el usuario).

En función de los resultados, depurar los perfiles de búsqueda.

4.3.2. Las redes neuronales

Las Redes Neuronales Artificiales o simplemente Redes Neuronales[99] son modelos informáticos inspirados en la estructura a bajo nivel del cerebro. Están inspiradas en los procesos de percepción, recuerdo, clasificación y decisión del conocimiento que realiza el cerebro biológico. Consisten en grandes cantidades de unidades de procesamiento sencillas llamadas neuronas, conectadas por enlaces de varias fuerzas.

Matemáticamente podemos representar una neurona simplificada por un valor (que debe ser superado para que se active) y una lista de enlaces que conectan a otras neuronas y sus fuerzas asociadas. Las señales de entrada a una neurona se multiplican por sus fuerzas ("pesos") asociadas y después se suman. El resultado se llama el nivel de activación de la neurona. Si el nivel de activación supera el valor de la neurona, ésta se activa, y una señal se envía a cada neurona que tiene conectada.

---

[99] Cf. A. MARTÍN VEGA. Las redes de neuronas artificiales en la recuperación de información. Algunas fuentes para su estudio. EN Los profesionales ante el reto del siglo XXI: integración y calidad. *IV Jornadas españolas de documentación automatizada*. *Documat 94* (Gijón, 6,7,y 8 de octubre de 1994) Actas. Oviedo: Universidad, 1994. pp 403-410.

Se diferencian de los sistemas expertos, en que las redes neuronales establecen sus propias reglas en virtud de las adaptaciones o los cambios en los pesos de las conexiones. Su establecimiento supone un considerable avance respecto a los modelos anteriores, y trata de resolver aquellos problemas que no pueden presentarse en términos concretos y exactos para los que la programación convencional ofrece soluciones limitadas. En estos sistemas, la manera de almacenar la información determina los resultados de la recuperación. La memoria de estas redes es distribuida, los pesos de las conexiones son las unidades de memoria de red y el proceso de almacenamiento se reconoce como aprendizaje que puede no ser revisado (donde la propia red se encarga de extraer propiedades y características de las entradas para clasificarlas y ofrecer oportunas respuestas. A priori, es un método viable para la R.I., pero debe estar supervisado por un entrenador, o reforzado. Aún hay pocos trabajos que apliquen las redes neuronales a la R.I.

El browsing, visto con anterioridad en los modelos que Belkin, es una aplicación de redes neuronales a la R.I[100].

4.3.3. Los algoritmos genéticos

Los algoritmos genéticos son programas computacionales cuyo fin es imitar el proceso de selección natural que según la teoría de Darwin rige el curso de la evolución. El proceso de selección natural descrito de una manera sencilla consiste en una población que se multiplica por medio del intercambio de genes, de la nueva generación sólo sobreviven los más capaces de adaptarse a su medio ambiente, para así formar una nueva población "mejor" que la anterior. Este ciclo se repite a través del tiempo y es como han ido evolucionando todas las especies.

[100] Cf. A. LELU From data analysis to neural networks: new propects for efficient browsing through databases. *Journal of Documentation Science*, 1991 17 1-2

Sin embargo hay ocasiones en que se producen mutaciones en los individuos lo cual origina cambios drásticos en las características del individuo y se evita que se llegue a un "estancamiento", en la evolución.

Surgieron para buscar soluciones a problemas que no podían ser solucionados por métodos matemáticos o analíticos, y cuya única forma de resolverlos era a través del método ensayo-error dirigido, es decir ir probando donde se cree que se van a obtener mejores resultados. Al utilizar este método, se dieron cuenta de que este proceso era similar al proceso que seguía la naturaleza, así que intentó copiar su manera de operar, y de este modo se crearon los algoritmos genéticos.

Tienen el siguiente esquema[101]: Para comenzar una población necesitamos primero decidir el número de genes para cada individuo y el total de número de cromosomas, en la población inicial. Los cromosomas se combinan entre sí, generando individuos y proporcionando soluciones a los problemas planteados. En el proceso de selección natural, sobrevivirán aquellos individuos, es decir, las soluciones, con una mayor probabilidad de éxito. La función de adaptación es la encargada de moderar el entorno. En estos algoritmos, las soluciones posibles a un problema se desarrollan de manera paralela, la mejor de esas soluciones se elige y se replica, mientras que la peor se rechaza. Las réplicas de la solución forman la población, en la que se generan nuevas soluciones. La población se adapta a los cambios con las características elegidas en la solución. Algunos individuos pueden ser “peores” que los “padres”, o mejores.

[101] H. CHEN (1995) op. cit.

Los algoritmos genéticos tienen diversas aplicaciones a la R.I[102],[103] por ejemplo en la construcción de preguntas para la realimentación por relevancia[104], indización de documentos[105], compresión de datos[106], recuperación de documentos[107], filtrado de documentos[108]... etc.

En un sistema de recuperación cada gen (bit), representa cierta palabra clave o concepto en el cromosoma (cadena de caracteres). La localización de cierto gen decide la existencia (valor 1) o inexistencia (valor 0) de un concepto. Un cromosoma, representa un documento que contiene múltiples conceptos. La población inicial contiene un conjunto de documentos que son juzgados como relevantes. El objetivo de este tipo de algoritmos es encontrar un conjunto óptimo de documentos en el cual la mejor coincidencia sea la que el investigador necesita.

En la realimentación de la relevancia[109] se asignan pesos a los documentos. En este sentido lo que hacen los algoritmos genéticos es empezar por elegir términos de los usuarios sin pesos y generar variantes de las preguntas asignando

---

[102] O. CORDON, F. DE MOYA, M. C. ZARCO. Breve estudio sobre la aplicación de algoritmos genéticos a la recuperación de Información En M. J. LÓPEZ HUERTAS Y J. C. FERNÁNDEZ MOLINA *La representación y la organización del conocimientos en sus distintas perspectiva, su influencia en la Recuperación de la Información: Actas del IV Congreso ISKO-España EOCONSID'99 22-24 de abril de 1999*. Granada, 1999 p. 179-185

[103] Cf. H. CHEN and J. KIN GANNET: a machine learning approach to document retrieval. *Journal or Management Information Systems* 1995 11 p 7-41

[104] M.P. SMITH and M.SMITH The use of genetic programming to build Boolean queries for text retrieval throught relevance feedback. *Journal of documentation* 23 (6) 1997 p. 423-431

[105] Cf. A. JOHNSON and F. FOTOUHI. Adaptative indexing in very large databases. *Journal of Database Management*. 6 1995 n 6 p. 4-12

[106] S. GRUMBACH And F. TAHI A new challenge for compression algorithms: genetic sequences. *Information Procesing and Management* 1994 30 (6) p. 875-886

[107] J. KULKARNI and H.R. PARSII Information resource matrix for production and intelligent manufacturing using genetic algorithms techniques Computer and Industrial Engeniering 1992 23 p 483-485

[108] R. M. LOSEE Learning syntactic rules and tags with genetic algorithms for information retrieval and filering: an empirical basis for gramatical rules. *Information processing and Management* 1996 32 (2), p. 185-197.

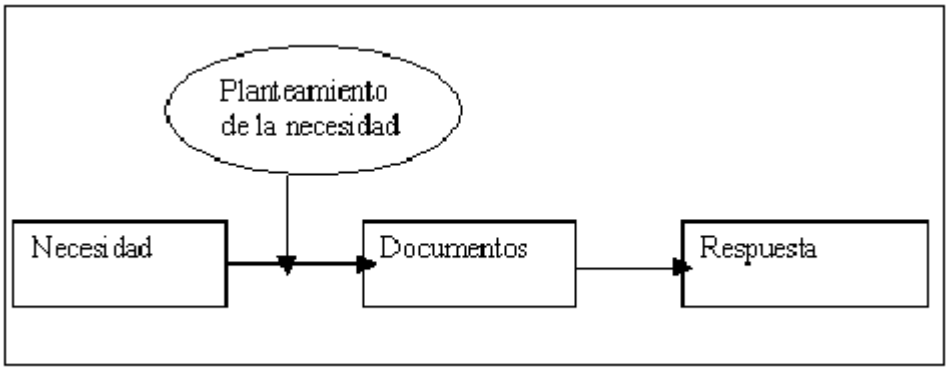
[109] J. YANG, R. R. KORFHAGE Adaptative information retrieval system in vector model. *Symposium on Document Analysis and Information Retrieval*. Las Vegas p. 134-150

pesos de manera aleatoria a los términos. Cada pregunta variante es un vector de pesos correspondiente al vector de los términos de la pregunta. Cada una de estas preguntas nuevas se usa en la base de datos, recuperando el conjunto de los documentos que son entonces evaluados.

5. La evaluación en recuperación de la información.

Para terminar con la revisión bibliográfica de los modelos, vamos a hacer un recorrido por los principales conceptos y medidas de evaluación utilizadas en R.I.

Cuando se produce una necesidad informativa, mediante una estrategia de búsqueda se interroga al conjunto de documentos, con el fin de obtener una respuesta a esa necesidad. Para saber en qué medida la respuesta es satisfactoria o no, tenemos que evaluar los resultados. Desde este punto de vista, la evaluación es la etapa final de la creación de un sistema.



Dibujo 4 Necesidad informativa

Dentro de la evaluación nos podemos encontrar con dos enfoques: el tradicional o algorítmico y el orientado a los usuarios, pero ambos no son

excluyentes sino que como señala Ingwersen[110] son perfectamente complementarios.

El principal problema que presenta la evaluación en R.I. es que hay variedad de interpretaciones en los conceptos, así por ejemplo, el de relevancia que es sobre el que se calculan la mayor parte de las medidas, como veremos a continuación, tiene distintas interpretaciones.

Hay trabajos importantes de revisión bibliográfica sobre las medidas de evaluación, aunque algunos, tal vez un poco antiguos, a pesar de ello, remitimos para una mayor información a Keen (1966[111], 1971[112]), Robetsoon (1969[113]), Tague (1996)[114] y a Harter[115] (1997), para ampliar y obtener una visión más completa de los distintos métodos que se han venido utilizando en el campo de la R.I. Aquí simplemente haremos un pequeño resumen de la historia de la evaluación en R.I. y de las medidas más importantes.

El antecedente de los experimentos de evaluación está en el trabajo realizado por la ASTIA (*Armed Services Technical Information Ag*ency) y el *College of Aeronautic*s sobre la recuperación de documentos representados con

---

[110] P. INGWERSEN *Information Retrieval interaction*. London:Taylor Graham, 1992.

[111] E. M KEEN Measures and Averaging Methods Used in Performance Testing Indexing System. Crandfield, Eng.,Aslib Crandfield Project 1966 citado por LANCASTER, W. F. and WARNER, A. J. *Information Retrieval Tod*ay. Arlington: Information Resources Press, 1993

[112] E. M KEEN Evaluation parameters. En SALTON, G. (ed), *The SMART retrieval system Experimentes in automatic document processing*. New Jersey:Prentice-Hall, 1971 p 74-111citado por YAO , J. J. Measuring Effectiveness Bases on User Preference of Documents. *Journal of the American Society for Information Science* 46, (2), 1995 p. 133-145

[113] S. E. ROBERTSON, The Parametric Description of Retrieval Test *Journal of Documentation*, 25 (2) June 1969 p. 93-107 citado por LANCASTER, W. F. and WARNER, A. J. *Information Retrieval Tod*ay Arlington: Information Resources Press, 1993

[114] J.M. TAGUE-SUCLIFFE (1996) op. cit.

[115] S. P. HARTER, HERT, C. A. Evaluation of Information Retrieval Systems: Approaches, Iusues, and Methods *Annual Review of Information Science and Technology* 1997 *(ARIST)* vol 32, pp. 3-94

unitérminos extraídos del título y el resumen.[116] En este experimento realizado en 1953 fue donde se utilizó por primera vez el concepto de relevancia, aunque éste ya había sido formulado en la década anterior[117].

El primer trabajo de evaluación propiamente dicho fue el desarrollado por Crandfield. La importancia de los trabajos desarrollados por este autor, radica en que fue el primero estableció la metodología de la evaluación y las herramientas que debían emplearse. Estas herramientas están formadas por :

- Una colección de documentos de la que se extraen las preguntas.
- Los juicios de relevancia.
- Medidas de precisión y exhaustividad de cara al análisis de los resultados.

Estas mismas herramientas son las que se utilizan en las TREC.

El trabajo de Crandfiel se divide en dos etapas: Crandfield I (1957-1962) y Crandfield II (1963-1966). En el primero comparó cuatro sistemas de indización y el segundo evaluó 33 tipos de lenguajes de indización con distinta terminología y estructura.

Los experimentos de Crandfield, como señala Olvera[118], “constituyeron una aportación fundamental en el campo de la evaluación ya que se pasa de una

[116] D. ELLIS (1990) op. cit.

[117] T. SARACEVIC (1975) p. 321-343

[118] Mª D. OLVERA LOBO (1999) (a) op. cit.

aproximación especulativa en los diseños de los sistemas de recuperación de información, a una concepción empírica y experimental".

5.1. La relevancia

5.1.1. Concepto de relevancia

Al hablar de la evaluación en R.I. resulta imprescindible explicar, aunque sea mínimamente el concepto de relevancia. Esta medida es importante porque está en la base del resto de las medidas que tradicionalmente se vienen aplicando en R.I., y aunque el concepto se formuló entre los años 30-40, no se utilizó experimentalmente hasta el test de Crandfield; a pesar de ello, aún no está suficientemente estudiada la validez del uso de los juicios de relevancia.

El concepto de relevancia se ha estudiado desde distintos puntos de vista: filosófico, psicológico, semántico, documental. Estos enfoques los podemos resumir en dos tendencias: la relevancia objetiva y la subjetiva. La primera hace hincapié en los sistemas, normalmente define cómo la materia de la información recuperada coincide con la de la pregunta. La subjetiva, según Swanson[119], es la que tiene en cuenta al usuario. Dentro de este enfoque está la relevancia mirada desde el punto de vista situacional, como proponen los trabajos de Schamber, Einsenberg y Nilo[120], y el de Wilson[121]. Así para los primeros, la relevancia se refiere a la utilidad, o potencial uso de los materiales recuperados, con relación a la satisfacción de los objetivos, el interés, el trabajo o los problemas intrínsecos

[119] D. R. SWANSON Subjetive versus objetive relevance in bibliografic retrieval system. *Library Quartely* 1986 56, p. 389-398

[120] L. SCHAMBERG, M.B. EINSEBERG, and M. S. NILO A. re-examination of relevance: toward a dynamic, situational definition Information Procesing and Management, 1990, (6) p. 755-775. Citado por BORLUM, P. and INGWERSEN, P. The development of method for evaluation of interactive Information Retrieval System. *Journal of Documentation* 1997 53 (3) p. 225-250

[121] P. WILSON Situacional relevance Information *Storage and Retrieval* 1973 9 p. 457-469

del usuario. Para Schamber, este enfoque desde el punto de vista del usuario resulta muy interesante[122]. La relevancia subjetiva, por su parte, la estudia desde el punto de vista de la información nueva que consigue un usuario de un documento. Según este concepto, la información conocida no es relevante. Esto ha sido estudiado por Boyce[123]. Hay autores a caballo entre estas dos tendencias, para los que la relevancia tiene un componente objetivo y otro subjetivo. Así Barry[124], determina si un documento es relevante en función de siete criterios (1. Información que contiene un documento; 2 experiencia previa del usuario; 3 creencias y preferencias del usuario; 4 otras informaciones y fuentes; 5 fuentes del documento; 6 documento como entidad física; 7 situación de los usuarios) de los cuales dos son objetivos (1 y 5) y cinco subjetivos.

Harter[125] indica que el principal problema de los estudios sobre los factores que afectan a la relevancia es que se han hecho de manera intuitiva.

Muy ligado al concepto de relevancia está el de pertinencia; con frecuencia se entremezclan y confunden. Según Korfhage[126], **relevancia** es la medida de *cómo una pregunta se ajusta a un documento,* y **pertinencia** *es la medida de cómo un documento se ajusta a una necesidad informativa.* Es decir, según este autor, la diferencia entre uno y otro radica en cómo expresamos la necesidad de información, por lo tanto, a la hora de establecer la relevancia tenemos que tener en cuenta la doble dificultad que lleva implícita la pregunta ya que tiene que ser el reflejo de la necesidad informativa (de ella dependerá la pertinencia) y al mismo tiempo ser adecuada para buscar los documentos, ya que

---

[122] Cf. L. SCHAMBER, Relevance and Information behaviours. *Annual Review of Information Science and Technology (ARIST)* 1994 29 p. 3-48

[123] B. BOYCE Beyond Topically: A two storage view of relevance and retrieval process *Information procesing and Management* 1992 18 p. 105-109

[124] C.L. BARRY User –defined Relevance Criteria: An Exploratory Study *Journal of the American Society for Information Science* 1994 45 (3) p. 149-159

[125] S. P. HARTER Variations in Relevance Assessment and Measurement of Retrieval Effectiveness. *Journal of the American Society for Information Science* 47 (1) 1996 p. 37-49

[126] R. R KORFHAGE (1997) op. cit.

la relevancia va a depender de la formulación concreta de la demanda informativa. A pesar de que Korfhage establece esta distinción, no todos los autores siguen esta línea, sino que algunos los utilizan como sinónimos. En el caso de los trabajos en español muchas veces se han traducido los dos términos indistintamente para referirse a los dos conceptos. Nosotros seguiremos la distinción de Korfhage, de este modo, para nosotros un documento será relevante para una pregunta, siempre que el contenido del documento conteste a ella, independientemente de si la información que tiene el documento resuelve la necesidad informativa o no del que la plantea. La valoración de la pertinencia es mucho más difícil de realizar ya que es el propio usuario el único que sabe si un documento se ajusta a su necesidad o no. Aquí entran en juego los factores que Schamber, Einsebber, Nilo[127], y Wilson[128] llaman relevancia desde el punto de vista situacional o del usuario.

5.1.2. El cálculo de la relevancia

Para calcular la relevancia, lo más habitual es establecer valores binarios: un documento es relevante, es decir, sirve como respuesta a nuestra pregunta, (valor 1) o no sirve (valor 0), aunque también se puede fijar una gradación, como la de Cuadra y Katter[129], que establecen una escala ordinal para medir la relevancia de los documentos. El problema de determinar una escala es que no hay una guía clara para elaborarla. Por ejemplo Keen[130], usa cuatro valores de escala, para dividir del más relevante al menos relevante. Saracevic[131] da tres

[127] L. SCHAMBER (1990) op. cit.

[128] P. WILSON (1973) op. cit.

[129] A. C. CUADRA, and R.V. KATTER, Opening the blok box of "relevance". *Journal of documentation* 1967 23 (4) p 291-303

[130] E. M. KEEN (1971) op. cit.

[131] T. SARACEVIC, [et al.] A study of information seeking and retrieving, background and metodology. *Journal of the American Society for Information Science* , 39 (3) p. 161-176

valores a su escala: relevante, parcialmente relevante y no relevante, pero en la práctica distinguir entre un documento relevante y uno parcialmente relevante es muy difícil.

Existen dos métodos para calcular la relevancia, uno manual y otro conocido como *polling*[132]:

1. Manual: consiste en la exploración de los documentos uno a uno para saber si se adecúan o no como respuesta a una pregunta. Muchas veces establecer la relevancia de un documento a una pregunta determinada resulta difícil y los especialistas no se ponen de acuerdo, por ello, es conveniente que los juicios los haga más de uno, y a ser posible un número impar de especialistas, para evitar la subjetividad. El principal problema que presenta este método, es que en colecciones muy grandes, hay que invertir gran cantidad de tiempo, lo que supone mucho dinero para realizar esta operación y esto no siempre es posible. Además, algunas bases de datos son más especializadas que otras, lo que hace necesario contar con un número mayor o menor de especialistas. Para solventar estos problemas se crean las colecciones experimentales, donde se fija de antemano qué documentos son relevantes para cada pregunta. Estas colecciones suelen tener un tamaño medio y suelen pertenecer a una misma área temática o muy próxima para que no sea necesaria la intervención de muchos especialistas.

Un ejemplo de una colección manual es la de Crandfield. En este caso se buscaron los artículos y se les pidió a los autores que elaboraran preguntas cuya respuesta fuera su artículo y también se les pidió que citaran otros artículos que correspondieran a esa misma pregunta que ellos habían formulado. Con las preguntas y los artículos citados por los autores se elaboró la base de datos y la colección de preguntas.

[132] No hemos encontrado ningún término en español que describa este tipo de cálculo por lo que optamos por mantener el término en inglés.

2. *Polling*: cuando las bases de datos son muy grandes, y no es posible evaluar uno a uno los documentos, para determinar cuáles son los documentos, se recurre al "polling". Lo que se hace es analizar de manera manual un número determinado de documentos recuperados con distintos sistemas, este número suele ser elevado (100 ó 200 documentos) y se corresponde con los primeros documentos recuperados con cada sistema. Este conjunto de documentos es el que de manera manual analizan los expertos, que son los encargados de decir en último término si son relevantes o no. Este sistema asume que la gran mayoría de los documentos relevantes son encontrados, si no por todos los sistemas, sí al menos por alguno de ellos, y los no recuperados pueden considerarse como no relevantes[133]. De esta manera no es necesario evaluar toda la base de datos, pero aún así el sistema es fiable ya que el número de documentos que se suele examinar es elevado. Este sistema es el que se viene utilizando en las TREC desde 1994[134].

5.2. Principales medidas de evaluación

Una vez definido el concepto de relevancia y relacionando éste con si un documento es recuperado o no, podemos establecer una serie de medidas que nos servirán para evaluar los sistemas de recuperación.

Los documentos recuperados pueden ser los relevantes, es decir, los correctos, los que se adecuan a la pregunta. O no relevantes para dicha pregunta, éstos son los que introducen ruido. Respecto a los documentos no recuperados, puede que éstos sean relevantes a la pregunta, es este caso perderíamos la

[133] G. KOWALSKI (1997) op. cit. p. 233-245

[134] D. HARMAN Overview of the Third Text Retrieval Conference (TREC-3) [en línea] <trec.nist.gov/pubs/trec3/t3_proceedings.html > [consultado el 10/01/01]

información que contienen, o no relevantes, que son los que el sistema rechaza con buen criterio.

Esto lo podemos expresar en una tabla de doble entrada:

	RELEVANTE	**NO RELEVANTE**	
RECUPERADO	a (correctos)	b (ruido)	a+b
NO RECUPERADO	c (perdidos)	d (rechazados correctamente)	c+d
	a+c	b+d	a+b+c+d

Tabla 1 Distribución de documentos

En función de estos cuatro parámetros se establecen las medidas de evaluación.

5.2.1. La precisión

Este concepto fue definido por Kent[135] en 1955, como *factor de pertinencia*. Hay otros autores que se refieren a él, como *ratio de aceptación*. Para Salton[136], la precisión es la proporción de material recuperado realmente relevante, del total de los documentos recuperados. A esta definición Frakes[137] añade que el resultado de esta operación está entre 0 y 1. Así, la recuperación

[135] A. KENT et al. Machine literature searching. VIII. Operational Criteria for Designing Information Retrieval Systems American Documentation Abril 1955 6 (2) p. 93-101

[136] G. SALTON (1983) op cit.

[137] W. B. FRAKES (1992) op. cit.

perfecta es en la que únicamente se recuperan los documentos relevantes y por lo tanto tiene un valor de 1. Para Pérez Álvarez Ossorio[138], esta medida se expresa en porcentajes. Según Kowalski[139] es la media de precisión de la búsqueda, evalúa directamente la correlación de la pregunta con la base de datos e indirectamente sirve para ver cómo es de completo el algoritmo de indización. Si el algoritmo de indización tiende a generalizar teniendo un umbral alto en los términos de índice o al usar los conceptos de indización, entonces la precisión es baja, no importa cómo sea el algoritmo de similaridad entre la pregunta y el índice. Nosotros basaremos nuestros cálculos en la ecuación generalmente admitida[140]:

$$\text{Precisión} = \frac{\text{Documentos relevantes recuperados}}{\text{Documentos recuperados}}$$

Ecuación 8 Precisión. Salton

Esta medida está relacionada con dos conceptos, el de ruido y el de silencio informativo. De este modo, cuanto más se acerque el valor de la precisión a 0, mayor será el número de documentos recuperados que no le sirvan al usuario y por lo tanto el ruido que encontrará será mayor.

Su representación gráfica se hace marcando en el eje de las *x* el número de documentos y en las de las *y*, la precisión de 0 a 1, de modo que los sistemas más precisos son aquellos que en su gráfica describen una curva con valores altos al principio y que van decreciendo. Comparando las distintas curvas de los sistemas, podemos hacernos una idea clara de cuáles son más precisos.

[138] J. R. PÉREZ ÁLVAREZ OSSORIO (1990) op. cit.

[139] G. KOWALSKI (1997) op. cit.

[140] G. SALTON (1983) op cit.

5.2.2. La exhaustividad

Junto con la precisión es el concepto más utilizado en la evaluación de los sistemas de recuperación.

Muchos autores, por influencia del término inglés la denominan *"recall"* o *"rellamada"*. Es la proporción de material relevante recuperado, del total de los documentos que son relevantes en la base de datos, independientemente de que éstos, se recuperen o no. Esta medida es inversamente proporcional a la precisión. Fue formulada, al igual que la de precisión por Kent[141] en 1955, con el nombre de *factor de exhaustividad*. Años más tarde, Swet[142] la llamó *probabilidad condicional de un item*, y Goffman y Newil[143] la denominaron *sensibilidad* (sensibility).

La ecuación tomada de Salton es[144]:

$$\text{Exhaustividad} = \frac{\text{Documentos relevantes recuperados}}{\text{Documentos relevantes}}$$

Ecuación 9 Exhaustividad. Salton

Si el resultado de este cálculo tiene como valor 1, tendremos la exhaustividad máxima, ya que hemos encontrado todo lo relevante que había en la base de datos, por lo tanto no tendremos ni ruido ni silencio informativo: la recuperación será perfecta.

Para alcanzar una exhaustividad alta, es necesario utilizar como índice términos generales de alta frecuencia, es decir, que aparezcan en muchos

---

[141] A. KENT (1955) op. cit.

[142] J. A. SWETS Information retrieval Systems *Science*, 141 (3577): July 1963 p. 245-250 citado por LANCASTER 1993 op. cit.

[143] GOFFMAN and NEWILL *Methogology for test and evaluation of information retrieval systems*. Information Storage and Retrieval (1964) 3 p. 19-25

[144] G. SALTON (1983) op. cit.

documentos de la colección. Para alcanzar una precisión alta, es necesario que los términos aparezcan con frecuencia alta, pero en pocos documentos y con nula en el resto. De este modo recuperaremos sólo los relevantes y rechazaremos los no relevantes. Aunque para el usuario la situación ideal es una precisión y exhaustividad alta, lo que Cooper denomina *utilidad teórica*[145], y esto es imposible. Como puede deducirse de esto, ambas medidas son contrarias por lo que habrá que llegar a un equilibrio entre ambas.

Al igual que la precisión también podemos representarla gráficamente, para ello en el eje de las *x* marcamos el número de documentos y en el de las *y* la exhaustividad de 0 a 1. El comportamiento normal de esta gráfica, es que la curva vaya aumentando. Los sistemas serán más exhaustivos cuando alcancen al principio valores altos (próximos a 1), y después vayan disminuyendo.

Korfhage señala principalmente dos objeciones a los sistemas que se basan en la precisión y en la exhaustividad. El primero de ellos es que mientras que la precisión se puede determinar, la exhaustividad no, ya que para calcularla necesitamos previamente el número de documentos relevantes. Para el cálculo de la exhaustividad se suelen utilizar métodos estadísticos por lo generalmente es un método aproximado. El segundo de los puntos que señala Korfhage es que la exhaustividad y la precisión son igualmente significativas para los usuarios. Mientras que unos prefieren una precisión mayor, otros prefieren una exhaustividad más alta, y ambas cosas es imposible tenerlas al mismo tiempo.

L.T. Su[146] afirma que ni la precisión ni la exhaustividad son factores altamente significativos en la determinación de cómo los sistemas de recuperación

---

[145] S. COOPER The Paradoxical Role of Unesamined Documents in the Evaluation of Retrieval Effectiveness. Information Processing and Management 1976 12 p. 367-375 Citado por C. A. HERT Understading information retrieval interaction: theoretical and practical implications. Greenwich: Ablex Publishing Corporation, 1997.

[146] L. T. SU The Relevance of Recall and Precision in User Evaluation. 1994 *Journal of the American Society for Information Science* 1994 45 (3) p. 207-217

satisfacen a los usuarios, y si no son importantes para los usuarios, entonces son medidas pobres para los sistemas de recuperación.

A pesar de las opiniones de estos autores estas medidas son las que más se utilizan para evaluar los sistemas de recuperación.

Podemos poner en relación la precisión y la exhaustividad en una sola gráfica, para ello, marcamos en el eje de las *x* la exhaustividad y para cada valor de ésta marcamos en el de las *y* el valor de precisión que le corresponde. De este modo relacionamos ambas medidas. Viendo la trayectoria de las curvas, podemos ver cómo se relacionan las medidas de precisión y exhaustividad en cada sistema y cuál es el más efectivo. La principal cualidad de esta gráfica, es su claridad para establecer comparaciones[147]. El sistema más efectivo es aquel cuyo punto en el gráfico dista más del origen.

En 1983 Salton y MacGill, sugirieron un método para la evaluación del sistema proponiendo salidas ordenadas de los documentos en las respuestas. De este modo, la precisión y la exhaustividad dependían del valor de corte, es decir, del punto a partir del cual se considera que al usuario ya no le interesan los documentos. Este criterio Blair[148] lo denomina "*umbral de futilidad*". La precisión y la exhaustividad se calcula para cada posición en la lista de documentos recuperados.

[147] LESPINASSE, K. (1997) op. cit.

[148] BLAIR Searching bases in large interactive document retrieval systems *Journal of the American Society for Information Science* 1980 (31) 4 p. 271-277

5.2.3. Medidas complementarias para la precisión y la exhaustividad

Existen otra serie de medidas complementarias a la precisión y a la exhaustividad.

5.2.3.1. Complemento del ratio de precisión

También se le denomina "*factor de ruido*". Consiste en los documentos no relevantes recuperados partido por los recuperados.

$$\text{Complemento del ratio de precisión} = \frac{\text{Documentos no relevantes recuperados}}{\text{Documentos recuperados}}$$

Ecuación 10 Complemento del ratio de precisión

5.2.3.2. Complemento del ratio de exhaustividad

Su ecuación se calcula dividiendo los documentos relevantes no recuperados entre el total de los documentos relevantes.

El primero en formularlo fue Swets 1963 que lo denominó probabilidad condicional de una pérdida. En 1965 Fairthore[149] lo denomino ratio del esnobismo. ("snobbery ratio")

$$\text{Complemento de exhaustividad} = \frac{\text{Documentos no relevantes recuperados}}{\text{Documentos relevantes}}$$

Ecuación 11 Complemento del ratio de exhaustividad

[149] FAIRTHORE Notas no publicadas

5.2.3.3. El índice de irrelevancia

Este índice[150] se obtiene de dividir los documentos recuperados no relevantes a la pregunta entre el total de los documentos contenidos en la colección. Como muchas de las medidas anteriores fue formulada en primer lugar por Swets en 1963, que se refirió a él como *probabilidad condicional de bajada falsa* (*conditional probability of false dr*op). Cleverdom, Mills and Keen[151]la llamaron posteriormente *fallout*. También ha sido denominada "*desechado*" (*discard*).

$$\text{El índice de irrelevancia} = \frac{\text{Documentos no relevantes no recuperados}}{\text{Documentos no relevantes}}$$

Ecuación 12 Índice de irrelevancia

Según Kowalski[152] con esta medida podemos definir con qué efectividad está actuando un sistema de recuperación. Esta medida es el inverso de la exhaustividad y nunca nos encontraremos con un resultado de 0/0, a menos que todos los documentos sean relevantes para la búsqueda.

5.2.3.4. Complemento del índice de irrelevancia

Swets en 1963, lo denominó *"probabilidad condicional de una correcta respuesta negativa*" (condictional probability of a correct rejection). Goffman and

[150] K-I. YU, P. SCHEIBE, F. NORDBY, The FDF Query Generation Workbench. Trec-3 [en línea] TREC3 http://trec.nist.gov/pubs/trec3/t3_proceedings.html [consultado el 10/01/01]

[151] C. W. CLEVERDON, J. MILLS and E. M. KEEN Factors Determining the perfomance indexing Systems. ASLIB Crandfiel proyect 1966

[152] G. KOWALSKI 1997 op. cit.

Newill la llamaron *"especificidad"*. Se calcula dividiendo los documentos no relevantes no recuperados entre el total de los documentos no relevantes:

$$\text{Complemento del índice de irrelevancia} = \frac{\text{Documentos relevantes}}{\text{Total de documentos}}$$

Ecuación 13 Complemento del índice de irrelevancia

Existen otra serie de medidas que ponen en relación las medidas anteriores como son:

5.2.3.4.1. Generalidad

La generalidad[153] sirve para calcular la densidad de documentos relevantes. Se calcula dividiendo los documentos relevantes entre el total de los documentos de la base. Esta medida se relaciona directamente con la pregunta[154].

$$\text{Generalidad} = \frac{\text{Documentos recuperados}}{\text{Número de documentos}}$$

Ecuación 14 Generalidad

La precisión, la exhaustividad, el índice de irrelevancia y la generalidad se relacionan mediante la siguiente ecuación:

$$\frac{P}{Ir} = \frac{P/(1-P)}{G/(1-G)}$$

Ecuación 15 Relación entre precisión, exhaustividad, y generalidad

[153] R. R. KORFHAGE (1997) op. cit.

[154] J. J. YAO (1995) op. cit.

Donde P/(1-P) es el ratio de los documentos relevantes recuperados partido el de los no relevantes recuperados.

G/(1-G) es el ratio de los documentos relevantes en la colección partido los documentos no relevantes en la colección.

P/ir es la ejecución de la recuperación en los documentos relevantes entre la ejecución de la recuperación en los documentos no relevantes. Es deseable tener el primero de los dos alto.

5.2.3.4.2. La medida de F

Sirve para corregir el error de la *Distancia*, en los casos en el E y P se compensan. Su ecuación es:

$$F_b = \frac{(B^2 + 1) \bullet P \bullet R}{B^2 \bullet P + R}$$

Ecuación 16 Medida de F

Donde *B* es un valor preestablecido, teniendo en cuenta que si *B* es igual a uno, estamos dando la misma importancia a P que a E, si *B* mayor que uno de damos más importancia a E y si es menor de damos más importancia a P.

5.2.3.5. La longitud de búsqueda esperada

Es el número de documentos no buscados que el usuario puede esperar examinar antes de encontrar el número de documentos deseados[155].

[155] W. S COOPER Expected Search Lenght: A single measures of retrieval effectiveness based on the weak ordering action retrieval systems. American Documentation 1968; 19 p. 30-41 Citado por HARTER (1997)

5.2.4. Medidas relacionadas con el usuario

A pesar de tener todas estas medidas no podemos perder la perspectiva del usuario ya que es la razón de ser de la existencia del sistema. La efectividad de un sistema es una medida ajena al propio sistema que relaciona la satisfacción del usuario con la salida que el sistema proporciona. Medir la satisfacción del usuario resulta muy importante, pero resulta complicado y es menos objetivo que las medidas vistas anteriormente, por eso estas medidas se han ido dejando de lado.

Las medidas orientadas al usuario fueron propuestas por Keen en 1971[156] y son:

- **Ratio de cobertura**: es la proporción de documentos relevantes conocidos por el usuario que son actualmente recuperados.
- **Ratio de novedad**: proporción de documentos relevantes recuperados que previamente son conocidos por el usuario
- **Exhaustividad relativa**: ratio de documentos relevantes recuperados, examinados por el usuario, partido por el número de documentos que el usuario quiere examinar.

Supongamos que el usuario conoce 15 documentos relevantes, y el sistema recupera 10 relevantes, incluyendo 4 documentos que son conocidos por el usuario. El ratio de cobertura sería 4/15 es decir 26,6%. De aquí el usuario puede inferir que hay aproximadamente 38 documentos relevantes, aproximadamente cuatro veces el número de documentos recuperados. Si el usuario ha visto 6 nuevos documentos relevantes añadidos a esos 15 previamente

[156] E. M. KEEN (1971) op. cit.

conocidos, podemos estimar que la base de datos contiene 16 ó 17 documentos relevantes que él nunca ha visto y a partir de aquí puede intentar recuperarlos, modificando, si lo considera oportuno, su estrategia de búsqueda.

Siguiendo con el ejemplo, el ratio de novedad sería 6/10. Un ratio de cobertura alto, podría dar al usuario alguna confianza en que los sistemas localicen todos los documentos relevantes. También sugiere que el sistema es efectivo en la localización de documentos desconocidos para el usuario. Del ejemplo anterior, el usuario puede inferir que aproximadamente el 60% de algún grupo de documentos relevantes recuperados para esta pregunta y esta base de datos, en particular, no será previamente conocida. Por supuesto, al usuario no le interesa saber que puede recuperar, aquellos documentos que él ya conoce, por lo tanto, es deseable que el ratio de novedad sea alto. En cuanto a la exhaustividad relativa, puede referirse más directamente a la cuestión de cómo el usuario quiere algunos documentos. Supongamos que el sistema presenta 20 documentos al usuario y que éste quiere 5 documentos relevantes. Si solo hay 3 documentos relevantes entre los 20, la exhaustividad relativa será 3/5, el usuario solo obtiene 3 de los 5 que busca. Si por el contrario, hay 5 o más documentos relevantes entre los 20, entonces, presumiblemente el usuario podrá abandonar después de encontrar los 5 deseados con una exhaustividad relativa de 5/5 es decir de 1. Si la exhaustividad relativa es de 1, la medida falla al referirse los esfuerzos a localizar los documentos.

Podría ser que el usuario encuentre los documentos entre los primeros 5 ó 6 examinados o podría ser que necesitara examinar los 20, por lo tanto esto nos da pie para definir una nueva medida: **esfuerzo de exhaustividad**, que es el ratio del número de documentos relevantes deseados partido por el número de documentos examinados para encontrar el número de documentos relevantes deseados. Esta medida asume que la colección contiene el número de documentos relevantes deseado y que el sistema de recuperación permite al usuario localizarlos todos, lo cual aunque es deseable no siempre es posible. Este ratio puede ir de 1, si los documentos relevantes deseados son los primeros documentos

examinados por él, a próximo a 0, si el usuario necesita examinar un gran número de documentos para encontrar los pocos que desea.

Otras medidas relacionadas con el usuario son la utilidad y satisfacción. De las medidas vistas hasta ahora, éstas son las más subjetivas, por lo que habrá que valorarlas con mucho cuidado. La satisfacción pone énfasis en la coincidencia entre lo que el usuario quiere y lo que el usuario recibe. La realidad es que los usuarios quieren una exhaustividad alta con una precisión alta también lo cual en términos teóricos es incompatible.

6. La recuperación de la información en español: experimentos más significativos.

En uno de los apartado anteriores, al hablar de la historia de la R.I., hacíamos referencia a que esta disciplina tan sólo lleva poco más de medio siglo. En el caso español se limita a un par de décadas, lo que nos hace presuponer que el volumen global de trabajos en estos temas sea escaso.

La mayor parte de los trabajos de investigación en este campo, se dan en el área anglosajona[157], allí, según un artículo de Félix de Moya, la R.I. representa un 25% del total de la investigación en biblioteconomía, en España, haciendo un estudio sobre las investigaciones presentadas en DOCUMAT, tan solo un 15%[158] de los trabajos se refiere a las labores de almacenamiento y recuperación. De estos

[157] F. MOYA ANEGÓN. (2000) op. cit.

[158] A. B. RÍOS HILARIO. Metodología, técnicas y estrategias de investigación en las Jornadas Españolas de Documentación Automatizada (1981-1996*)* IV Jornadas Españolas de Documentación. En *Los sistemas de información al servicio de la sociedad: actas de las jornadas. VI Jornadas españolas de Documentación*. Valencia: FESABID, 1998.

trabajos de investigación en biblioteconomía y documentación un porcentaje elevado son trabajos de revisión bibliográfica; esto se debe a que se trata de una disciplina joven[159].

Teniendo esto en cuenta, no será de extrañar que al limitar el campo de investigación a la recuperación de la información en español, el número de trabajos descienda en un porcentaje aún mayor, ya que aunque el español es la tercera lengua más hablada, existen muy pocas herramientas para la recuperación de la información realizadas para dicho idioma. Quizá esto se deba, entre otros motivos, como indican los profesores Carretero y Rodríguez[160], a la poca influencia tecnológica de los países de habla hispánica. A esto hay que añadir la falta de colecciones experimentales, tipo el Test de Cranfield, que existe por ejemplo para el Inglés, y que permitiría la comparación de sistemas y la evaluación de los mismos.

6.1. Los experimentos en las TREC

Los experimentos más importantes en el campo de la recuperación de la información en español, son los que se realizaron entre los años 1994, 1995 y 1996 en las TREC 3, 4 y 5 respectivamente. La importancia de estos experimentos radica en que sus resultados se pueden comparar, ya que utilizan las mismas colecciones documentales, donde previamente se ha establecido la relevancia de los documentos con relación a las preguntas y las mismas medidas de evaluación. A pesar de esto, el principal inconveniente de las actas de las TREC es que los experimentos en español ocupan un espacio muy pequeño, son pocos los trabajos

---

[159] E. DELGADO LÓPEZ CÓZAR Diagnóstico de la investigación en Biblioteconomía y Documentación en España (1976-1996): Estado embrionario. EN *Journal of Spanish Research on Information Science/Revista de Investigación Iberoamericana en Ciencia de la Información y Documenta*ción. Vol 1 (1) 2000 p. 79-93

[160] J. CARRETERO, S. RODRÍGUEZ Building lexical tools to manage information written in Spanish *Journal of Information Science* 1996 22, (5). p 391-399

que hay y en la mayor parte de los artículos que hacen referencia al español, apenas especifican cómo han realizado los experimentos. En la mayoría de los casos se limitan a decir que adaptaron el trabajo realizado para el inglés, ya que parten de la base de que los sistemas de recuperación son independientes de los idiomas, por lo que simplemente dicen que se adaptan las lista de palabras vacías con la que trabajan o las terminaciones en el caso de los sufijos, pero no suelen incluir dichas listas. Un inconveniente paralelo a esto, es que las colecciones de documentos con sus criterios de relevancia no son de libre copia con lo cual no es fácil el acceso a ello, lo que impide comparar experimentos que se realicen en otros lugares.

Por ser los experimentos más importantes para el español, trataré de explicar brevemente qué es lo que los distintos grupos han ido realizando en las TREC-3, 4 y 5. Algunos grupos de trabajo participaron más de un año, otros sólo lo hicieron una vez, la exposición que haré a continuación será por grupos de trabajo, ya que en algunos casos los trabajos se basan en los que realizaron el año antes. Algunos grupos no especifican apenas nada de cómo realizaron sus tareas de recuperación, por lo que el detalle que se da aquí es mínimo.

6.1.1. Universidad de Dublin

La Universidad de Dublin participó en las TREC-3, 4 y 5. En sus trabajos ha ido adaptando distintas técnicas que servían para el inglés.

El primer año que participó, basaron su trabajo en la recuperación por triagramas[161]. Para ello partieron del experimento de Cavnar[162] de las TREC-2,

[161] A. F. SMEATON, R. O'DONNELL, F. KELLEDY. Indexing Structures Derived from Syntax in TREC-3: System Description 1994 [en línea] <http://trec.nist.gov/pubs/trec3/t3_proceedings.html> [Consultado el 24/07/00]

[162] W. CAVNAR (1993) op. cit.

intentando de este modo mejorar los pobres resultados que se habían obtenido. Lo primero fue la normalización de las letras, para ello los caracteres acentuados fueron sustituidos por los no acentuados, los dígrafos (ch, ll) se consideraron un solo carácter. Tradujeron la lista de palabras vacías del lematizador de Porter, que incluye también grupos de palabras. El suprimir estas palabras hizo reducir el tamaño del texto en un 35%. Después de analizar la frecuencia de los trigramas se dieron cuenta de que no se seguía la distribución de Zipf[163]. Aplicaron el mismo proceso a la preguntas. Se creó un fichero invertido con los trigramas y se determinó su frecuencia de aparición. El esquema de pesado fue tf * idf. Calcularon la frecuencia de los trigramas y las preguntas y establecieron el umbral a partir de cual se descartaron algunos trigramas. La conclusión a la que llegó este grupo es que los resultados con trigramas son peores que aquellos trabajos con lematización fuerte.

El problema que tiene este trabajo es que no especifica cuáles son las palabras vacías que utiliza, sólo indica que trabaja con la lista del lematizador de Porter pero no si con la versión española o con la inglesa, y en este caso si la adapta o simplemente la traduce.

El año siguiente, este mismo grupo volvió a participar, en esta ocasión su tarea pretendía probar varios post-tagger[164] para el español, desarrollados con anterioridad a este experimento. Ninguno de ellos había sido evaluado previamente en cuanto a la exactitud de la lematización. Se utilizó el SMART que había sido preparado para permitir que los documentos fueran indizados por la

---

[163] Cf. La loi de Ziphf. Linguistique et Statistique de L' Enciclopedie Universalis Version 3.0 Sur CD-ROM [en línea] <http://users.info.unicaen.fr./~guguet/java/zipfeu.html [consultado el 13/01/01]

W.L. Zipf Law and the Structure and Evolutioon of Languages" (letter to Editors) Complexiti 2 (5) 12-13 1997. [en línea] http://linkage.rokfeller/edu/wli/pub/comp98_zipf.html [Consultado el 13/01/01]

[164] A. F. SMEATON, R. O'DONNELL, F. KELLEDY. Thresholding Posting Lists, Query Expansion with WordNet and POS Tagging of Spanish" 1995 [en línea] http://trec.nist.gov/pubs/trec4/t4_proceedings.html [Consultado el 24/07/00]

base de las palabras y asociados a las categorías gramaticales. La recuperación se basó en el SMART y el tradicional tf * idf para el pesado de los términos, ofreciendo una salida ordenada. Para la recuperación se etiquetaron los temas del mismo modo que los documentos y la salida fue por categorías gramaticales. Al no haber un conjunto de temas con sus juicios de relevancia fiables para los datos en español, se decidió que los adjetivos se pesaran doblemente con el esquema tf*idf. Los resultados de este experimento fueron peores que los obtenidos por otros grupos, esto quizá puede ser debido a que estos post tagger no obtenían los lemas precisos.

El principal inconveniente de este trabajo es que al no haber evaluado previamente la eficacia de los post-tagger no se sabe si el sistema no funciona porque la aplicación de los post tagger a la recuperación del español no es adecuada o simplemente los post tagger no tienen un buen rendimiento.

El último año que participó este grupo, tenía como objetivo evaluar la ejecución del nuevo lematizador para el español desarrollado por Porter[165]. Aunque el esquema del lematizador para el español es prácticamente igual que el del inglés, las diferencias consisten en que en este caso la longitud de las palabras se mide en sílabas pero sin tener en cuenta una pequeña lista de prefijos. La diferencia más importante con respecto al algoritmo de Porter es en los verbos, ya que en español hay que tener en cuenta las cinco personas (la segunda es igual para el singular y para el plural), el tiempo y el modo. Esto se complica con el gran número de verbos irregulares que hay. Respecto a las palabras vacías, se tradujo la lista del lematizador de Porter y se amplió con algunas palabras más, procedentes de categorías gramaticales vacías. Hicieron tres tareas probando con los términos lematizados, formas completas y con los términos lematizado de las preguntas y haciendo una expansión de las mismas teniendo en cuenta los diez

[165] F. KELLEDY. and A. F. SMEATON. TREC-5 Experiments at Dublin City University: Query Space Reduction Spanish and Character Shaape Encoding 1996 [en línea] <http://trec.nist.gov/pubs/trec5/t5_proceedings.html> [Consultado el 24/07/00]

primeros documentos recuperados. Entre las conclusiones a las que llegó este grupo está que el uso de las formas sin lematizar produce mejores resultados que el uso de formas cortas en las preguntas. La expansión manual de las preguntas de formas iniciales cortas, mejora la recuperación respecto al uso simple de formas cortas con precisiones altas al final de la escala.

Los problemas de este trabajo son que no especifica la lista de palabras vacías, tampoco lo hace con los finales que tiene en cuenta el lematizador, ni los prefijos que no se tienen en cuenta a la hora de medir las palabras. Respecto a los verbos no especifica si lo que hace es identificar modelos o por el contrario son formas verbales.

6.1.2. Instituto de Investigación Medioambiental de Michigan

El Instituto de Investigación Medioambiental de Michigan, tan solo participó en las TREC-3; en esta ocasión presentó un trabajo basado en cuatrigramas[166], utilizando también esquemas de pesado de términos según los modelos tradicionales. Este grupo pretendió aunar en un mismo sistemas las ventajas de los n-gramas que no requieren conocimiento lingüístico, y no hay que tener en cuenta las palabras vacías (aunque hay algunos experimentos que sí las suprimen), con las del modelo de espacio vectorial (que permite representar fácilmente el contenido del documento y la pregunta facilitando así su comparación). El tamaño elegido para los gramas fue de cuatro. Entre las conclusiones a las que llegó este grupo es que los cuatrigramas mantienen mucha de la información que los sistemas que trabajan con palabras vacías y con la lematización pierden, y esta información tiene una frecuencia de aparición alta, lo que nos indica que es representativa. Las mejoras frente a estos sistemas son que

[166] W. CAVNAR Using An N-Gram-Based Document Representation With A Vector Processing Retrieval Model 1994 [en línea] <http://trec.nist.gov/pubs/trec3/t3_proceedings.html> [Consultado el 24/07/00]

en los errores ortográficos no introducen tanto ruido y gran ventaja es que es independiente del idioma.

El problema de este trabajo es que no da los resultados de las medidas de precisión y exhaustividad.

6.1.3. Universidad de Cornell

La Universidad de Cornell participo durante los tres años consecutivos en que se realizaron las tareas con el español.

En 1994[167], presentó una aplicación del SMART para el español, esto necesitó de tres tareas, construir un SMART de 8 bits, crear las reglas de lematización para el español, en este sentido solo se trabajó con plurales, los cambios fueron para los plurales: de *"as"*, a *"a"* de *"es"* a *"e"*, de *"os"* a *"o"* y la *"z"* final se cambió por *"c"*. La última tarea que realizaron fue establecer una lista de palabras vacías en función de la frecuencia alta de aparición, de la inicial de 800 términos se redujo a 324. La colección se indizó teniendo en cuenta tanto las palabras vacías como las reglas de lematización.

Los problemas de este trabajo son que no indica la lista de palabras vacías, ni los criterios de frecuencia de los términos o la manera de eliminación de la lista inicial, tampoco especifica los resultados en términos de precisión y exhaustividad por franjas de documentos

Al año siguiente[168] este mismo grupo repitió el trabajo realizado el año antes. En este expandieron por 50 términos la recuperación de 20 documentos. La

---

[167] C. BUCKLEY, G. SALTON, J. ALLAN, A. SINGHAL Automatic Query Expansion Using SMART: TREC 3 1994 [en línea] <http://trec.nist.gov/pubs/trec3/t3_proceedings.html> [Consultado el 24/07/00

[168] C. BUCKLEY, A. SINGHAL, M. MITRA New Retrieval Approaches Using SMART: TREC 4 1995 [en línea]<http://trec.nist.gov/pubs/trec4/t4_proceedings.html> [Consultado el 24/07/00]

conclusión de este trabajo fue que el SMART es independiente de la lengua que utilice, ya que en este caso las operaciones realizadas fueron exactamente las mismas que para el inglés y obtuvieron buenos resultados. En este experimento probaron con dos vectores. No se tuvieron en cuenta los acentos. Tampoco se tuvieron en cuenta las frases. En el SMART estas frases se formaron mediante aproximaciones estadísticas derivadas de la frecuencia de aparición conjunta. Para el español esto en principio se puede hacer igual que para el inglés con lo cual si esto se tuviera en cuenta se produciría una mejora de entre un 12 y un 15%.

Los resultados para el español están por encima de la media realizando las mismas tareas que para el inglés

El último año de participación del grupo de Cornell[169], utilizaron un lematizador simple que quitaba los finales de las palabras para normalizar la frecuencia de aparición conjunta. Los documentos fueron inicialmente ordenados y vueltos a ordenar basándose en la realimentación de la relevancia de Rocchio, asumiendo que los documentos relevantes son los que aparecen en los primeros lugares. La aproximación de Cornell también usó palabras no vacías con un índice de aparición suficientemente frecuente. A parte de la efectividad en la recuperación es suficientemente notable el esfuerzo aportado para permitir manejar el SMART con un corpus en español muy pequeño, lo que fue incluido en el módulo de normalización.

[169] C. BUCKLEY, A. SINGHAL, M. MITRA Using Query Zoning and Correlation within SMART: TREC 5 1996 [en línea]<http://trec.nist.gov/pubs/trec5/t5_proceedings.html> [Consultado el 24/07/00]

6.1.4. Universidad de Masachussets

La Universidad de Masachusset participó durante tres años consecutivos. El primer año[170] centró su experimento en la evaluación del efecto del proceso morfológico, para ello fue necesario hacer algunas pequeñas modificaciones para aplicar INQUERY[171] a colecciones de documentos en español. Se creó una lista de palabras vacías de manera manual. Se desarrolló una lematización para el español y se modificó el interfaz gráfico. El lematizador usado está basado en el de Porter, pero es más complejo que el empleado para el inglés porque en español hay más tiempos verbales y más verbos irregulares que en inglés. El sistema para el inglés tenía un post-tagger pero en el caso del español esta parte se inhabilitó por no existir un programa de estas características para el español. Las expresiones vacías de contenido no se utilizaron. El trabajo tuvo dos partes:

SIN002: basado en un procesado automático completo de los temas. El pesado de los términos se hizo teniendo en cuenta el pesado de los documentos en inglés

SIN001: este sistema es semiautomático. El usuario podía hacer suprimir términos de las preguntas, agruparlos con operadores de proximidad o ajustar el peso de los términos en la pregunta.

A la luz de los resultados parece en principio que se pueden aplicar con éxito las mismas técnicas para el español que para el inglés

---

[170] J. BROGLIO, J. P CALLAN, W. B CROFT, D. W. NACHBAR. Document Retrieval and Routing Using the INQUERY System 1994 [en línea] <http://trec.nist.gov/pubs/trec3/t3_proceedings.html> [Consultado el 24/07/00]

[171] Sistema estadístico de estimación de la probabilidad de aparición de un concepto mediante es esquema tf*idf

Al año siguiente este grupo trató de mejorar el trabajo realizado el año antes[172], en esta ocasión incorporaron la técnica Infinder para la expansión de preguntas y se centraron en la comparación con los resultados del año anterior. Al no haber un part-of- speech tagger para el español, los temas fueron analizados con un reconocedor de sintagmas nominales. Las secuencias de los nombres y de los pares de nombres-adjetivos fueron elegidos por el operador #PHRASE. No se tradujeron las frases vacías al español pero sí se suprimieron automáticamente las contenidas en la tabla siguiente.

Español	**Inglés**
Evidencia	Evidence of
Hay	Are/is there
Indicaciones de	Indications of
Cúales son	Which are
Cómo van	How is
Tendrá	Will it be/ have
Información sobre	Information about

Tabla 2 Palabras vacías de la U. de Masachussets TREC-4

Como puede observarse dicha lista no tiene demasiado sentido ya que es muy corta, tampoco pueden considerarse frases en español "indicaciones de" o "hay", aunque en inglés correspondan a varias palabras.También se introdujo una lista de finales de palabras que indicaban que dichas palabras eran nombres; usaron once finales:

[172] J. ALLAN, L. BALLESTEROS, J. P CALLAN., W. B. CROFT, Z. LU Recent Experiments with INQUERY 1995 [en línea] <http://trec.nist.gov/pubs/trec4/t4_proceedings.html> [Consultado el 24/07/00]

FINALES	EJEMPLO
-dor	Matador
-d	Verdad
-ata	Corbata
-z	Arroz
-[sz]mo	Capitalismo
-miento	Conocimiento
[cs][ií]a	Democracia
-[cgnstx]i [oó]n	Lección
-az [oó]n	Corazón
-cida	Conocida
-i[ae]nte	Pariente

Tabla 3 Finales utilizados por la U. de Masachusset TREC-4

La lista de finales es también bastante parca. No indica si hace alguna transformación o si simplemente suprime esos finales. El otro problema de la lista es que no todas las terminaciones elegidas son del todo correctas. Por ejemplo, la terminación en *-z* no siempre indica que se trata de un nombre: así, "*fugaz*", que termina en *z* y es un adjetivo, y todos los imperativos que terminan en *–d*, son verbos, y según los criterios de este trabajo serían etiquetados como nombres.

En su investigación señalan la ambigüedad del español, ya que una palabra puede ser al mismo tiempo un nombre y una forma verbal, como es el caso de *denuncia*, sustantivo pero también tercera persona del singular del presente de indicativo del verbo *denunciar*. Sin embargo esto se puede deducir empleando normas sintácticas, por ejemplo teniendo en cuenta si va precedido o

no de un artículo. Para evitar esta ambigüedad, sería necesario introducir normas que distinguieran las diferentes categorías gramaticales.

Se utilizaron los nombres y sintagmas nominales seleccionados por el reconocedor para identificar frases. Para la expansión de preguntas los sintagmas nominales fueron definidos de manera simple, doble o triple en función de la probabilidad de aparición conjunta. Se hicieron dos experimentos:

SIN010: con INQUERY

SIN011: con la versión modificada. La precisión obtenida con este experimento es ligeramente inferior.

Se usó un proceso similar al utilizado para el inglés para generar la base de preguntas para la recuperación para los temas en español.

Al año siguiente[173] trataron de mejorar el trabajo de las TREC-4, para ello establecieron la combinación del análisis global y la realimentación local al generar la expansión de preguntas basada en el análisis local del contexto, esencialmente en la expansión de preguntas por términos que aparecen en documentos en los primeros puestos de recuperación si aparecen cerca de términos en las preguntas. Este grupo usó un proceso sofisticado de normalización desarrollado para el trabajo de las TREC-4.

[173] J. ALLAN, L. BALLESTEROS, J. P. CALLAN., W. B. CROFT, Z. LU INQUERY at TREC-5 1996 [en línea] <http://trec.nist.gov/pubs/trec5/t5_proceedings.html> [Consultado el 24/07/00]

6.1.5. Universidad de Berkely

La Universidad de Berkely (California) participó en las TREC-4 y en las 5. En su primera participación[174], basaron su trabajo en las dos aproximaciones conflictivas a la lematización en las TREC-3. La afirmación de Cornell, que mantenía que sólo era necesario lematizar de manera muy básica y tener una lista de palabras vacías que podía elaborarse con las raíces más frecuentes, revisada manualmente para elegir que raíces se debían mantener, frente a la afirmación de Massachusetts, que mantenía que con un lematizador más sofisticado se podían obtener mejoras significativas. A pesar de que Berkeley consideraba que no era posible mejorar el rendimiento, emprendió la tarea de desarrollar un lematizador para el español. Éste, comenzaba encontrando los finales más largos, cuando se encontraba un final, el programa convertía el final c en z, formalizaba también las formas verbales. Los verbos son más complejos de lematizar, hay que tener en cuenta las tres conjugaciones y la división entre verbos regulares e irregulares. En el caso del resto de las categorías gramaticales distintas de los verbos es mucho más fácil aunque hay ciertas irregularidades. Se tradujo la lista de palabras vacías del SMART y se amplió para conseguir sus posibles variantes, la lista que se consiguió fue de 10.000 palabras.

El método de trabajo consistió en elaboración de manera manual de las preguntas, para posteriormente traducirlas al inglés buscando en la base de datos Melvyl News, volvieron a formular las preguntas en inglés basándose en estos buscadores y las tradujeron de nuevo al español.

La conclusión de los que hicieron el experimento fue que el lematizador era tan bueno para el inglés como para el español.

---

[174] F. C. GEY, J. A. CHEN, M. HE and JASON Logistic Regression at TREC4: Probabilistic Retrieval from Full Text Document Collections 1995 [en línea] <http://trec.nist.gov/pubs/trec4/t4_proceedings.html> [Consultado el 24/07/00]

Al año siguiente[175] volvieron a trabajar con el mismo algoritmo que en las TREC-4. Los esfuerzos de este trabajo fueron en la línea de mejorar el lematizador morfológico de las TREC-4, para ello se añadieron más términos de formas verbales regulares e irregulares. Se contemplaron 184.496 verbos que fueron reducidos a 3.375 lemas. Se identificaron los acrónimos y se excluyeron del proceso de lematización. Este trabajo suprime los acentos en la normalización. Parece que los resultados en este caso son superiores a los del año anterior.

El problema que encontramos en este trabajo es que da muy pocos detalles.

6.1.6. Universidad Central de Florida

La Universidad Central de Florida tan solo participó en 1995[176]. Entre las aportaciones de este trabajo está la creación de preguntas en español de manera manual construyendo para ello una amplia lista de sinónimos.

6.1.7. Equipo de David A. Grossman

El equipo de trabajo dirigido por David A. Grossman, participó en las TREC-4 y 5. El primer año[177], basaron su trabajo en los juicios de relevancia de

---

[175] F. C. GEY, A. CHEN, J. HE, L. XU, and J. MEGGS, Term importance boolean conjunct training, negative term, and foreing language retrieval: probabilistic algorithm at TREC-5 1996 [en línea] <http://trec.nist.gov/pubs/trec4/t4_proceedings.html> [Consultado el 24/07/00]

[176] D. HARMAN Overview of the Fourth Text Retrieval Conference (TREC-4) 1995 [en línea] <http://trec.nist.gov/pubs/trec4/t4_proceedings.html>[Consultado el 24/07/00]

J. D. DRISCOLL. Multi-lingual Text Filtering Using Semantic Modeling, (Praxis Technologies), S. Abbott, K. Hu, M. Miller, G. Thesis (University of Central Florida)

[177] D. A. GROSSMAN, D. O. HOLMES , O. FRIEDER, M. D. NGUYEN AND. C. E. KINGSBURY Improving Accuracy and Run-Time Performance for TREC-*4* 1995 [en línea] http://trec.nist.gov/pubs/trec4/t4_proceedings.html [Consultado el 24/07/00]

las TREC-3. Experimentaron con trigramas, 4-gramas y 5-gramas. Los resultados de los 3 y 4 gramas fueron muy similares. Los mejores fueron los obtenidos con gramas de tamaño 5. Para el desarrollo de este trabajo construyeron una lista de palabras vacías basándose en los 500 términos más frecuentes de un texto, y aplicaron el sistema de espacio vectorial.

En las conclusiones de dicho experimento no se hace ninguna referencia al caso español.

Al año siguiente[178], intentaron un esquema mejor de pesado, tal que los términos añadidos por la realimentación de la relevancia fueran pesados de diferente forma que aquellos que estaban en la pregunta original. Se usó la realimentación de la relevancia para el inglés, el español y el chino. Para el español utilizaron un prototipo relacional para obtener resultados automáticos. Desarrollaron una lista con 500 palabras vacías obtenidas de una lista de frecuencias y consultaron con especialistas para determinar los términos que debían incluir. Se identificaron los 10 documentos recuperados en primer lugar usando la medida del coseno. Los términos se ordenaron por n*idf, y los términos contenidos en los diez primeros documentos recuperados se incorporaron para la realimentación de las preguntas.

6.1.8. Departamento de defensa

El Departamento de defensa sólo participó en las TREC-4 [179]. Para esta ocasión, desarrolló un sistema de espacio vectorial que trabaja con la técnica de n-gramas de tamaño 5 para la categorización de documentos denominado

[178] D. A. GROSSMAN, J. REICHART, A. CHOWDHURY, C. LUNDQUIST, D. HOLMES, O. FRIEDER Using Relevance Feedback within the Relational Model for TREC-5 1996 [en línea] http://trec.nist.gov/pubs/trec5/t5_proceedings.html [Consultado el 24/07/00]

[179] S. HUFFMAN Acquaintance: Language-Independent Document Categorization by N-Grams 1995 [en línea] <http://trec.nist.gov/pubs/trec4/t4_proceedings.html> [Consultado el 24/07/00]

"Acquaintance". Esta técnica fue utilizada básicamente de la misma manera que para el inglés. Los problemas encontrados para el español fueron similares a los del inglés, la descripción de los temas es bastante corta y no proporciona suficiente información. El inconveniente es que los que realizaron el experimento no sabían español. La evolución del sistema es bastante pobre.

6.1.9. Universidad del Estado de Nuevo México

La Universidad del Estado de Nuevo México, participó en 1994 y 1995; el trabajo del primer año[180], pretendía que las preguntas elaboradas en un idioma se pudieran usar para la recuperación de documentos en varios idiomas. Aunque todos los documentos de una colección puedan ser traducidos a un único idioma, una aproximación más eficiente es simplificar la traducción de las preguntas a cada una de las lenguas del documento. Investigaron los distintos métodos de traducción de preguntas. Se usaron 5 métodos:

Traducción término a término usando un diccionario bilingüe

Uso de un corpus paralelo para las frecuencias altas de los términos

Usa de un corpus paralelo para localizar estadísticamente términos significativos

Utilización de la técnica LS5 de corpus paralelo

Salvo el sistema manual el resto introducía un factor bastante alto de ruido.

---

[180] M. DAVIS, T. DUNNING A TREC Evaluation of Query Translation Methods For Multi-Lingual Text Retrieval 1995 [en línea] <http://trec.nist.gov/pubs/trec4/t4_proceedings.html> [Consultado el 24/07/00]

El siguiente año, presentaron el sistema *RECUERDO*[181] desarrollado por CRL, es una modificación del *SMART*. Entre los nuevos desarrollos cuenta con un lematizador basado en el modelo de Porter que contiene 145 reglas para normalizar los términos en español. La complejidad de los verbos irregulares fue parcialmente manipulada, aunque se decidió hacer sin especificar los paradigmas de los verbos irregulares precisamente para mantener la velocidad del algoritmo de lematización. La efectividad de esta aproximación ha sido sólo testada con el esquema del experimento de recuperación. El sistema es capaz de indizar alrededor de 200Mb por hora, en español o en inglés y crea índices de la mitad de tamaño que el fichero original Así mismo, puede ejecutar la expansión de términos para encontrar un subconjunto de términos. Para la traducción usa un diccionario bilingüe, donde los homógrafos son conflactados después de hecha una normalización basada en el algoritmo de Porter. Los duplicados equivalentes no se suprimen del conjunto de términos conflactados. Los equivalentes en español fueron normalizados y lematizados con la variante española del algoritmo de Porter. Entre las conclusiones que se obtuvieron se dedujo que la desambiguación de términos en un conjunto de equivalentes suministrado por un diccionario de transferencia bilingüe, puede resultar una mejora sustancial sobre la mayoría de los sistemas CLTR vistos hasta la fecha. La alta ejecución proviene de la desambiguación del corpus para la traducción de preguntas que necesitan una traducción especializada, quizá una aproximación interactiva a la traducción de nueva terminología.

[181] M. DAVIS New Experiments In Cross-Language Text Retrieval At NMSU's Computing Research Lab 1996 [en línea] <http://trec.nist.gov/pubs/trec5/t5_proceedings.html> [Consultado el 24/07/00]

6.1.10. El Centro Xerox

El centro Xerox, participó en 1995 y 1996. En las TREC-4[182], se centró en el efecto del análisis del lenguaje en un sistema de R.I. El proceso que sigue es, comenzar por el *part of speech tagger*, para etiquetar las palabras, cada palabra etiquetada se lematiza, y se le añaden campos con la siguiente información:

f1 forma lematizada.

f2 sintagmas nominales lematizados.

f3 verbos lematizados.

f4 frases lematizadas sin espacios en blanco que son sustituidos por guiones.

Con esta información se realizaron dos tareas, una basada en las formas individuales lematizadas y otra en los sintagmas nominales lematizados que corresponde a usar el texto lematizado y doblar el pesos de los componentes en los sintagmas. La indización de los sintagmas nominales por su componentes más que tratarlos como unidades simples produce una ligera mejoría en los test preliminares. Las preguntas en español son cortas, por esta razón se espera que la expansión produzca beneficios.

Los resultados de este experimento mostraron que la lematización usando los casos de morfología flexiva ayuda en la mayor parte de los casos. Sin embargo, cuando las preguntas son largas para menos de 20 de documentos examinados no hay mejoras. Doblando el pesado de los sintagmas obtenemos mejoras más importantes.

---

[182] M. HEARST, J. PEDERSEN, P.,PIROLLI, H. SCHUTZE, G. GREFENSTETTE, D. A. HULL. Xerox Site Report: Four TREC-4 Tracks 1995 [en línea] <http://trec.nist.gov/pubs/trec4/t4_proceedings.html> [Consultado el 24/07/00]

El experimento concluyó que este lematizador no obtenía resultados tan buenos para el español como para el inglés. Otra de las conclusiones fue que las herramientas lingüísticas tampoco producen beneficios espectaculares a la hora de resolver los problemas de la R.I.

Al año siguiente[183] usaron un *part-of-speech tagger* para identificar e indizar pares de nombres, como alternativa a la simple yuxtaposición de pares de palabras no vacías o "sintagmas estadísticos" usados por otros grupos como Cornell. En este trabajo el texto es separado en palabras clave, etiquetado y después se le pasa un lematizador lo que significa que los términos no ambiguos son reducidos a su raíz. Los acentos se suprimieron, dado que hay pocas palabras que varíen su significado en función de la acentuación no influye demasiado. El texto resultante es indizado y recuperado con el *SMART*. El grupo Xerox usó un lematizador para reducir las palabras a sus raíces como alternativa a los lematizadores "fuertes" empleados en los procesos de R.I. Usaron un lematizador morfológico para lematizar el texto de la versión inglesa de los temas y traducirlos automáticamente cada raíz resultante a su equivalente en español para recuperar basándose en un diccionario de búsquedas. Este experimento fue el primer intento en las TREC por hacer recuperación multilingüe. Los resultados para el español son mejores que los obtenidos para el inglés

6.1.11. Equipo de Ross Wilkinson

El equipo de Ross Wilkinson tan solo participó en las TREC-4[184]. Estaban interesados en ver cómo se podía combinar el resultado de los

[183] D. A. HULL, G. GREFENSTETTE, M. HEARST, H . SCHUTZE, J. PEDERSEN, P. PIROLLI Xerox TREC-5 Site Report: Routing, Filtering, NLP and Spanish Tracks 1995 [en línea] <http://trec.nist.gov/pubs/trec5/t5_proceedings.html> [Consultado el 24/07/00]

[184] R. WILKINSON, J. ZOBEL, R. SACKS-DAVIS Similarity Measures for Short Queries 1995 [en línea] <http://trec.nist.gov/pubs/trec4/t4_proceedings.html> [Consultado el 24/07/00]

experimentos de la lista de palabras vacías, y palabras vacías y lematización conjuntamente. Para ello desarrollaron una lista de 316 palabras vacías y un lematizador que suprimía 30 finales de palabras, principalmente los sufijos de los verbos regulares, se implementó un mecanismo para que limitara las combinaciones con longitudes de raíces menores de 10 caracteres, para de esta manera, reducir el número de términos de índice. Los mejores resultados fueron los obtenidos por este último experimento. Este grupo demostró que el tamaño de la pregunta influyen en la recuperación. Este trabajo parece indicar que la misma metodología de trabajo para el inglés funciona de manera similar para el español.

6.1.12. Universidad de Maryland

El trabajo de la Universidad de Maryland[185] de las TREC-5, describe cómo se usa la colección de documentos TREC y los juicios de relevancia para evaluar la ejecución de dos técnicas que se han implementado: recuperaciones ad-hoc a través del lenguaje donde los documentos están una lengua y las preguntas en otra. El trabajo concluye que el efecto de cambio de dominio será inherente a las técnicas basadas en corpus tales como Cross-language latent Semantic Indexing al menos en colecciones de texto traducidos que usan al menos las mismas técnicas para la evaluación la habilidad para caracterizar el efecto de cambio de dominio es por otro lado artificial.

6.1.13. Universidad George Mason

La Universidad George Mason[186] al igual que el grupo de Cornell usó un algoritmo de recuperación en dos pasos, incorporando la estimación de la

---

[185] D. W. OARD Alignment of Spanish and English TREC Topic Descriptions 1996 [en línea] <http://trec.nist.gov/pubs/trec5/t5_proceedings.html> [Consultado el 24/07/00]

[186] D. A. GROSSMAN Using Relevance Feedback within the Retrieval Model for Trec-5 1996 [en línea] <http://trec.nist.gov/pubs/trec5/t5_proceedings.html> [Consultado el 24/07/00]

relevancia en dos pasos. Los términos de los temas originales se pesaron mediante el esquema tf* idf para generar el primer documentos del ranking y después los términos de los 10 mejores fueron ordenado por su valor n*idf y los 10 mejores de ellos fueron añadidos a las preguntas para el segundo paso de la recuperación.

Se usó una lista de palabras vacías con 500 términos. La implementación de GMU fue una base de datos relacional.

6.1.14 Comparación de los experimentos TREC para el español

A continuación mostramos unas tablas comparativas de todos los experimentos TREC que acabamos de reseñar. El problema que tiene la comparación es que no se da en todos los trabajos el mismo nivel de detalle.

GRUPOS DE TRABAJO	BASADO	N-GRAMAS	PALABRAS VACÍAS	POST TAGGER	SMART	TAREAS	ESQUEMA DE PESADO	PROBLEMAS
Dublin TREC-3 (1994)	TREC- 2	Sí, tamaño 3	Lista de Porter (254)	No	No	Normalización de los acentos	tf*idf	No indica la lista de palabras vacías No especifica que palabras vacías son las de más de una palabra
Dublin TREC- 4 (1995)		No	No indica	UMASS y NMSU	Sí	Indiza por las bases asociadas a categorías gramaticales. Los adjetivos son doblemente pesados	tf*idf	La no evaluaron previamente los post tagger.
Dublin TREC-5 (1996)		No	Sí, la traducción de la lista de Porter completada con otros términos	No	Sí	Lematización de los términos de los documentos y de las preguntas Expansión de preguntas teniendo en cuenta los 10 documentos primeros recuperados.	tf*idf	No indica la lista de prefijos que no tiene en cuenta para medir las palabras no indica la lista de palabras vacías no especifica los sufijos
Instituto Medioambiental de Michigan TREC-3 (1994)		Sí, tamaño 4	No son necesarias	No	No	Adaptación del modelo de espacio vectorial para que trabaje con cuatrigramas	tf*idf	No da los resultados de las medidas de precisión y exhaustividad

Tabla 4 Comparación de experimentos Trec (parte 1)

GRUPOS DE TRABAJO	BASADO	N-GRAMAS	PALABRAS VACÍAS	POST TAGGER	SMART	TAREAS	ESQUEMA DE PESADO	PROBLEMAS
Cornell TREC-3 (1994)		No	Sí (342 términos) De elaboración propia	No	Adaptado a 8 bits	Construir un SMART de 8 bits adaptar las reglas de lematización para el español establecer una lista de palabras vacías	No indica cual utiliza	No indica la lista de palabras vacías. No especifica los resultados en términos de precisión y exhaustividad por franjas de documentos
Cornell TREC-4 (1995)	En el experimento que el mismo grupo realizó para las TREC-3	No	Sí (342 términos) De elaboración propia	No	Sí	No tienen cuenta ni acentos ni frases	No especifica	No da detalles de los resultados para el español.
Cornell TREC-5 (1996)	No especifica	No	No especifica	No	Sí con una adaptación para trabajar con corpus en español	Mediante un lematizador simple se normalizan los términos	No especifica	Pocos detalles

Tabla 5 Comparación de experimentos Trec (parte 2)

GRUPOS DE TRABAJO	BASADO	N-GRAMAS	PALABRAS VACÍAS	POST TAGGER	SMART	TAREAS	ESQUEMA DE PESADO	PROBLEMAS
Masachussets TREC-3 (1994)	Basado en un trabajo hecho para el inglés	No	Sí de elaboración propia, no indica el número	No, aunque la versión inglesa de este sistema si la tiene	No	Lematizador para el español	No especifica	Pocos detalles
Masachussets TREC-4 (1995)	Basado en las TREC-3	No	Lista con siete "frases"	No	No	Técnica Infinder de expansión de preguntas	No especifica	Lista de palabras vacías muy corta Errores de etiquetado de finales
Masachussets TREC-5 (1996)	Basado en las TREC-4	No	No especifica	No	No	Expansión de preguntas	No especifica	Mismos problemas que en las Trec-4
Berkely TREC-4 (1995)	Basado en las diferencias establecidas entre los trabajos de Cornell y Masachussets de las TREC-3	No	La del SMART ampliada a 10.000, pero no indica la lista ni como la establece.	No	Sí	Elaboración de manera manual de las preguntas	No especifica	Pocos detalles
Berkely TREC-5 (1996)	TREC-4	No	Sí	No	No	Normaliza los acentos, identifica acrónimos y los suprime antes de la lematización	No especifica	No especifica las palabras vacías que utiliza. Pocos detalles

Tabla 6 Comparación de los experimentos Trec (parte3)

GRUPOS DE TRABAJO	BASADO	N-GRAMAS	PALABRAS VACÍAS	POST TAGGER	SMART	TAREAS	ESQUEMA DE PESADO	PROBLEMAS
Universidad Central de Florida TREC-4 (1995)	No especifica	No	No especifica	No	No	Construcción de preguntas en español, construcción de lista de sinónimos en español,	No especifica	Muy pocos detalles
David A Grossmann TREC-4 (1995)	En los juicios de relevancia de las TREC-3	De tamaño 3, 4, 5.	Si, 500 términos	No	No	Sistema de espacio vectorial	No especifica	Sin referencia a los resultados para el español.
David A Grossmann TREC-5 (1996)	TREC-4	Sí	Si, 500 términos	No especifica	No	Pesado de los términos dependiendo del lugar donde aparezcan	Distintos esquemas de pesado dependiendo de si los términos son los de la preguntas o los añadidos en la realimentación de las preguntas	No da detalles de los resultados para el español.

Tabla 7 Comparación de experimentos Trec (parte 4)

GRUPOS DE TRABAJO	BASADO	N-GRAMAS	PALABRAS VACÍAS	POST TAGGER	SMART	TAREAS	ESQUEMA DE PESADO	PROBLEMAS
Departamento de Defensa TREC-4 (1995)	En un trabajo para el inglés	Sí de tamaño 5	No especifica	No	No	Uso de n-gramas	No especifica	Muy pocas especificaciones
Nuevo México TREC- 4 (1995)	No especifica	No	No especifica	No	Sí pero modificado	Traducción de las preguntas con un diccionario bilingüe	No especifica	Pocos detalles
Nuevo México TREC-5 (1996)	Porter	No	No especifica	Sí	Sí, modificado	Normalización de los términos y los traduce con un diccionario bilingüe	No especifica	Pocos detalles
Xerox TREC- 4 (1995)	No especifica	No	No especifica	Sí	No	Utiliza un lematizador para la normalización	No especifica	Pocos detalles
Xerox TREC-5 (1995)	TREC-4	No	No especifica	Sí	Sí	Utiliza un part of speech tagger para la normalización	No especifica	Pocos detalles

Tabla 8 Comparación de experimentos Trec (parte 5)

GRUPOS DE TRABAJO	BASADO	N-GRAMAS	PALABRAS VACÍAS	POST TAGGER	SMART	TAREAS	ESQUEMA DE PESADO	PROBLEMAS
Ross Wilkinson TREC- 4 (1995)	No especifica	No	316 términos	No	No	Utiliza un lematizador	No especifica	No da la lista de palabras vacías
Maryland TREC-5 (1996)	No especifica	No	No especifica	No	No	Suprime palabras vacías y lematiza	No especifica	Pocos detalles
George Mason TREC- 6(1995)	No especifica	No	500 términos	No	No	Se pesan los documentos y se establecen los diez mejor recuperados y con ellos se expanden las preguntas	Tf*idf	No especifica resultados

Tabla 9 Comparación de experimentos Trec (parte 6)

6.2. Experimentos de R.I. para el español fuera de las TREC

Aparte de los experimentos de las TREC, en España se han realizado algunos experimentos con el español, que aplican conocimiento lingüístico a la R.I. El primero de ellos, es una aproximación a la lematización[187], el segundo, compara la lematización, con la recuperación basada en n-gramas, de distinto tamaño. A continuación explicaremos brevemente en qué han consistido estos trabajos.

El primero de ellos es una aplicación del algoritmo de Porter para el español. Nos basamos en las pautas del trabajo que Grossman[188] realizó para las TREC-5, adaptando al español el conocimiento lingüístico, es decir, la lista de terminaciones y la de las palabras vacías. La lista que utilizamos tenía un total de 260 terminaciones; en ella se incluía tanto las terminaciones flexivas que expresan género y número como los sufijos derivativos con sus variantes alomórficas. No se tuvieron en cuenta desinencias verbales. La lista de palabras vacías, la elaboramos basándonos principalmente en la categorías gramaticales vacías de contenido (preposiciones, conjunciones...), aunque también se incluyeron numerales y algunos verbos, nombres y adjetivos de uso muy frecuente.

Para probar el experimento se crearon quince preguntas y se utilizó la base de datos DATATHÈKE[189], seleccionando de ella artículos exclusivamente en español y con una longitud mínima prefijada, evitando así documentos con un campo resumen poco significativo. Las relevancias se calcularon de manera

[187] R. GÓMEZ DÍAZ. Op. cit

[188] D. A. GROSSMAN et al (1996) op. cit.

[189] Esta base de datos se puede consultar en http://milano.usal.es/dtt.htm Contiene un vaciado de artículos de materias relacionadas con la biblioteconomía, la documentación, los archivos, la traducción y la informática.

manual analizando artículo por artículo para ver cuales de ellos respondían a las preguntas, que previamente habíamos elaborado.

Se realizaron cinco experimentos. Todos seguían el mismo proceso en la indización de las preguntas y los documentos de la base de datos. Las diferencias entre los distintos experimentos, está en si simplemente se suprimen los plurales y se normaliza el género (Este experimento se denominó *plurales*) o si además se suprimen las terminaciones de las palabras, coincidiendo con la lista de sufijos, establecida previamente, dejando un tamaño de raíz de 4 caracteres (*raíz 4*) o de 5 (*raíz 5*). Se hicieron dos experimentos como test de control, el primero sin eliminar las palabras vacías (*sin nada*) y el segundo eliminándolas (*sin vacías*). Una vez hecha la lematización, se indizó la base de datos, teniendo en cuenta cada uno de los procesos de lematización y estableciendo la comparación con la pregunta. En el análisis de los resultados obtenidos solo se tuvieron en cuenta los primeros 60 documentos recuperados. Con estos 60 documentos, se calculó la precisión y la exhaustividad según las ecuaciones de Salton[190], y se realizaron las curvas de precisión, exhaustividad y precisión-exhaustividad para establecer las comparaciones pertinentes.

[190] Ver parte dedicada a precisión y exhaustividad dentro de las medidas de Evaluación.

Los resultados se ven en en el siguiente gráfico.

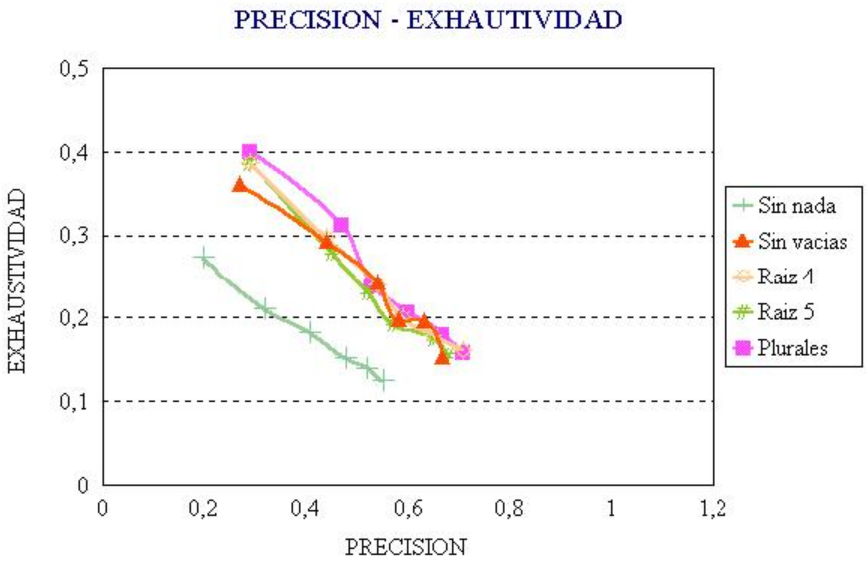


Gráfico 1 Resultados del Trabajo de Grado R. Gómez 1998

Según el gráfico, podemos ver cómo la simple supresión de las palabras vacías mejora los resultados frente al experimento que no las suprime. Los mejores resultados son los que simplemente suprimen plurales. En este caso se consigue una precisión media cercana al 0,4 para los primeros 10 documentos recuperados y una exhaustividad del 0,70 para los primeros 60 documentos recuperados.

Una de las conclusiones más importantes de este trabajo es que la simple supresión de las palabras vacías mejora la recuperación de información para el español, siendo éste es el primer trabajo para el español donde se demuestra esto. Otra de las conclusiones de este trabajo es que, para aumentar la exhaustividad de la búsqueda en español resulta mejor una simple supresión de desinencias flexivas que una supresión más radical. Respecto a esta conclusión, hay que considerar que

en este trabajo, no se trabajó con terminaciones verbales y es de esperar que los resultados varíen al tenerlas en cuenta.

La importancia de este estudio radica principalmente en que es el primer experimento realizado en España que aplica la lematización a la R.I. Su principal carencia es que este sistema no se compara con otro.

El otro trabajo realizado para el español[191], compara la R.I. basada en la lematización con la basada en n-gramas. El experimento de lematización es el visto anteriormente. En lo que se refiere a la aplicación de los n-gramas, se probó con distintos tamaños de n y se concluyó que para el español, los mejores resultados son los que utilizan un tamaño de n de 6 ó 7 caracteres. Este resultado era poco esperado ya que según los estudios hechos para el inglés, cuando el tamaño de n superaba el de 4 caracteres, la efectividad del sistema decaía. Este hecho viene a mostrar que los sistemas de recuperación no son independientes de la lengua de los documentos, premisa contraria a muchos de los trabajos realizados en las TREC.

De la comparación de los dos sistemas se puede concluir que los mejores resultados obtenidos con n-gramas son peores que los obtenidos con el mejor de los experimentos de lematización. Por lo que podemos afirmar que para la R.I. en español, es más adecuada la lematización, tanto la flexiva como la derivativa) que los n-gramas.

[191] C. G. FIGUEROLA (2000) op. cit.

Respecto a la aplicación de la lematización en español, queda por determinar cómo influyen los verbos, habrá que ver si lematizar los verbos mejora las tasas de precisión y exhaustividad en la recuperación.

II LA LEMATIZACIÓN

1. Introducción.

Los sistemas tradicionales de R.I. se basaban en la coincidencia entre los términos de la pregunta y los contenidos en el documento, (ya fuera en la indización o en el texto completo), y al hacer esto, se buscaban palabras exactamente iguales. El gran inconveniente de estos sistemas era que no se podían tener en cuenta las variantes morfológicas y, como afirma Hull[192], *"en la mayor parte de los casos estas variantes tienen interpretaciones semánticas que pueden tratar como equivalentes en aplicaciones a la R.I."* De aquí surge al interés por desarrollar algoritmos que reduzcan el número de variantes morfológicas de las palabras, y esto es precisamente lo que buscan los algoritmos de lematización.

2. Definición y problema de uso del término.

El término lematización no está admitido en el diccionario de la R.A.E., sin embargo lo vamos a emplear para expresar la acción de extraer la esencia, es decir, el tema o lema de una palabra. El término lema no está establecido en gramática sino que se trata de una etiqueta introducida por aplicaciones informáticas para hacer referencia a la base sobre la cual las distintas formas flexivas actúan. En inglés, esto se denomina "s*tem"*. La definición de *"stem"* hace referencia al tema de una palabra[193]. En español, la esencia de un término se suele obtener de la raíz. El lema aunque es próximo a la raíz y en algunos casos

[192] D. A HULL. Stemmings Algorithms: A Case Study for Detailed Evaluation J*ournal of the American Society for Information Science* 47 (1) 1996 p. 70-84

[193] Collins Spanish-English English –Spanish Diccionario español- inglés inglés- español. Colins Smith ed. Barcelon: Grijalbo, 1979 p 489.

coincide, no es siempre equivalente. Optamos por usar lema como base para crear el neologismo *lematización*.

Con la lematización pretendemos detectar las variantes tanto flexivas como derivativas de una palabra, tales como *bibliotecas*, *bibliotecario*, *bibliotecaria*, *bibliotecarios*... Esto puede implicar algoritmos más o menos complejos, desde los que simplemente juegan con las variantes de género y número (distinguiría entre *bibliotecario, bibliotecaria, bibliotecario*s y *bibliotecarias*), a los que combinan esto con una lista más o menos larga de sufijos y o prefijos (la lematización de estos últimos es poco frecuente). De este modo, todos los términos citados antes se reducirían al término *biblioteca*.

Normalmente los algoritmos de lematización no llevan asociado análisis morfológico, aunque contemplar esta posibilidad resolvería problemas de ambigüedad.

La lematización según Kroventz[194], puede verse desde distintos puntos de vista. Uno de ellos es como un método de expansión de preguntas: los términos contenidos en la pregunta inicial se sustituyen por sus lemas; de este modo aumenta el porcentaje de documentos con los que dicha pregunta puede casar. Otro punto de los puntos de vista está basado en la idea del cluster, donde cada uno de los cluster se basa en normas de conflación. El tercer punto de vista es que el que considera que la lematización puede utilizarse como normalización de los términos usados en la pregunta.

En función de estas posturas, los distintos autores han creado sus lematizadores, como veremos más adelante

---

[194] KROVENTZ Viewing morphology as inference process. *Proceedings of the 16 th ACM/SIGIR* Conference. New York: Association for Computing Machinery 1993 p. 191-202. [también en línea] Consultado el [20 12 1999] < ftp://ftp.cs.umass.edu/pub/techrept/techreport/1993/UM-CS-1993-036.ps>

Los primeros estudios sobre la lematización datan de los años 60[195]. En esta época el objetivo buscado se centraba en reducir el tamaño de los ficheros de índice, para de este modo reducir el espacio en disco; hoy en día, ésta no es la motivación que lleva a desarrollar estos sistemas, sino el mejorar la ejecución de la recuperación de la información al incrementar la exhaustividad de la misma.

Al hablar de lematización, hay que tener en cuenta que se trata de un proceso que intenta mejorar la ejecución de los sistemas de recuperación. No es un ejercicio etimológico o de gramática, ya que desde estos puntos de vista incurre en errores, sino que es un ejercicio de normalización lingüística mediante el cual las variantes de una palabra son reducidas a su forma común. En este sentido hay que tener en cuenta que el proceso de lematización no es perfecto.

La lematización se inscribe en el nivel morfológico[196] del procesamiento del lenguaje natural. Con este sistema, se trata de eliminar de manera automática los sufijos de las palabras, para obtener, en el mínimo de caracteres posibles, el máximo de información del término. De este modo podemos relacionar un término con otro u otros de su misma familia.

Para Lovins[197], un algoritmo de lematización es un procedimiento computacional por el cual se reducen todas las palabras a su forma común al quitar los sufijos de derivación y flexión. En este mismo sentido, Salton[198] afirma que una materia está representada por las primeras letras de una palabra, el final, representa la función sintáctica, por lo que si los finales se pueden cambiar, transformar o incluso suprimir, las formas variantes se pueden cambiar o transformar en formas más pequeñas, de manera que podremos reducir el tamaño

---

[195] C BELL and K.P JONES. Toward everyday languaje information retrieval system via minicomputer. J*ournal of the American Society for Information Science* 1979 30 334-338 op. cit

[196] La morfología es la sección de la lingüística que estudia y describe como están formadas las palabras en la lengua e incluye la inflexión, derivación y composición de las palabras.

[197] J. B. LOVINS (1968) op. cit.

[198] G. SALTON (1983) op. cit.

de los ficheros de índice, sin eliminar la esencia de cada palabra. Lo que se hace, es indizar por el lema de la palabra, de manera que al hacer la búsqueda se incrementa el éxito de casar correctamente los documentos con las preguntas. Al reducir la palabra a su lema correspondiente, podemos aumentar la exhaustividad en la recuperación, aunque al mismo tiempo reducimos la precisión de la búsqueda (recordemos que son medidas inversamente proporcionales). De cara a la R.I., la lematización es importante porque, en la mayor parte de los casos, la información que es semánticamente representativa para el usuario de un sistema, se contiene en los lemas del léxico del tema de interés, y los sufijos introducen modificaciones sutiles al concepto o simplemente expresan relaciones gramaticales que permiten que esta información se pueda expresar de manera gramaticalmente correcta. Por lo que si podemos controlar los lemas de las palabras tendremos más posibilidades de hacer una recuperación eficaz.

Al hablar de lematización es necesario introducir el término inglés "*conflation"*, que aquí está traducido por *conflación*[199], para referirnos a la representación de múltiples variantes morfológicas bajo una única forma (lema). Según esto, la lematización es una forma de conflactar automáticamente. La diferencia que existe entre una y otra es que para que se produzca lematización sólo se hace el tratamiento conjunto, es decir, sólo se hace la conflación con aquellas palabras que son semánticamente equivalentes, y además tienen en común el mismo lema, en cambio en la conflación éste último requisito no es necesario.

Para Paice[200] existen tres clases de relaciones entre palabras:

- Palabras con idéntica forma. Como es el caso de los homógrafos, (g*ato* (mamífero), *gato* (motor hidrálico).

---

[199] En español existe el término aunque ya en desuso con el significado de fundir. Diccionario de la leguna española.21ª ed. Madrid: R.A.E., 1992 p. 539

[200] C. D. PAICE Method for Evaluation of Stemming Algorithms Based on Error Counting. J*ournal of the American Society for Information Science* 1996 47 (8) p. 32-649.

- Palabras de diferente forma pero semánticamente equivalentes. Como son los sinónimos, (gato, minino).

- Palabras que son diferentes en forma y significado, (*gato, perro*).

Los casos interesantes para la lematización son los pertenecientes al grupo uno y algunos del dos.

El tratamiento conjunto que se hace al reducir términos distintos a formas comunes, es muy útil en R.I. para establecer la comparación entre el documento y la pregunta, ya que como indica Dawson[201], la conflación ahorra tiempo de indización, aunque hay que tener en cuenta que se producen errores que son difíciles de controlar, por lo que las palabras que han sido conflactadas deben usarse con cuidado.

También podemos considerar que la lematización es una manera de establecer un truncamiento por la derecha (ya que como veremos más adelante la lematización se aplica a los sufijos principalmente), aunque en el caso de la lematización la reducción a la forma común, es decir la conflación, se hace de manera automática, en función de una serie de criterios establecidos en reglas. En el caso de un simple truncamiento, el único criterio que se sigue es la longitud, bien sea del propio truncamiento o de la parte de palabra que queremos que quede. En este sentido, también hay que decir que los truncamientos en R.I., solamente se suelen aplicar a las preguntas, mientras que la lematización se aplica tanto a las preguntas como a los documentos.

[201] J. DAWSON. Suffix Removal and Word Conflation *Assoc Liter and Linguistic Computing Bulletin, Michelmas,* 1974 p. 33-46.

A pesar de la cantidad de experimentos que se han hecho con lematizadores aplicados a la R.I., Kraaij[202] señala que la eficacia del uso de la lematización en R.I. no está aún suficientemente demostrada, de este modo, mientras que Lenon[203] y Harman[204] afirman que la mera supresión de sufijos no produce una mejora notable en la ejecución de la recuperación, al menos para el inglés, Frakes[205] afirma que sí la produce; Popovic[206] también afirman lo mismo que Frakes, pero matiza que cuando se trata de idiomas con una complejidad morfológica mayor que la del inglés. Incluso Kroventz[207] que repite el experimento de Harman y obtiene los mismos resultados que ella, concluye que sí produce mejora. Hull[208] llega a la conclusión de que en general la lematización es siempre beneficiosa excepto en las preguntas largas con niveles bajos de precisión, pero no es capaz de mostrar diferencias significativas entre las técnicas simples de supresión de sufijos como son los algoritmos de Porter o Lovins y algoritmos con mayor conocimiento lingüístico. Niederman[209] sugiere que en su caso, la lematización produce mejoras muy importantes al aplicar la lematización al sistema MARS, el aumento de exhaustividad, compensa la caída de la precisión.

---

[202] W. KRAAIJ and R POLMANN. Viewing Stemming as Recall Enhacement [en línea] <http://rayuela.ieec.uned.es/~ircourse/doc/uplift/sigir96revised.ps> [consultado el 25-11-99]. Edición revisada de la de SIGIR'96.

[203] M. LENONE [et al] An Evaluation of same conflation algorithms for Information Retrieval *Journal of Information Science* 1981 3 p.177-188.

[204] D. HARMAN (1991) op. cit.

[205] W.B FRAKES Stemming Algorithms En W.B. FRAKES, R. BAEZA YATES (ed) *Information Retrieval: Data Structures and Algorithms*. Mexico: Prentice-Hall, 1992 p.131-161.

[206] M. POPOVIC and WILLETT The effectiveness of stemming for natural language access to Slovene textual data *Journal of the American Society for Information Science* 1992 43(5) p. 384-390.

[207] KROVENTZ (1993) op. cit.

[208] D. A. HULL (1996) op. cit.

[209] G.T. NIEDERMAN, G. THURMAIR & J. BÜTTEL MARS: a retrieval tool on the basis morphological analysis En C.J. van Rijsbergen (ed.). *Research and development in information retrieval*. Cambridge: C.U.P., 1985 citado por POPOVIC (1992) op. cit.

En la valoración de los resultados hay que tener en cuenta que las colecciones empleadas no son siempre las mismas con lo cual la comparación no es del todo válida.

A la hora de establecer lematización hay que tener en cuenta una serie de aspectos, como son:

- La complejidad morfológica de la lengua en la que se hace la lematización[210]. Así, en lenguas morfológicamente poco complejas como el inglés, la mejora de la exhaustividad no compensa por la pérdida de precisión; en cambio, en lengua mucho más complejas como es el esloveno[211] o el griego[212], sí produce mejora.

- La longitud de los documentos, ya que para Krovetz los mejores resultados son los obtenidos con documentos y preguntas cortas.

- El tamaño de la colección[213].

[210] W. KRAAIJ and R. POHLMANN Evaluation of Dutch stemming algorithm [en línea] <http://rayela.ieec.uned.es/~ircourse/doc/uplift/ > [consultado el 25-11-99]

[211] C. JACQUEMIN, and E. TZOUKERMAN NLP for term variant extraction: synergy between morphology, lexicon and Syntax. EN T. STRZALKOWSKI (ed) *Natural Language Information Retrieval*. Dordrcht: Kluwer Academic Publisher, 1999. p 25-75

[212] T.Z. KALAMBOUKIS. Suffix stripping with moderm Greek. *Program* 29 (3) 1995 p. 313-321.

[213] J. SAVOY A Stemming Procedure and Stopword List for General French Corpora. J*ournal of the American Society for Information Science* 50 (10) 1999 p. 944-952.

3. Tipos de algoritmos de lematización: clasificaciones.

Los algoritmos de lematización se pueden clasificar atendiendo a distintos criterios, como son el tipo de sufijos que trata, el modo de establecer la lematización o la conflación y el grado de conocimiento lingüístico que tienen.

3.1 Lematizadores simplemente flexivos y algo más que flexivos.

Esta es la clasificación que hace Harman[214]. Esta autora distingue entre los algoritmos que tan solo hacen lematización de plurales y algunas formas verbales como es su *"S Stemmer"* o el *"K-Stemmer"* de Kroventz, el *"Weak Stemmer"* de Savoy y el de Walker y Jones[215], entre otros, frente a otros esquemas más sofisticados que abarcan más sufijos. Los primeros, simplemente suprimen la "s" o "es" que marcan los plurales para hacer que la palabra pase al singular. Así para el español un sistema de este tipo en una palabra como *pies* suprimiría simplemente *s* para dejar *pie* mientas que en la palabra *reyes* eliminaría *es* para dejar *rey*. La complejidad de los algoritmos del segundo tipo es muy variada.

[214] D. HARMAN (1991) op. cit

[215] S. WALKER & R. M. JONES. Improving subject retrieval in on line catalogues: 1 stemming, automatic spelling correction and cross-references tables. British Library Reseach Paper 24 London. London Britihs Library. Citado por Popovic (1992) op. cit.

3.2 Cómo establecen la lematización.

Esta es la clasificación que hace Lovins[216]. Para este autor, hay dos tipos de algoritmos, los que se basan en la **iteración** y los que se basan en el **sufijo más largo**. En los primeros, los sufijos pertenecen a ciertas categorías y se pegan a la raíz en función de éstas. La ventaja de los algoritmos que trabajan con las categorías, es que requiere una lista más corta de sufijos que otros modelos.

El otro sistema del que habla este autor es el basado en **el sufijo más largo,** aquí los sufijos pertenecen a una única clase. Se ordenan en función de su longitud, de mayor a menor. En este caso no se requiere ningún conocimiento lingüístico previo sobre las distintas combinaciones de los finales[217].

Tanto un sistema como el otro puede obedecer a uno o a varios de los siguientes criterios para encontrar el mejor lema:

- Modo cuantitativo: la longitud del lema debe exceder de un número determinado.
- Modo cualitativo: el final del lema debe satisfacer una condición.
- Reconociendo normas: el ajuste se hace para mejorar la combinación de los lemas producidos por los algoritmos de supresión de sufijos.

[216] LOVINS, J.B. (1968) op. cit.

[217] R.R. KORFHAGE. *Information Storage and Retrieval*. New York. Wiley Computer publishing, 1997. p. 136.

A esta clasificación, Lovins añade dos atributos al contexto en el que se encuentra el sufijo, que puede ser **libre**, cuando no hay restricciones cuantitativas o cualitativas a la hora de establecer la supresión de finales, o de **contexto sensible**, cuando sí las hay. Generalmente el contexto suele ser sensible, aunque la única limitación que se use sea la de longitud mínima del lema final, teniendo en cuenta que en ningún caso el lema resultante podrá tener longitud cero. Este tipo de algoritmos requieren un conjunto de reglas.

3.3 Por el modo de establecer la conflación.

Esta es la clasificación de Frakes[218]. Según este autor la conflación se puede establecer de las siguientes formas.

1. Manual

2. Automática: (lematizadores)

    2.1. Supresión de afijos

        2.1.1. Coincidencia más larga

        2.1.2. Simple supresión

    2.2. Sucesor de variedad

    2.3. Tabla de búsquedas

    2.4. N-gramas

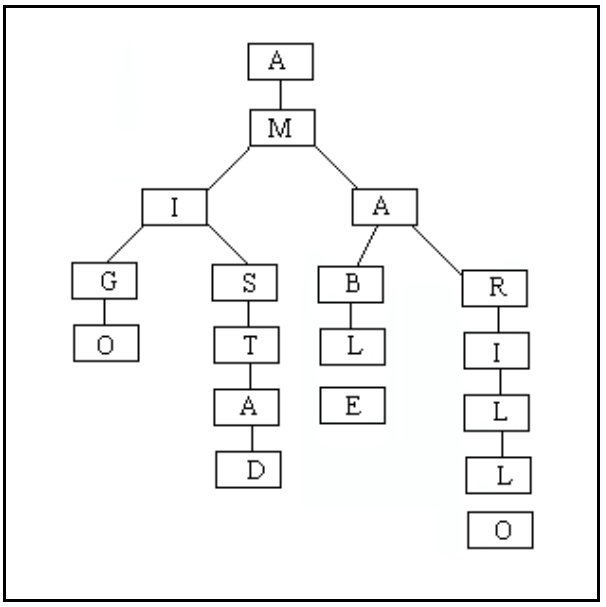
A continuación nos limitaremos a comentar los métodos automáticos:

[218] FRAKES (1992) Stemming ... op. cit.

2.1 La **supresión de afijos** busca eliminar los sufijos y/o prefijos de los términos, dejando su raíz. Estos algoritmos a veces también transforman el resultado de la raíz. Este método es uno de los más comunes. Dentro de aquí se puede jugar con los tamaños de la raíz y los sufijos. Así tendremos algoritmos que simplemente supriman los afijos (generalmente sufijos) o otros más complejos que también incluyan una serie de condiciones y reglas que se deben cumplir a la hora de ejecutarlos.

2.2 El **sucesor de variedad**. Ha sido investigado por Hafer y Weis [219]. En este sistema se usan las frecuencias de las secuencias de letras en el cuerpo del texto como base de lematización. El sucesor de variedad es el segmento de una palabra en un conjunto de palabras. Es el número de letras distintas que ocupa la longitud del segmento más un carácter. Un sucesor de variedad lo que hace es dividir la palabra en segmentos, y seleccionar uno de ellos como lema. La representación gráfica se hace en forma de árbol. Así por ejemplo para las palabras: *amigo, amistad, amarillo* y *amable* la representación sería:

[219] M. HAFER, and S. WEIS, Word Segmentation by Letter Succesor Varities *Information Storage and Retrieval*, () 10 197 p. 4371-385 citado por W.B. FRAKES (1992).



Dibujo 5 Sucesor de variedad

Por ejemplo el sucesor de variedad para las tres primeras letras de una palabra de cinco letras es el número de palabras que tienen las mismas tres letras pero con la cuarta diferente. En el caso del ejemplo "AM" lleva asociados dos hijos, luego el sucesor de variedad para "AM" es de dos y para "AMI" también es de dos.

Los sucesores de variedad tienen distintos métodos de trabajo:

2.2.1 Método de cierre: se selecciona un valor de corte que define la longitud del lema. Este valor varía para cada posible conjunto de palabras. El problema es cómo seleccionar este valor de cierre ya que si es demasiado pequeño o demasiado largo el corte será incorrecto.

2.2.2 Método de punta y altiplanicie: se parte un segmento, después un carácter cuyo sucesor de variedad exceda del carácter inmediatamente precedente y del carácter inmediatamente siguiente. Este método quita los valores de corte.

2.2.3 Método de palabra completa: se hace una ruptura después de un segmento determinado.

2.2.4 Método de entropía: tiene la ventaja de la distribución del sucesor de variedad. En este método mediante una Ecuación de búsqueda se selecciona un valor de búsqueda y se identifica un lindero siempre que se busca el valor de corte.

2.3 **Tabla de búsquedas**: este sistema consiste en introducir los términos conflactados en una tabla y hacer las búsquedas a través de aquí. Esto también se conoce como diccionario lematizador. Más que un sistema en sí, que también puede serlo, es un complemento a cualquier sistema de lematización. Normalmente las tablas de búsquedas es la parte del programa donde se archivan los resultados de la lematizaciones hechas previamente.

2.4 **los n-gramas**[220]: Frakes los incluye dentro de la lematización puesto que las palabras derivadas tendrán siempre una serie de n-gramas comunes con la palabra de la cual derivan, variando simplemente los que están fuera de la raíz, aunque este sistema no es estrictamente un lematizador porque simplemente juega con las cadenas de caracteres y no tiene en cuenta ninguna información semántica.

3.4 En función del conocimiento lingüístico.

Otra de las clasificaciones que podemos hacer es en función del conocimiento lingüístico que el sistema de lematización aplica. Según esto,

[220] Ver dentro de las aplicaciones del P.L.N. a la R.I. el apartado dedicado a los n-gramas.

podemos dividir los lematizadores en dos grandes grupos: los **algoritmos sencillos** y **los que aplican un mayor conocimiento lingüístico**.

Dentro de los primeros, están los métodos que simplemente aplican técnicas de corte sin tener en cuenta reglas que limiten el contexto donde se debe aplicar ese corte, o el número de reglas es muy pequeño. Dentro de este grupo está el *"S Stemmer"* de Harman.

Los algoritmos que aplican un mayor conocimiento lingüístico como son aquellos trabajos donde se intentan aplicar un mayor número de reglas o se establece la aplicación de las mismas en función de una serie de criterios morfológicos como es el caso del lematizador de Savoy[221] o el que nosotros propondremos más adelante.

En principio parece que esta clasificación que proponemos nosotros coincide con la de Harman, pero no es así. Aunque en la mayor parte de los casos la lematización flexiva requiere algoritmos más sencillos, no siempre es así ya dependiendo de la complejidad morfológica de la lengua que trate, el lematizador puede requerir un conjunto amplio de reglas, que delimiten el contexto donde se puede aplicar la supresión de sufijos, y sería un algoritmo con conocimiento lingüístico.

4. La necesidad de lematizar.

La necesidad de lematizar surge de intentar solucionar el problema de tener muchos términos distintos que expresen un mismo concepto; esto que es propio del lenguaje natural, aplicado a la R.I. es un problema. La aplicación de la

[221] J. SAVOY (1999) op. cit.

lematización a la R.I. nace la búsqueda de métodos que mejoren las tasas de exhaustividad de la recuperación en los sistemas que trabajan con lenguaje natural. Resulta obvio decir que las palabras que pertenecen a una misma familia están semánticamente relacionadas, pero estas relaciones no se ponen de manifiesto de manera automática en los sistemas convencionales de R.I. En un principio, como comenté al comienzo del capítulo, el objetivo de estos sistemas era reducir el número de palabras únicas en la indización, con el fin de disminuir el espacio ocupado en el disco[222], en un intento de ahorrar recursos. Hoy en día, dadas las capacidades de los ordenadores, el espacio en disco no tiene tanta importancia, pero sí el trabajar con listas más reducidas, porque implica un ahorro de tiempo no sólo en las tareas de indización, sino también en la consulta, al tener que hacer menos preguntas (las relaciones semánticas entre los términos derivados se hacen automáticamente).

La motivación más importante que los lleva a investigar en la lematización está relacionada con la frecuencia de aparición de los términos en los documentos. Normalmente cuando hacemos un estudio de frecuencias consideramos toda la cadena de caracteres, es decir que si en un texto aparece el mismo término en singular y en plural consideramos que son términos distintos en un lugar de uno solo, lo que puede hacer que se pierdan documentos relevantes. De este modo los estudios de frecuencia no están siendo fieles a la realidad, por lo que los sistemas basados en ellos arrastrarán un error que puede ser subsanable con algún método de normalización. La lematización puede ser uno de ellos ya que lo que precisamente propone es unir bajo un único término, palabras con un origen común, basándose en que todas guardan relación semántica.

[222] C BELL and K.P. JONES, (1979) op. cit.

5. Problemas de la lematización.

El primer inconveniente que presenta la lematización es inherente al propio lenguaje natural. De este modo, el mismo problema que presentan los homógrafos en cualquier contexto, se traslada aquí. Este problema puede mejorar al introducir un análisis morfológico y o sintáctico del término, ya que el distinguir las categorías gramaticales y las funciones dentro de cada oración, nos puede ayudar a establecer distinciones de significado, aunque esto no es siempre útil. Hay que tener en cuenta que hay parejas de palabras que etimológicamente están relacionadas pero con diferencias marcadas en su significado, como es el caso de *autoritario* y *autor*. También hay que tener en cuenta que añadir sufijos en ocasiones implica transformaciones irregulares, por ejemplo por el contexto donde está el sufijo y no siempre es fácil controlar esto mediante reglas.

El resto de los errores que se producen con la lematización, son propios del sistema empleado. Como señala Lovins, al manejar la palabra de manera individual se pierde toda la información sobre las relaciones gramaticales y semánticas. También los algoritmos pueden generar lemas lingüísticamente incorrectos. No podemos olvidar que hay veces que existen palabras distintas con raíces gráficamente iguales, por lo que si no somos capaces de controlar estas formas, estaremos introduciendo ruido en la recuperación. Otro de los errores es que al ganar exhaustividad perdemos precisión. En este sentido hay que procurar que la mayor exhaustividad compense esta pérdida de precisión.

Existen también errores derivados de cortar incorrectamente las palabras, uno es la infralematización, es decir, cuando cortamos un sufijo demasiado corto, no recuperamos todo lo esperado por lo que tenemos silencio informativo, o lo que es lo mismo, tasas bajas de exhaustividad. También puede ocurrir lo contrario si quitamos una parte demasiado larga (sobrelematización), es decir, dejamos un lema muy corto que coincidirá con el lema de gran número de palabras, cayendo

de este modo el índice de precisión. Dentro de la infralematización, Savoy[223] señala un tipo especial de error que es el que se da cuando se superponen unos caracteres en lugar de otros, y por tanto no se encuentra la raíz correcta. Por su parte, Korfhage[224] señala dos problemas. El primero de ellos es cuando las secuencias de letras parecen falsos finales. Por ejemplo en español la terminación -*able* en la palabra *sable* no contiene el sufijo -*ble*. En este caso el error se podría subsanar de tres maneras distintas, la primera es haciendo un análisis morfológico. La terminación -*ble* es propia de adjetivos, y *sable* es un nombre, luego no puede contener dicho sufijo, por lo tanto una vez analizada la palabra, el programa detectaría que dicha palabra no tiene que ser lematizada. Otra solución es introducir este término en una tabla de búsqueda, poniendo que el lema de *sable* es *sable*. Y el tercer método es poniendo una regla que indique una mínima longitud del lema resultante. Aunque para este caso hay tres posibles soluciones, no siempre es viable llevarlas a cabo.

El otro problema que señala Korfhage afecta a las palabras que tienen derivación irregular. La mejor solución para esos casos es incluir estas formas en una lista de excepciones. También puede darse el efecto contrario y no quitar un sufijo cuando realmente lo es. La solución a este problema suele ser más compleja.

Ahmad[225] habla de errores derivados de las irregularidades ortográficas, en su caso, del árabe, y también de aquellos errores que se producen al no aplicar las reglas en un orden determinado, estos errores están muy relacionados con el idioma del lematizador y en función de éste tendrán una importancia mayor (caso del malayo) o menor.

---

[223] J. SAVOY A Stemming Procedure and Stopword List for General French Corpora. J*ournal of the American Society for Information Science* 1999 50 (10) p. 944-952.

[224] R. R. KORFHAGE op. cit (1997)

[225] F. AHMAD, M. YUSOFF, T. Sembok Experiments with a Stemming Algorithm for Malay Word *Journal of the American Society for Information Science* 1996 47 (12) p. 909-918

Korventz[226], señala que el problema de la lematización es que al lematizar perdemos parte del sentido, por ejemplo si *"gravitatorio"*, que se refiere a la *fuerza de la gravedad*, lo reducimos a *grave*, con el sentido de *serio* estaríamos conflactando bajo un mismo término conceptos distintos, en este sentido el problema de los homógrafos se acentúa.

6. Principales algoritmos de lematización el para el inglés.

Los primeros estudios sobre la supresión de sufijos datan, como indicamos anteriormente, de la segunda mitad de los años 60[227], en el artículo de Lovins[228] es donde aparece el primer algoritmo de lematización propiamente dicho. Este trabajo, está en la base de los algoritmos diseñados por Salton[229], Dawson[230], Porter[231], y Paice[232]. Todos estos algoritmos se hicieron para el inglés y se basan en que hay que establecer un conjunto de terminaciones y otro de reglas, con el fin de hacer un tratamiento común para conjuntos de lemas y así mejorar la R.I.. Las diferencias entre unos y otros están en el número de sufijos y de reglas. Según la comparativa que hace Harman[233] de los algoritmos de Porter,

---

[226] KROVENTZ (1993) op. cit.

[227] C.f. RESNIKOFF and DOLVI The Nature of Affixing in Written English Part I *Mechanical Translation*, 1965 8 nº 3. y Part II Mechanical Translation and Computational Linguistic vol 9 n. 2 (1966) citado por LOVINS (1968) y DAWSON (1974) op. cit.

C.f. M. KAY and G.R. MARTINS, The MIND System: the Morpholocical. Analysis. *Program*, US Air Force Proyect Rand Report R.M. 6265/2 PR April (1970) citado por Dawson (1974)

[228] C.f. J. B. LOVINS (1968) op. cit.

[229] G. SALTON (1968) op. cit.

[230] J. DAWSON (1974) op. cit.

[231] M. F. PORTER. (1980) op. cit.

[232] C. PAICE Another Stemmer *ACM SIGIR Forun*, 1990 24 (3) p. 56-61

[233] D. HARMAN (1991) op. cit.

Lovins y un simple algoritmo de supresión de finales, el de Lovins es el que hace la reducción más fuerte. Frakes afirma que el de Porter es más compacto que el de Lovin, Salton y Dawson y según Paice[234] la ejecución de la recuperación del algoritmo de Lovins es comparable a la de algoritmos mayores.

Hemos incluido en este parte el algoritmo de Kroventz[235], porque este autor hace lematización flexiva y con ella obtiene resultados bastante buenos y nos ha parecido adecuado compararlo con los algoritmos clásicos.

6.1 Algoritmo de Lovins

En el algoritmo de Lovins, hay una lista de 260 finales de palabras divididos en once conjuntos ordenados por tamaño. Cada subconjunto va precedido por una categoría especial donde se da la longitud de los finales. Una vez que se han suprimido los sufijos, se compara la raíz con 34 reglas, entre las que se encuentran algunas restricciones relativas al contexto en el que los sufijos se deben mantener o suprimir, por lo que según la clasificación que este autor hace de los algoritmos[236] el suyo pertenece a los de *contexto sensible.* La lematización la hace en dos pasos, primero suprime los plurales, después del sufijo más largo de este modo se evita duplicar todos los sufijos con el consiguiente ahorro de espacio y de tiempo que ello implica y después el resto de los sufijos en función de las reglas.

[234] C. PAICE (1996) op cit.

[235] KROVENTZ (1993) op cit.

[236] Vid supra

6.2 Algoritmo de Salton

El SMART de Salton es una versión más potente del lematizador de Lovins[237]. Utiliza algunos sufijos diferentes de los de Lovins. Actúa del mismo modo que un algoritmo de supresión "S" aunque con un nivel de complejidad mayor. Quita el sufijo más largo posible. El lema restante es chequeado de nuevo para ver si se puede volver a suprimir algún sufijo más. El lema final tiene que cumplir una serie de condiciones relacionadas con la longitud del lema resultante. Al igual que el algoritmo de Lovins trabaja con 260 finales y además tiene una larga lista de excepciones.

6.3 Algoritmo de Dawson

El algoritmo de Dawson está basado en el de Lovins, aunque es más complejo que éste. A la lista inicial de 260 sufijos de Lovins, le añadió algunos otros y la completó al incluir todas las variantes hasta conseguir una lista con 1.200 finales, agrupados en 55 grupos. No se tienen en cuenta las flexiones verbales irregulares ya que ello dispararía el número de reglas con las que habría que trabajar. El orden que sigue en el chequeo es el mayor a menor, utilizando una serie de condiciones en función de las que un determinado sufijo se suprime o no. El código 0 implica que el sufijo puede ser suprimido siempre que la longitud del lema resultante sea de dos caracteres. Las condiciones básicas son: un número que representa la longitud mínima que debe quedar después de suprimir un sufijo, una letra o un grupo de ellas que indica que se debe quitar o no el sufijo dada una determinada cadena de caracteres. Todo esto se combina mediante paréntesis y operadores booleanos.

[237] D. HARMAN (1991) op. cit.

Ej:

(not m) and ((3 and nor (s ** or u) or 4 (and nor (a or e or o)))

No se suprime el sufijo después de *'m"* y si la longitud mínima es de 3 caracteres no se suprime el sufijo *s*** ó *u* , o si la longitud mínima es de 4 caracteres no se suprime el sufijo después de *a*, *e*, u *o*.

6.4 Algoritmo de Porter

El algoritmo de Porter[238] uno de los más importantes y también el más conocido por la repercusión que ha tenido su adaptación para crear los lematizadores en otros idiomas distintos del inglés, como es el caso del esloveno[239], el latín[240], el holandés[241], el español[242]... Porter demostró cómo su algoritmo de supresión de sufijos mejora frente a otros sistemas más complejos la ejecución de la recuperación en términos de exhaustividad. Frakes[243] demuestra que el algoritmo de Porter, da resultados comparables a los truncamientos manuales.

Tiene dos consideraciones importantes, una es que el sufijo que se elimine sea siempre el más largo y que la palabra cortada mantenga una

---

[238] Cf. *The Porter Stemming Algorithms* [en línea] <http://www.muscat.com/~martin/stem.html> [Consultado el 20/07/00]

[239] M. POPOVIC (1992) op. cit. 1992

[240] R. SCHINKE, M. GREENGRAS, M. A. ROBERTSON, P WILLETT. A Stemming Algorithm for Latin text databases. *Journal of Documentation* 1996 52 (2) p.172-187

[241] W. KRAAIJ, and R. POLMANN *Porter´s Steming algorithm for Dutch*. In L. Noordman and W. De Uroomen (eds) Informatiewetenschap 1994: Wertenschapplijke bijdragen aan de derde STINFON Conferentie p. 167-180 [on line] http://rayuela.ieec.uned.es/~ircourse/doc/uplift/ [Consultado el 25 del 11 de 1999]

[242] F KELLEDY (1995) op. cit.

[243] W. B. FRAKES Term Conflation for information retrieval C.J. van Rijsbergen (ed.). *Research and development in information retrieval*. Cambridge: C.U.P., 1984 p. 383-390 citado por POPOVIC (1992) op. cit.

determinada longitud. Teniendo esto en cuenta, cada palabra puede ser lematizada tantas veces como se considere necesario. El algoritmo se basa en:

La medida (m) de la raíz, se basa en la alternancia de vocales (a, e, i, o, u, y (en el caso de que vaya precedida de una consonante)) y consonantes, (todas las letras que no son vocales)

La alternancia de vocales y consonantes se expresa:

$[C] (VC)^m [V]$

donde

[C] : Consonante susceptible de aparecer

(VC) : Conjunto de vocal/consonante

m : medida de cada palabra o parte de palabra

[V] : vocal susceptible de aparecer

Cuando m es igual a 0 la palabra es nula.

Condiciones de la raíz

*<x> la raíz termina con la letra x

v la raíz contiene una vocal

*d la raíz termina en doble consonante

*o la raíz termina con una secuencia del tipo consonante-vocal-consonante, donde el final de la consonante no es *w*, *x*, o *y*

Estas condiciones se pueden combinar entre sí y con la longitud de m mediante operadores booleanos.

Reglas:

$S_1 \longrightarrow S_2$

Si una palabra termina con el sufijo 1, el lema que antes era S_1 debe satisfacer una serie de condiciones, en este caso se podrá reemplazar con S_2. La condición general está relacionada con la medida de m.

Las reglas se dividen en pasos que definen el orden en que se aplican dichas reglas.

El algoritmo de Porter está a caballo entre la simple supresión de plurales y el algoritmo de Lovins. La diferencia principal con el Lovins es que en lugar de comenzar por suprimir el sufijo más largo, suprime los sufijos cortos en varios pasos. Reconoce menos sufijos y no tiene lista de excepciones.

Este algoritmo sustituye con éxito el análisis morfológico en el inglés para establecer los índices en R.I., sin embargo no es fácil encontrar algoritmos simples y con éxito para lenguas más flexivas.

Los problemas que presenta son que es poco difícil de entender y de modificar. También tiene problemas de infralematización y sobrelematización derivados de la mala realización de la conflación. A pesar de esto la evaluación de la precisión y la exhaustividad son altas

6.5 Algoritmo de Kroventz

Kroventz parte de que la ejecución de la recuperación puede mejorar indizando los documentos por significados, no por palabras, ya que cuando el sentido de las palabras en el documento no coincide con el de las preguntas,

aunque se recupere el documento, este será irrelevante. Para este autor, Porter no ha tenido en cuenta esta idea.

El algoritmo de Kroventz es sólo para terminaciones flexivas, actúa en tres pasos: primero convierte los plurales en singulares, convirtiendo los participios (terminación en *-ed*) a presente y eliminando la forma del gerundio (*-ing*). Dentro de los plurales para el inglés hay tres formas, *-ies* que se remplazará por *y* ; *-es* que se suprime la *s* y *–s* que se suprime la s siempre que el final no sea *–ous* o doble *s*

En los tres casos se chequea que el resultado coincida con la entrada de un diccionario. Con este lematizador se incrementan los niveles de exhaustividad y se mejora la precisión. Kroventz pretendía que los resultados fueran mejores que un simple truncamiento, que la conflación se hiciera solamente con las palabras pertenecientes al grupo uno, según la clasificación de Paice[244], relacionadas, con la mayor cobertura posible, con unos resultado al menos tan buenos como los de Porter y que además desambiguara los términos.

Este autor compara los resultados obtenidos con los obtenidos por Lovin y Porter concluyendo que simplemente eliminando los sufijos flexivos los resultados son mejores.

6.6 Comparación de algoritmos para el inglés

En la siguiente tabla mostramos la comparación de los algoritmos más importantes realizados para el inglés:

[244] Vid supra.

	LOVINS	SALTON	DAWSON	PÒRTER	KROVENTZ
AÑO	1968	1968	1974	1980	1993
FINALES	260 finales	260 finales	55 conjuntos de finales	60 finales	3 finales
Nº REGLAS	34	No especifica	No especifica	No especifica	No especifica
PRINCIPIO	Suprime el sufijo más largo	No especifica	Suprime el sufijo más largo	Suprime plurales y después el sufijo más largo	De plural a singular; de participio y gerundio a infinitivo
LONGITUD MÍNIMA	Juega con diversos tamaños	2 caracteres	2 caracteres	Al menos un conjunto formado por una vocal y una consonante	No
TIPO DE CONTEXTO	Sensible	No especifica	Sensible	Sensible: restricciones cuantitativas y cualitativas	No
PASOS QUE SIGUE	Dos: suprime primero los plurales y después el resto de las reglas.	No especifica	No especifica	Suprime plurales y después el sufijo más largo	No
DICCIONARIO	No	No especifica	Sí	Sí	No
ITERATIVO	SI			Sí	Sí

Tabla 10 Comparación de algoritmos de lematización para el inglés.

7. La lematización en otros idiomas distintos del inglés.

Aparte de los algoritmos para el inglés, se han ido desarrollando, principalmente en la década de los 90, una serie de algoritmos para idiomas distintos del inglés. Aunque hay más trabajos, hemos elegido aquí solo un grupo representativo de ellos. La principal diferencia que hay entre ellos son las propias características de los idiomas, y aunque todos coinciden en que son lenguas morfológicamente más complejas que el inglés, entre ellos hay distinto grado de complejidad. De este modo, el holandés y el esloveno[245] tienen tres desinencias distintas para expresar los géneros (masculino, femenino y neutro) una más que en francés. En holandés, latín, esloveno griego y árabe, hay casos agrupados en declinaciones para expresar la función sintáctica. La mayor complejidad morfológica del francés respecto del inglés, es que los nombres, adjetivos, pronombres y determinantes, se flexionan no sólo para establecer la distinción de número, como en el resto de los idiomas, sino que también para expresar el género, al igual que ocurre en español. Parte de la complejidad del programa para el árabe es el sentido de la lectura (derecha-izquierda). El malayo, tiene cuatro clases de afijos (prefijos, sufijos, pares de prefijos-sufijos e interfijos). En cuanto al número, todas las lenguas aquí elegidas, distinguen entre singular y plural, el esloveno y el árabe también utilizan el dual para los nombres, adjetivos y verbos. Esto multiplica el número de desinencias a tener en cuenta. Esta misma información la hemos resumido en la siguiente tabla:

[245] F. JAKOPIN *¿Quieres hablar esloveno?*. Ljbliana: Slovenska izlejeniska matica, 1962

	HOLANDÉS	FRANCÉS	LATÍN	ESLOVENO	GRIEGO	MALAYO
GÉNERO	Masculino y femenino y neutro	Masculino y femenino	Masculino y femenino y neutro	Masculino y femenino y dual	Masculino y femenino y neutro	Masculino y femenino
NÚMERO	Singular y plural	Singular y plural	Singular y plural	Singular, dual y plural	Singular y plural	Singular y plural
DECLINACIONES	Sí	No	Sí	Sí	Sí	No
TIPOS DE AFIJOS	Prefijos, sufijos, interfijos	Prefijos y sufijos	Prefijos y sufijos	Prefijos y sufijos	Prefijos y sufijos	Prefijos, sufijos, interfijos, pares de sufijos-prefijos

Tabla 11 Comparación de los idiomas

Dadas estas características tan distintas de las lenguas, y una vez analizados los artículos donde se estudian los algoritmos para la lematización en francés[246], holandés[247], esloveno[248], malayo[249], árabe[250], latín[251] y el griego[252], y viendo que funcionan, con mayor o menor efectividad pero en general con buenos resultados, cabe pensar que en español puede funcionar un algoritmo de lematización adaptado a las características morfológicas del español.

El problema que presentan estos trabajos para su comparación es, aparte de las propias características de cada idioma, como acabamos de mostrar, que no en todos los trabajos se da el mismo nivel de detalle de cómo funcionan o cómo han creado los algoritmos. Aún así intentaré establecer una comparación entre ellos.

Los algoritmos para el esloveno, holandés, y latín se basan el trabajo de Porter, para los dos últimos, concretamente en el esquema que aparece en Frakes[253]. En el caso del francés, malayo y el árabe no se especifica que se basen en Porter, aunque coinciden con el esquema de este autor ya que buscan de la terminación más larga según una lista establecida previamente. De lo artículos aquí seleccionados, el trabajo para el griego es el único que se basa en el SMART de Salton.

---

[246] J. SAVOY (1993) op. cit.

[247] W. KRAAIJ (1994) op. cit

[248] M. POPOVIC (1992) op. cit.

[249] F. AHMAD, M. T. YUSO (1996) op. cit.

[250] H. ABU-SALEM, M. AL-OMARI, M. EVENS, Stemming Methodologies Over Individual Queri words for an Arabic Information Retrieval System *Journal of the American Society for Information Science* 1999 50 (6) p. 524-529

[251] R. SCHIKE (1996) op. cit.

[252] T.Z. KALAMBOUKIS. (1995) op. cit.

[253] FRAKES (1992) (b) op. cit

Todos trabajan con sufijos, pero en el caso del holandés también lo hace con algunos prefijos e interfijos para los verbos; el del francés es el único que trabaja con algunos prefijos (para los casos de términos científicos). En este sentido el trabajo para el malayo es el más complejo, ya que trabaja con los cuatro tipos de afijos que hay en dicha lengua. El latín añade una novedad respecto a los anteriores y es que trata algunos sufijos enclíticos.

El número de sufijos y reglas que contemplan, no aporta nada a la comparación puesto que esto va en función de la propia estructura morfológica de cada idioma.

Respecto a la evaluación, el holandés, el malayo y el latín evalúan el lematizador y el esloveno, el francés, el griego y el árabe evalúan la aplicación de la lematización a la recuperación. En este último caso, sería bueno que estos sistemas se compararan con otros sistemas, por ejemplo con métodos de recuperación booleana, de n-gramas o métodos probabilísticos, pero ninguno de los trabajos lo hace, simplemente comparan con no aplicar lematización a la recuperación; en este sentido, la ausencia de colecciones como la de Crandfield para el inglés hace que sea más dificil la comparación, ya que hay que invertir tiempo en la creación de la colección y establecer los juicios de relevancia. Popovic para comparar su algoritmo lo que hace es traducir el corpus de textos del esloveno al inglés y aplicar el algoritmo de Porter para establecer una comparación entre los dos. En el trabajo de Savoy lo que se hace es comparar un lematizador leve (*weak stemmer*), parecido al *S Stemmer* que propone Harman, con su algoritmo basado en categorías gramaticales. En el trabajo para el griego lo que se hace es aplicar el mismo algoritmo a dos colecciones distintas. En el

primero de sus experimentos[254], y el otro lo que hace es comparar varios experimentos para el griego[255].

En la siguiente tabla vemos la comparación de los algoritmos aquí reseñados. Hay que tener en cuenta las diferencias de los idiomas.

[254] T.Z. KALAMBOUKIS (1995) op. cit.

[255] S. NICOLAIDIS, T. Z. KALAMBOUKIS Evaluation of stemming algorithms with moder greek . (en prensa)

	HOLANDÉS	ESLOVENO	FRANCÉS	MALAYO	ARABE	LATIN	GRIEGO	ESPAÑOL	
								DER	FLE
AÑO	1994	1992	1993/1999	1993	1992/1994/1995	1996	1995	2001	
BASADO	Porter (Frakes)	Porter	No especifica	No especifica	No especifica	Porter (Frakes)	SMART (Salton)	Porter	
TRATA	Prefijos Sufijos infijos	Sufijos	Prefijos Sufijos	Prefijos Sufijos Prefijos Sufijos infijos	Sufijos	Sufijos Sufijos endíticos	Sufijos	Sufijos	
NÚMERO DE SUFIJOS	No especifica	5276	No especifica	No especifica	No especifica	90 de nombres y adjetivos 26 de verbos	5 flexivos	230 flex y der	88 flex
NÚMERO DE REGLAS	98	No especifica	35	121	No especifica	No especifica	No especifica	3692	2700
EVALUACIÓN	Según Paice. Evalúa el algoritmo pero no su aplicación a la R.I.	Salton. Evalúa la aplicación a la recuperación	Salton. Evalúa la aplicación a la recuperación	Según Paice. Evalúa el algoritmo pero no su aplicación a la R.I.	Salton. Evalúa la aplicación a la recuperación	No especifica	Salton. Evalúa la R.I.	Salton Evalua la aplicación a la recuperación	
TIPO DE SUFIJOS	No especifica	No especifica	Flexivos y derivativos	No especifica	No especifica	Flexivos, derivativos y endíticos	Flexivos y derivativos	Flexivos y derivativos	Flexivos
MODO DE ACTUAR	No especifica	Reglas de contexto sensible	1º elimina los sufijos flexivos 2º derivativos por las categorías gramaticales	Las reglas se aplican en orden alfabético.	No especifica	6 pasos	Primero los flexivos y después los derivativos	Aplica las reglas correspondientes al sufijo más largo	

Tabla 12 Comparación de los algoritmos distintos del inglés

8. La evaluación de los sistemas de lematización.

A la hora de evaluar la lematización, hay dos tipos de parámetros, los que afectan al propio algoritmo de lematización, como son la corrección de la lematización y la correcta ejecución de la compresión, y los que afectan a la aplicación de la lematización a la recuperación. Hay un parámetro más, que se inscribe en ambos grupos: el tiempo.

Considerar todos estos criterios no es fácil y, de hecho, pocos de los trabajos encontrados trabajan con estos cuatro parámetros.

8.1 Corrección de la lematización

Es decir, si la conflación se produce correctamente. Para calcular esto, Paice[256], propone como medidas la infralematización y la sobrelematización. Con ellas es posible establecer una comparación cuantitativa entre los distintos métodos de lematización. Para ello se establece una lista de grupos de palabras semánticamente relacionadas, el lematizador ideal será aquel que lematiza todas las palabras de un grupo con el mismo lema. Si un grupo contiene más de un lema se están produciendo errores de infralematización y la exhaustividad cae. Si el lema de cierto grupo aparece también como lema de otros grupos se está produciendo sobrelematización por lo que estamos perdiendo precisión. Un buen lematizador es el que tiene tasas bajas de error tanto por sobrelematización como por infralematización. Paice establece una serie de medidas para saber en qué medida se relacionan las palabras que pertenecen a un mismo grupo y con las de otros grupos mediante el *"deseo de unión total*", *"uniones totales no deseadas"* , *"numero total de errores no archivados"* ... A pesar de la complejidad de estas

[256] C.P. PAICE (1996) op. cit.

medidas resulta difícil establecer que tipo de errores son los menos perjudiciales ya que según Paice no se puede afirmar que siempre la infralematización sea mejor que la sobrelematización y viceversa.

Cualquier lematizador debe evitar el máximo de errores de cualquiera de estos dos tipos, pero esto no siempre es fácil. En función del tipo de lematizador se darán más errores de un tipo o de otro. De este modo, mientas que un lematizador ligero (como el *S Stemmer* de Harman) permite más errores de infralematización, un lematizador fuerte, al suprimir todos los finales largos, permitirá más errores de sobrelematización

8.2 Correcta ejecución de la compresión

Una de las ventajas de la lematización es que reduce el tamaño de los ficheros que contienen la información al reducir las palabras a una forma común. Habrá por tanto que calcular en qué porcentaje se reduce el tamaño de la información. Para calcular la tasa de compresión bastará con dividir el número de palabras inicial entre el número de lemas resultante. Si queremos calcular la tasa de compresión absoluta y el número de palabras únicas entre el número de lemas resultantes de la lematización si queremos que sea relativa.

8.3 Efectividad en la recuperación

Jacquemin[257] afirma que la lematización debe ser evaluada en relación con la recuperación de la información, ya que en una sobrelematización puede

[257] C. JACQUEMIN and E. TZOUKERMAN. NLP term variant extraction synergy between morphology, lexicon, and sitax. EN T. Stzalkowski (ed) *Natural Languaje Information Retrieval.* Klwer Academic Publisher, 1999 p. 25-74

corresponder a una correcta relación lingüística pero irrelevante en la recuperación de información. Salton aplica este parámetro y para calcularlo se basa en las medidas tradicionales de R.I.: precisión y exhaustividad[258]. En este sentido Hull[259] afirma que las medidas propuestas por Salton no son las más adecuadas por lo que también propone un análisis estadístico detallado.

8.4 Tiempo

El último parámetro, el tiempo, como decíamos al principio afecta tanto a la propia lematización como a la aplicación a la R.I. En este sentido, podemos medir el tiempo que se tarda en lematizar y compararlo con el que se tarda en preparar la base de datos para la recuperación con otro sistema, por ejemplo aplicando un lenguaje controlado como base para establecer la indización posterior. También podemos medir el tiempo que se tarda en hacer las tareas propias de la recuperación: indización, recuperación.

A la hora de utilizar este parámetro, hay que tener en cuenta que también dependerá de la complejidad del algoritmo (tamaño de las listas de sufijos y número de reglas) y del procesador que utilicemos, por lo que dado que la capacidad de éstos está continua evolución este parámetro no resulta de gran importancia.

[258] Ver la parte dedicada a las medidas de evaluación.

[259] D. A. HULL (1996) op. cit.

III EL LEMATIZADOR

1 Objetivos.

El principal objetivo de este trabajo es la creación de un lematizador que extraiga automáticamente los lemas de las palabras para posteriormente aplicarlo a un sistema de recuperación que trabaje con textos en lenguaje natural en español, es decir, sin que el lenguaje de dichos textos haya sido traducido a un lenguaje controlado.

Se intentará determinar, la supresión de qué sufijos es más eficaz de cara a la recuperación de información; si la de los sufijos flexivos, es decir, aquélla que sólo contempla plurales, género y desinencias verbales, que denominaremos **lematización flexiva**, o aquélla en la que también se incluyen, además de los flexivos, los sufijos derivativos **lematización derivativa**. Posteriormente se intentará mostrar si de este modo aumenta la eficacia de la recuperación respecto a los sistemas que no aplican la lematización.

También queremos probar, cómo inciden las palabras vacías en la recuperación de la información, qué listas son más adecuadas, si las que sólo incluyen categorías gramaticales vacías de contenido, o por el contrario, aquellas que además incluyen algunos verbos, adverbios, adjetivos y nombres, con una frecuencia alta de aparición. Para esto, estudiaremos las categorías de palabras vacías de contenido y usaremos un estudio de frecuencias, para conocer cuáles son las que aparecen más veces, y compararemos los resultados obtenidos con ambas listas[260].

[260] Aunque lo referente a las palabras vacías podríamos haberlo incluido en un capítulo independiente, ya que no está necesariamente relacionado con el lematizador, hemos preferido incluirlo en este capítulo junto con los experimentos.

Finalmente se establecerán comparaciones para ver con qué tipo de lematización, con qué lista de palabras vacías se obtienen mejores resultados en recuperación.

2 Antecedentes del trabajo.

El antecedente de este trabajo, como explicamos con anterioridad, está en el trabajo de Grado de Licenciatura[261], que a su vez se basó en el trabajo que Grossman[262] presentó en la TREC-5. A su vez este trabajo es una aplicación del algoritmo de Porter[263] para el español, por lo tanto, se puede considerar éste último, como el precedente de nuestro trabajo. Entre las conclusiones del citado trabajo de Grado, está que *el conocimiento que aplican los trabajos realizados en las TREC para la recuperación en español es escaso* y también que *el algoritmo de Porter para nuestro idioma no obtiene resultados buenos*. Por ello, este trabajo pretende mostrar si es posible mejorar los resultados de la aplicación del algoritmo de Porter al tener en cuenta un mayor conocimiento lingüístico con el fin de obtener mejores resultados en la recuperación de textos en español, y de este modo mejorar nuestros resultados de 1998.

3 La formación de palabras en español.

Antes de pasar al desarrollo del lematizador, consideramos necesario tener en cuenta una serie de aspectos básicos sobre la formación de palabras, ya

[261] R. GÓMEZ DÍAZ (1998) op. cit

[262] D. A. GROSSMAN (1996) op. cit.

[263] The Porter Stemming Algorithms op. cit.

que si conocemos cómo están formadas, será más fácil descomponerlas. Este conocimiento es necesario y previo a la elaboración del lematizador.

3.1 Mecanismos de formación de palabras en español

La formación de palabras se encarga del estudio de los medios de los que dispone una lengua para enriquecer su vocabulario. En español tenemos principalmente tres mecanismos:

La **composición** consiste en la combinación de lexemas independientes, para formar un nuevo término. Como es el caso de *limpiaparabrisas*, *sacacorchos*, *abrecartas* o *cabizbajo.*

La **derivación:** procedimiento de formación de una palabra nueva, mediante adición, supresión o intercambio de afijos[264], que puede ser mediante prefijos (**anti**demócrata, **pre**concebir), sufijos (corre**dor**, zapat**ero**), ya sea usando uno de los dos o los dos al mismo tiempo (parasíntesis[265]) (**em**bell**ecer**) o interfijos[266], en función de la posición respecto a la base sobre la cual actúen.

Un tipo especial de derivación es la regresiva o derivación cero, mediante ésta, las nuevas palabras resultantes poseen un cuerpo formal inferior al que poseían en su forma primitiva, cuando lo habitual es que se produzca lo contrario, así *pelea* procede de *pelear.*

---

[264] F. LÁZARO CARRETER. (1987) op. cit

[265] La parasíntesis es un tipo de derivación especial. En español es bastante productiva en la formación de verbos. Para que se dé, tiene que haber prefijación y sufijación simultáneamente, sin que exista la forma sin prefijar. Así de *blando* obtenemos el verbo *ablandar*, sin que exista *blandar o de *bello*, *embellecer* y no *bellecer. Cf. A. DARMOSTETES. *Traité de la formation des mots composès*. Paris: E Bouillon, 1974

[266] Los interfijos en español están asociados a la derivación apreciativa (Pabl-it-o, lej-it-os). Cf J. PENA La palabra: estructura y procesos morfológicos. *Verba* 1991 n 18 p. 69-128

La **importación de términos extranjeros**, éstos pueden adaptarse con terminaciones propias de nuestra lengua, o bien pueden integrarse en el idioma sin ningún tipo de adaptación. Así nos encontramos con palabras admitidas por la Real Academia de la lengua como *whisky* o *güisqui* o con la adaptación de términos como *zapeo*, del inglés "zaping".

Hay otros métodos menos productivos en el caso del español pero que también se deben de tener en cuenta, como son los acortamientos de palabras (*boli* en lugar de *bolígrafo*), la acronimia o la creación de nuevas palabras tomando como base siglas[267].

En este trabajo explicaremos brevemente los mecanismos de derivación centrándonos en la sufijación, por ser el método más productivo en español y el que se va a aplicar al sistema de recuperación de información.

3.2 Dificultades del estudio de la derivación en español

Antes de comenzar con el estudio de la derivación en español, hay que tener en cuenta las siguientes dificultades:

- El **establecimiento de la lista de sufijos**. No se trata simplemente de hacer una nómina de sufijos, sino que hay que tener en cuenta los alomorfos[268], que unos autores consideran como un solo sufijo (caso de *sor/tor*), mientras que otros autores los consideran sufijos distintos.

[267] Cf. M. CASADO VELARDE Otros procedimientos morfológicos: acortamientos, formación de siglas y acrónimos. EN BOSQUE, I. DEMONTE, V. (dir) (1999) op. cit.

[268] Alomorfos: "variantes de los sufijos con un mismo origen y forma parecida y con el mismo sentido que la forma principal" R. ALMELA PÉREZ, *Procedimientos de formación de palabras en Español*. Barcelona: Ariel, 1999 p. 104

Este es el motivo por el que los distintos especialistas en la materia no coinciden en sus listas de sufijos.

- La **segmentación**: resulta difícil establecer el corte entre la base y el sufijo cuando los derivados son verbales y se mantiene la vocal temática del verbo. (*hospedaje* → *hospedar*, ¿cuál sería el sufijo -aje o -je?.

- La **alternancia entre vocales de abertura media y mínima** (*cont**e**ner*→ *cont**i**nencia*), para explicar estos casos hay que hacer estudios diacrónicos.

- **Modificaciones consonánticas** que se producen bien por el contexto o porque la combinación resultante es inexistente en español. (*co**g**er*→ *co**j**o)*

- **Elisión de elementos** morfológicamente pertenecientes a la base (haplologia) *humilde*→ *humildad* y no *humilidad.

- **Irregularidades en la sufijación**: que puede ser debida a la diversidad de valores de un mismo sufijo[269]. De este modo -ero puede expresar persona (*zapatero*), lugar (*granero*), objeto (*cenicero*). También puede producirse el efecto contrario y varios sufijos expresan la misma idea o valor, así para expresar "color próximo a" tenemos los sufijos -ento, (*amarillento*), -ino (*blanquecino*), -uzco (*bancuzco*); -enco (*azulenco*); -ado (*azulado*)

- **Variantes dialectales**: -ico, -ino, -in, -illo... (*muchachico, muchachino, muchachin, muchachillo*)

[269] F. MONGE Aspectos de la sufijación en español. *Revista española de lingüística* 1996, 26 (1) p 43-56

Al comienzo de esta parte, al enunciar los mecanismos de la derivación en español, indicábamos la diferencia de los afijos en función de la posición respecto a la base, así, mientras los prefijos se anteponen a la base los sufijos se posponen a ella. Además de esta diferencia existen otra serie de divergencias entre ellos[270].

Mientras que los prefijos nunca modifican la clase gramatical de la base, los sufijos pueden o no hacerlo, incluso variar el género en el caso de la sufijación apreciativa. Así en caso de los prefijos, el nuevo término tendrá la misma categoría que el término del cual procede, de este modo el prefijo *des-* aplicado sobre el verbo *hacer*, nos da otro verbo: *deshacer*. En cambio, en el caso de los sufijos, tomando como base el verbo *correr,* obtenemos el nombre *corredor,* o del nombre femenino *casa*, al añadirle el sufijo -ón, obtenemos como resultado el *caserón*, que es un nombre (no ha habido cambio de categoría gramatical pero sí lo ha habido de género) masculino.

En ocasiones los prefijos pueden tener significado de por sí, es decir son piezas que pueden aparecer de manera independiente como entradas en el diccionario (hiper, ultra).

Otra diferencia es que los sufijos poseen mayor capacidad gramatical y menor capacidad léxica que los prefijos, es decir, que una las palabra con la misma raíz y distintos sufijos guardan relación semántica, así podemos expresar una acción (*correr*) y el agente para dicha acción, (*corredor*) cambiando el sufijo. Mientras que con los prefijos no variamos la categoría gramatical pero sí se modifica sustancialmente la significación del radical, "*huelga*", "*antihuelguista*".

---

[270] J. A. MIRANDA. *La formación de palabras en español*. Salamanca: Ed. del Colegio de España, 1994.

3.3 Clasificación de los sufijos

A la hora de clasificar los sufijos podemos atender a distintos criterios:

En función del uso: la división aquí sería entre sufijos productivos, que siguen generando palabras y los fosilizados, que han caído en desuso.

En función de la base sobre la cual actúan: así hablamos de sufijos denominales si toman como base un nombre, deadjetivales, si es un adjetivo, deverbales, en el caso de los verbos, y deadverbiales en el caso de los adverbios[271].

Flexivos y derivativos: los derivativos, aplicados a una base léxica crean una nueva palabra. La creatividad es inherente a estos sufijos. En español los derivativos son más que los flexivos. Estos sufijos pueden aumentar o caer en desuso. La flexión por el contrario no crea nuevas unidades léxicas. Comprende un conjunto cerrado y limitado, en él no influyen modas, como puede ocurrir con los derivativos[272].

Otra distinción, es la que divide los sufijos entre los que los que son apreciativos y los que no lo son. Los sufijos apreciativos son un tipo especial de sufijos, incluso existen dudas de si pertenecen a la derivación flexiva o a la derivativa, ya que tienen características de ambos grupos. Así, al igual que en los flexivos, no cambian la clase de palabra de la base: *casa* (n) → *casita* (n). El mismo tipo de sufijo puede actuar con distintas categorías gramaticales *casa* (n) → *casita* (n) *cerca* (adv) → *cerquita* (adv) *blanco* (adj) → *blanquito* (adj). Esto no ocurre con todos los sufijos derivativos aunque con algunos sí. A diferencia de

[271] I. BOSQUE La morfología EN ABAD, F. GARCÍA BERRIO, A. *Introducción a la Lingüística*. Madrid: Alambra, 1993

[272] S. VARELA ORTEGA. *Fundamentos de morfología*. Madrid: Síntesis, 1992

los sufijos flexivos, los apreciativos no van al final de la palabra, mientras que los flexivos siempre van en posición final. Los flexivos por definición no alteran la categoría gramatical de la base, aportan nociones de género, persona y número, mientras que los apreciativos por naturaleza están capacitados para modificar la categoría derivante. Los sufijos flexivos son instrumentales y los derivativos tiene valor sémico.

Finalmente la sufijación apreciativa tiene dos características propias y que la distingue tanto de la flexiva como de la derivativa .

Las palabras creadas por este proceso son con frecuencia lexicalizadas, es decir, adquieren nuevos significados alejándose del inicial. Una *casilla* no siempre es una casa pequeña.

Aportan connotaciones afectivas al lenguaje, provocando alteración semántica de la base de modo subjetivo, así los diminutivos expresan pequeñez o afectividad, los aumentativos amplían dimensiones, expresan grandiosidad o fealdad y los peyorativos desagrado o ridiculez.

3.4 Procesos de sufijación

Dentro de la sufijación se dan cuatro procesos en función de la categoría gramatical resultante

- **Nominalización**: tomando como base adjetivos (*bello* → *belleza*), verbos (pelear→pelea) y otros nombres (*ceniza* → *cenicero*) se forman nombres, este proceso es el más productivo en español.

- **Verbalización**: se forman verbos a partir de adjetivos (*redondo*→ *redondear*), de nombres (*batalla* → *batallar*); o incluso de adverbios (*cerca* → *acercar*) y de pronombres (*tú* → *tutear*). También puede tomar como base otro verbo, en este caso es muy frecuente que se

produzca prefijación y sufijación al mismo tiempo (parasíntesis[273]) (*dormir → adormilar*).

- **Adjetivación**: los adjetivos se forman a partir de nombres (*gusto→gustoso*), verbos (*transportar→tansportable*).

- **Adverbialización**: es el menos productivo en español y se reduce a la formación de adverbios a través de otros adverbios (*cerca →cerquita*) (con sufijos apreciativos principalmente) o tomando como base adjetivos más el sufijo –mente (*bueno → buenamente*).

3.5 Reglas de sufijación

Cuando hablamos de reglas en sufijación estamos hablando de predicción de comportamiento, pero debemos tener en cuenta que estas previsiones se incumplen con mucha frecuencia. Así Lang[274] afirma que *"debido a las carencias que presentan los paradigmas derivativos, los teóricos han tratado con prudencia el tema relativo a las reglas de formación de palabras, llegando incluso a proponer su sustitución por el concepto de 'restricciones derivativas' que impiden la espontánea aplicación de reglas morfológicas básicas"*. La prueba más palpable de esto es la gran variedad de alomórficos que hay. No podemos explicar por qué el sufijo -ción actúa con determinados verbos y sin embargo no lo hace con otros. *Obtener > obtención*, en cambio *Mover* > *movición; y lo mismo ocurre con el sufijo miento *obtener* > *obtenimiento, *mover>movimiento,* sin recurrir a argumentos etimológicos, fonéticos, lexémicos... y en ocasiones ni siquiera éstos sirven para explicarlo.

---

[273] Cf. F. A. LÁZARO MORA Sobre la parasíntesis en español. *Dicenda. Cuadernos de filología hispánica*, 5 Ed. Madrid: Universidad Complutense de Madrid, 1986 p. 221-235.

[274] M. F. LANG *Formación de palabras en español: Morfología derivativa productiva en el léxico moderno* Madrid: Cátedra, 1992. p. 51

Hay que tener en cuenta, que las reglas dependen de cada lengua, de este modo en español destacan las siguientes[275].

- Gran parte de los sufijos pueden aplicarse a bases de distintas categorías gramaticales[276] (*casa* (nombre) → *casita*; *cerca* (adverbio) → *cerquita*).

- Cuando un prefijo y un sufijo actúan sobre la misma base, esto puede darse necesariamente al mismo tiempo (parasíntesis) (*embellecer*) o en momento distintos (*precalentar, calentamiento; precalentamiento*).

- Los pronombres y adverbios también pueden funcionar como bases (*le*→ *leísmo*; *cerca* → *acercar*), aunque esto no es muy frecuente.

- Los sufijos pueden actuar sobre bases que su vez contienen sufijos (*nación*→ *nacional* → *nacionalizar*→ *nacionalización*).

- Una reglas son más productivas que otras, así las nominalizaciones y las verbalizaciones son más productivas frente a las adverbializaciones, que son muy escasas.

- Un mismo fenómeno puede interpretarse de varias formas.

- Puede haber caso únicos (hápax) (*ejecutar*→ *ejecutivo* en lugar de *ejecutativo).

- El bloqueo condiciona las reglas: el bloqueo sucede cuando un sufijo no se pega a una determinada base porque el significado del sufijo está contenido en la base (*tener*→*tención) en cambio existen formas como *retención*.

---

[275] R. ALMELA PÉREZ (1999) op. cit.

[276] F. MONGE. (1996) op. cit.

Las reglas siguen una escala de regularidad que va de lo más regular a lo irregular debido a un conjunto de factores de raíz fonética, lexémica, morfémica... por eso tenemos que tener en cuenta la estructura y etimología de la base, el acento, la configuración fonética de la base.

Hay un grupo de terminaciones que están a caballo entre la sufijación y la composición, son los llamados sufijoides de origen culto entre los que se encuentran -*logo, -logía, -cracia*[277]

4 Consideraciones previas a la creación del lematizador.

Antes de pasar al desarrollo del lematizador fue necesario ir tomando una serie de decisiones sobre los acentos, los prefijos, la estructura de las palabras, la elección de los sufijos y los criterios para la selección de los lemas.

4.1 Los acentos

Antes de la elaboración del lematizador se estudió la posibilidad de incluir o suprimir las tildes. En español, a pesar de que las reglas gramaticales son muy claras al respecto[278], sobre todo en documentos electrónicos muchas veces se encuentran errores, al omitir erróneamente las tildes. Con éstas, hubiéramos podido desambiguar casos en los que la misma palabra es vacía o no, como por ejemplo *de* (preposición) y *dé* (verbo dar). Los ejemplos de este tipo son escasos.

[277] R. ALMELA PÉREZ (1999) op. cit.

[278] Cf. E. ALARCOS LLORACH. *Gramática de la lengua española*. 7ª reimp. Madrid: Espasa Calpe, 1995 p. 44-48

El otro supuesto que nos permitiría desambiguar son los tiempos verbales: *amara* (pretérito imperfecto del subjuntivo) y *amará* (futuro simple de indicativo). En el caso de los verbos para el lematizador y dado que ambos casos serán reducidos a la raíz, el suprimir las tildes simplifica las reglas, la información gramatical que indica la tilde no es necesaria y la información semántica no varía.

Por estos motivos, las reglas no contemplan las tildes, y antes de proceder a la lematización, se normalizan las palabras, suprimiendo todos los signos de acentuación y puntuación, reduciendo de este modo el tamaño de los ficheros de índice.

4.2 Los prefijos

La mayor parte de los algoritmos de lematización, estudiados con anterioridad al desarrollo de nuestro lematizador, no tratan los prefijos, ya que éstos, como vimos anteriormente, suelen cambiar radicalmente el significado del término. Este es el motivo por el que Paice[279] sugiere que no se consideren en el tratamiento de textos, aunque no siempre se produce un cambio radical del significado, lo que lleva a Savoy[280] a tenerlos en cuenta, sobre todo para materias técnicas como pueden ser la medicina o la química.

La información semántica que tienen los prefijos depende de cada idioma, así en español, podemos distinguir entre dos tipos de prefijos, los que aportan información semántica y los que no la aportan, a este grupo pertenecen las formas parasintéticas, es decir aquellas formas que se forman con un prefijo y un sufijo al mismo tiempo, sin que exista la forma sufijada sin la prefijada. De ese modo con el adjetivo *bello*, el sufijo *–ecer* y el prefijo *em-* formamos el verbo

[279] C. D. PAICE *Information Retrieval and the Computer*. London: McDonal and Janes, 1977

[280] J. SAVOY (1993) op. cit.

embellecer, pero no existe la forma **bellecer*. Estos casos, son típicos en los verbos denominales y deadjetivales, es decir los verbos procedentes de nombres y adjetivos respectivamente.

Por estos motivos decidimos omitir los prefijos.

4.3 La estructura de las palabras

La estructura de las palabras puede estudiarse desde distintos enfoques en función de las partes que se considere que tiene una palabra. Para el lematizador consideraremos que las palabras tienen una estructura bipartita[281], es decir, se componen de una base y uno o varios sufijos, ya sea éstos flexivos o derivativos, pegados a esa base. Así una base puede llevar al mismo tiempo un sufijo derivativo y otro flexivo al mismo tiempo (muy frecuente en el uso de los plurales), o varios derivativos al mismo tiempo (*nacion-al-iza-ción*). Por eso, el programa será recursivo, es decir, irá quitando los sufijos uno a uno. De este modo, si el programa tiene que lematizar la palabra *zapateros* primero reducirá la palabra hasta el singular (*zapatero*) y después a la palabra de la que deriva (*zapato*). También podría haberse hecho de manera que quitara terminaciones (ya fuera un solo sufijo o varios agrupados al mismo tiempo). Así para quitar tratar la palabra *zapateros* consideraría la terminación *-eros*, pero esta opción hubiera multiplicado el número de terminaciones posibles ya que habría que haber tenido en cuenta todas las combinaciones posibles de sufijos.

El programa actuará teniendo en cuenta las palabras, es decir, que en aquellos casos en los que la palabra y la base no coincidan, lo primero que hará el programa será reducir a la base, quitando tantos sufijos como sea necesario y después aplicar las reglas de pegado. De este modo en el término relojero, la base

[281] R. ALMELA PÉREZ (1999) op. cit.

está incluida en la palabra, por lo tanto solo necesitaremos una regla que elimine el sufijo:

RELOJERO - ERO = RELOJ

Si esta misma palabra estuviera en plural:

RELOJEROS - S = RELOJERO - ERO = RELOJ

En cambio con la palabra *animador* tendremos que hacer una operación más, para dejar la base, ya que en este caso palabra y base no coinciden.

ANIMADOR - DOR + R= ANIMAR

Con el plural ocurre lo mismo:

ANIMADORES - ES =ANIMADOR - DOR + R= ANIMAR

Los casos de parasíntesis serán considerados como excepciones por lo que se incluirán en el diccionario del programa, de este modo las palabras derivadas de formas parasintéticas obtendrán como salida la palabra sobre la que se formó el verbo. Así la base del termino *embellecer*, y de *embellecimiento,* es *bello*. Sólo se han considerado así las formas de las que de un adjetivo sale un verbo y no existe gran cambio semántico.

4.4 La elección de los sufijos

Como acabamos de explicar, decidimos que el programa se basará en los sufijos en lugar de en las terminaciones. La única excepción que hemos hecho es en el caso de los verbos donde hemos agrupado la desinencia de persona y la de tiempo con el fin de que el programa fuera más rápido.

Al elegir los sufijos, nos encontramos con dos posibilidades, la primera era hacer las reglas para cada sufijo con sus variantes alomórficas, bien de manera conjunta o por separado, o bien hacer una selección de los sufijos. Elegir la primera opción nos hubiera llevado a tener en cuenta muchos sufijos, algunos de ellos prácticamente en desuso. Además hay que tener en consideración que si los sufijos son poco productivos, lo que ocurre es que se crean reglas para pocos casos, cuando resulta más fácil y también más práctico incluir esos casos como excepciones. De la lista inicial de sufijos de la parte de "*Desarrollos gramaticales*" del diccionario de María Moliner en CD-Rom[282], fuimos eligiendo aquellos sufijos con sus variantes alomórficas más productivas. Se han contemplado finalmente 230 sufijos en total. Dependiendo de los casos los sufijos y sus variantes se han agrupado bajo las mismas reglas o bajo reglas distintas, siempre buscando la mayor simplicidad. De este modo mientras que *tor/sor* se han considerado sufijos distintos, *-ptible* y *-stible* están agrupados bajo el sufijo *–ble.*

A continuación indicamos la lista de terminaciones contempladas, hay que tener en cuenta que según lo indicado anteriormente, cuando bajo el mismo sufijo se encuentran agrupadas variantes solo aparece la entrada del que los agrupa.

[282] M. MOLINER Diccionario de uso del español ed. en CD -Rom. Madrid Gredos, 1996

4.4.1 Lista de todos los sufijos

Eesta lista será utilizada, según veremos más adelante por la lematización derivativa.

A	ASTRO	CULO
ACEO	ATA	D
ACHO	ATICO	DA
ACO	ATO	DAD
AGA	AZ	DERAS
AINA	AZGO	DERO
AIS	AZO	DIO
AJE	AZON	DIZO
AJO	BA	DO
AL	BAIS	DOR
ALES	BAMOS	DURA
AMOS	BAN	DURIA
AN	BAS	E
ANCO	BILIDAD	EAR
ANEO	BLE	ECER
ANO	BUNDO	ECILLO
ANZA	CECITO	ECITO
AR	CIDA	EDA
ARIO	CIDIO	EDAL
AS	CILLO	EIS
ASTA	CION	EJO

ELA	ESEN	IERA
EMOS	ESES	IERAIS
EN	ETA	IERAMOS
ENCO	ETE	IERAN
ENDO	EZ	IERAS
ENGO	EZA	IERE
ENO	EZNO	IEREIS
ENSE	EZO	IEREMOS
ENTO	FILIA	IEREN
EÑO	FILO	IERES
ERA	FONO	IERON
ERE	GRAFIA	IESE
EREIS	I	IESEIS
EREMOS	IA	IESEMOS
EREN	IAIS	IESEN
ERES	IAMOS	IESES
ERIA	IAN	IFICAR
ERIO	IANO	IJO
ERIZA	IAS	IL
ERO	ICA	ILLO
ES	ICIA	IMOS
ESCA	ICIE	IN
ESE	ICIO	INO
ESEIS	ICO	IO
ESEMOS	IEGO	IS

ISIMO	MOS	RAZO
ISMO	N	RE
ISO	NCIA	REIS
ISTA	NDERO	REMOS
ISTE	NDO	RIA
ISTEIS	NOS	RIAIS
ISTICO	NTE	RIAMOS
ITARIO	O	RIAN
ITIS	OIDE	RIAS
ITO	ON	RIZAR
ITUD	ONA	RON
IVO	ORRO	S
IZAL	OS	SCO
IZAR	OSIS	SE
IZO	OSO	SEIS
LA	OTA	SEMOS
LO	OTE	SEN
LOGIA	OY	SES
LOGO	PATIA	SIDOR
LOS	QUECILLO	SION
MANIA	RA	SITIVO
ME	RAIS	SIVO
MENTE	RAMOS	SOR
MENTO	RAN	STE
MIENTO	RAS	STEIS

TAD	TORIO	UITO
TARIO	TUD	UNO
TE	TURA	URA
TIVO	UCHO	ZCO
TO	UCO	
TOR	UDO	

4.4.2 Lista de los sufijos flexivos

Después hicimos una selección de los sufijos anteriores para elegir sólo las desinencias verbales, el plural y la normalización de género[283]. Para una mayor claridad hemos preferido dar las dos listas, aunque todos los que están en esta lista también están en la anterior.

A	BAS	ENDO
AIS	D	ERE
AMOS	DA	EREIS
AN	DAD	EREMOS
AS	DO	EREN
BA	E	ERES
BAIS	EIS	ERIA
BAMOS	EMOS	ES
BAN	EN	ESE

[283] Estos sufijos serán los empleados para la lematización flexiva.

ESEIS
ESEMOS
ESEN
ESES
I
IA
IAIS
IAMOS
IAN
IAS
IERA
IERAIS
IERAMOS
IERAN
IERAS
IERE
IEREIS
IEREMOS
IEREN
IERES
IERON
IESE
IESEIS
IESEMOS
IESEN
IESES
IMOS
IO
ISO
ISTEIS
LA
LAS
LO
LOS
ME
N
NDO
NOS
O
OS
RA
RAIS
RAMOS
RAN
RAS
RE
REIS
REMOS
RIA
RIAIS
RIAMOS
RIAN
RIAS
RON
S
SE
SEIS
SEMOS
SEN
SES
STE
STEIS
TE

4.5 Criterios de selección de los lemas

Con el fin de obtener la mayor uniformidad posible, establecimos los criterios para elegir los lemas tanto para el procedimiento manual como para la creación de las reglas. De este modo, en los casos de los adjetivos, sustantivos y adverbios deverbales, se ha elegido el verbo en infinitivo como lema, con el fin de simplificar reglas. Cuando los verbos son claramente denominales (sufijo en *–ecer* por ejemplo), el lema, es el sustantivo correspondiente. Cuando el lema es un adjetivo o un sustantivo, se elige su forma masculino singular.

Sobre este punto, es necesario advertir que no se ha hecho un estudio filológico de la historia de cada término para ver cuál es el que históricamente apareció en primer lugar, ya que este trabajo está fuera de nuestro alcance y complicaría de manera innecesaria el trabajo que aquí se pretendía.

5. Las palabras vacías.

5.1 Introducción

Desde los comienzos de los trabajos de la R.I., Luhn[284] afirmó que las palabras con frecuencias de aparición alta no eran buenas como términos de índice, por lo que pronto se empezaron a discriminar evitando así ruido innecesario, disminuyendo el tamaño de los ficheros invertidos, lo que hacía que aumentara el espacio libre[285] y la velocidad de procesado, sin dañar la efectividad

[284] H. P. LUHN, (1957) op. cit.

[285] Por ejemplo en español se calcula que aproximadamente entre el 20 y el 30% de las palabras que componen un documento son las diez palabras más frecuentes. Almacenamiento y recuperación de la información textual. [en línea]. http://protos.dis.ulpgc.es/docencia/seminarios/rit/analisis_lexico/sld034.htm> [consultado el 10-03-01]

en la recuperación, ya que dichas palabras no aportan información para la R.I. al no ayudar en la distinción entre documentos relevantes y no relevantes. A este tipo de palabras se las denominó palabras vacías o diccionario negativo.

Partiendo de la idea de Luhn, en el Trabajo de Grado llegamos a conclusión de que "la supresión de las palabras vacías aumenta la precisión y la exhaustividad de las búsquedas frente a los sistemas que no las suprimen".[286] Por eso, en este trabajo, decidimos ver sí ampliando la lista y variando el momento de supresión de las palabras vacías podíamos mejorar los resultados.

5.2 Criterios de creación de listas de palabras vacías

Hay dos criterios básicos a la hora de establecer una lista de este tipo. El primero de ellos es la frecuencia alta de aparición, tal y como demostró Luhn en sus trabajos, pero hay que tener en cuenta que algunas de las más frecuentes son demasiado importantes para ser excluidas como términos de índice[287]. El otro criterio está basado en las categorías gramaticales, ya que determinadas categorías no solo no aportan información sobre el contenido del documento, aunque sean necesarias para facilitar la legibilidad del mismo, como son las conjunciones, preposiciones, determinantes, algunos adverbios... sino que además introducen ruido documental en la recuperación. Estos y otros criterios, fueron analizados en 1999 por Jacques Savoy [288]. En dicho trabajo se establecieron los siguientes criterios para crear una lista de palabras vacías:

- Palabras con alta frecuencia de aparición.
- Numerales, tanto ordinales como cardinales.

[286] R. GÓMEZ 1998. op. cit p. 146

[287] Almacenamiento y recuperación de la información textual. [en línea] op. cit.

[288] J. SAVOY. (1999) op. cit.

- Nombres y adjetivos muy usados (dependiendo de la materia que se trate). Así en una base de datos que trate sobre medicina, probablemente el término paciente, sea vacío de contenido.

- Posesivos, pronombres y conjunciones.

- Algunas formas verbales que no aportan información semántica (ser/estar/haber)

En este trabajo, como ya indicamos antes, hemos aplicado varias listas. La primera contiene las palabras en función de su categorías gramatical. Por lo tanto esta lista está formada por preposiciones, conjunciones, artículos, posesivos... también hemos añadido algunos numerales y el infinitivo de ser, estar y haber. A esta lista la hemos denominado *vacías leve.* En total la lista se compone de 264 palabras.

La segunda lista, además de los términos anteriores incluye también palabras con una alta aparición en textos, con un total de 682 palabras. La hemos denominado *vacías fuerte.* Para elaborarla, nos basamos en el estudio de frecuencias hecho por José Ramón Alameda y Fernando Cuetos[289]. Este diccionario ofrece el estudio de la frecuencia de aparición de 2.000.000 de palabras, recogidas de 606 textos escritos pertenecientes a distintos géneros literarios y publicados desde 1978 a 1993. La muestra tiene la siguiente composición: 50% de novela, 25% de prensa, 15% de ensayo y 10% de divulgación científico-técnica. Del diccionario se tomaron las 220 palabras con un índice más alto de aparición, muchas de ellas son también vacías en función de contenido. Esta lista se ha ido completando con palabras, que aunque no tienen una frecuencia de aparición tan elevada, no aportan información para determinar

[289] J. R. ALAMEDA, F. CUETOS. *Diccionario de frecuencias de la Unidades lingüísticas del castellano*. Vol I-II. Oviedo: Universidad, 1995.

la relevancia de un documento respecto a una pregunta, como son los números. Se incluyeron también interjecciones, a pesar de que no es habitual encontrarlas en textos escritos, y palabras latinas de uso habitual en expresiones hechas del tipo *"modus operandi"*. También están los verbos *haber*, *ser*, *estar* en todas sus formas simples conjugadas, ya que si las palabras vacías se suprimen antes de la lematización es necesario considerar todas las formas variantes y el hacerlo así implica considerar todas las formas variantes. Se eligió hacerlo de esto modo porque aunque haya que tener en cuenta una lista de palabras vacías mayor, al eliminar las palabras antes de la lematización se reduce considerablemente el número de palabras a lematizar con el considerable ahorro de tiempo[290]. Si las palabras vacías se suprimen después de la lematización basta con incluir una forma en lugar de todas las variantes de dicha forma.

A continuación indicamos las dos listas[291].

5.2.1 Lista de vacías fuerte

A	ADEMAS
ABA	ADONDE
ABUR	AFIRMO
ACA	AGORA
ACHIS	AGUR
ACTUALMENTE	AHE
AD	AHI
ADELANTE	AHO

[290] Ver la fase 1 de los experimentos.

[291] Como se podrá observar hemos eliminado los acentos.

AHORA
AJA
AJAJA
AJAJAY
AJO
AL
ALA
ALE
ALELUYA
ALGO
ALGUN
ALGUNA
ALGUNAS
ALGUNO
ALGUNOS
ALLA
ALLI
AMBAS
AND
ANGELA
ANIMO
ANTE
ANTES
APARTE
APENAS
APROXIMADAMENTE
AQUEL
AQUELLA
AQUELLAS
AQUELLO
AQUELLOS
AQUESTA
AQUI
ARRE
ARTICULO
ASI
AUN
AUNQUE
AUPA
AX
AYAYAY
AYME
BAH
BAJO
BASTANTE
BELLI
BIEN
BIS
BONA
BRIOS

BUENA
BUENAS
BUENO
BUENOS
CA
CABE
CARACHO
CARAMBA
CARAPE
CARAY
CASI
CASPITA
CASUS
CATORCE
CAUTELAM
CE
CERCA
CHAO
CHAPO
CHAU
CHITO
CHITON
CHO
CIEN
CIENTO
CIENTOS
CIERTO
CINCO
CINCUENTA
COMO
CON
CONMIGO
CONQUE
CONSIGO
CONTIGO
CONTRA
CONTRARIIS
CORCHOLIS
COSAS
CRIBAS
CUAL
CUALES
CUALESQUIER
CUALESQUIERA
CUALQUIER
CUALQUIERA
CUAN
CUANDO
CUANTA
CUANTAS

CUANTO
CUANTOS
CUARENTA
CUARTO
CUATRO
CUERPO
CURIA
CUYA
CUYAS
CUYO
CUYOS
DADO
DE
DEBAJO
DECIMO
DEL
DELANTE
DELLA
DELLO
DEMASIADO
DENTRAMBAS
DENTRAMBOS
DENTRO
DEO
DES
DESA
DESDE
DESDEL
DESE
DESO
DESPUES
DESQUE
DESTA
DESTE
DESTO
DESTOTRA
DESTOTRO
DETRAS
DIA
DIAS
DICE
DIE
DIECINUEVE
DIECIOCHO
DIECISEIS
DIECISIETE
DIEM
DIEZ
DIJO
DIVINIS

DIVINUM
DO
DOCE
DOMO
DONDE
DOQUIER
DOS
DURANTE
E
EA
ECCLESIAE
EFESIOS
EFIGIES
EJEM
EL
ELE
ELLA
ELLAS
ELLO
ELLOS
EMBARGO
EMPERO
EN
ENCIMA
ENSEGUIDA
ENTONCES
ENTRAMBOS
ENTRE
ENTRETANTO
EPA
ERAIS
ERAMOS
ERAN
ERAS
ERES
ERGO
ES
ESA
ESAS
ESE
ESO
ESOS
ESTA
ESTABA
ESTABAIS
ESTABAS
ESTADO
ESTAIS
ESTAMOS
ESTAN

ESTAS
ESTE
ESTO
ESTOS
ESTOY
ET
ETC
EVOHE
EX
FACIE
FERENDAE
FORMA
FRENTE
FU
FUE
FUERA
FUERAIS
FUERAMOS
FUERAN
FUERAS
FUERON
FUESE
FUESEIS
FUESEMOS
FUESEN
FUESES
FUI
FUIMOS
FUISTE
FUISTEIS
GENERIS
GRAN
GRATIA
GUA
GUALA
GUARTE
GUAY
GUAYAS
HA
HABEIS
HABIA
HABIAIS
HABIAMOS
HABIAN
HABIAS
HABIDO
HABIENDO
HABRAN
HABRAS
HABRE

HABREIS
HABREMOS
HABRIA
HABRIAIS
HABRIAMOS
HABRIAN
HABRIAS
HACE
HACIA
HAE
HALE
HAN
HAS
HASTA
HAY
HAYA
HAYAN
HAYAS
HAYEIS
HAYEMOS
HE
HECHO
HEMOS
HI
HOC
HOMBRE
HOMBRES
HOMINEM
HOY
HU
HUBE
HUBIERA
HUBIERAMOS
HUBIERAN
HUBIERAS
HUBIERE
HUBIEREIS
HUBIEREMOS
HUBIEREN
HUBIERES
HUBIERON
HUBIESE
HUBIESEIS
HUBIESEN
HUBIESES
HUBIMOS
HUBISTE
HUBISTEIS
HUF
HUIFA

HUM
HURRA
HUY
IBA
IDEM
IJUJU
IN
INCLUSO
JA
JAJAY
JAMAS
JAU
JE
JI
JO
JOLIN
JU
JUNTO
JUSTO
LA
LADO
LAS
LATAE
LATERE
LE
LEJOS
LES
LITEM
LO
LOS
LUCRANDO
LUEGO
LUGAR
LUZ
MADIOS
MADRE
MAES
MAGAR
MAGNUM
MAGUER
MAGUERA
MALO
MAMOLA
MAÑANA
MANERA
MANO
MANOS
MARE
MAS
ME

MECACHIS
MENOS
MERA
MERAMENTE
MERIDIEM
MI
MIA
MIALMAS
MIAS
MIENTRAS
MIL
MILES
MILLON
MILLONES
MINUTA
MIO
MIQUIS
MIS
MISMA
MISMAS
MISMO
MISMOS
MOMENTO
MORTIS
MOSTE
MUCHA
MUCHAS
MUCHO
MUCHOS
MUERTE
MUNDO
MUY
NADIE
NE
NEMINE
NEN
NI
NIMIS
NIN
NINGUNA
NINGUNAS
NINGUNO
NINGUNOS
NO
NOCHE
NON
NOS
NOSOTRAS
NOSOTROS
NOT

NOVENTA
NUESA
NUESTRA
NUESTRAS
NUESTRO
NUESTROS
NUEVA
NUEVAS
NUEVE
NUNCA
NUTUM
O
OCHENTA
OCHO
OCTAVO
OH
OJOS
OLE
ONCE
ONDE
OR
ORA
ORBI
OS
OTRA
OTRAS
OTRO
OTRORA
OTROS
OTROSI
PACE
PADRE
PANE
PARA
PARDIEZ
PARDIOBRE
PARECE
PARI
PARTE
PARTIBUS
PATAPLUM
PCHE
PCHS
PECCATA
PECTORE
PEOR
PER
PERO
PERPETUAM
PESIA

PETTO
POCAS
POCO
POCOS
PODIA
POR
PORQUE
PORVIDA
PRIMERA
PRIMERO
PRINCIPIO
PRIORI
PRO
PROMPTU
PRONTO
PU
PUCHA
PUEDE
PUERTA
PUES
PUESTO
PUF
QUA
QUE
QUEM
QUID
QUIEN
QUIENES
QUIER
QUINCE
QUINTO
QUIZA
QUIZAS
QUO
RECORCHOLIS
REDIEZ
REDIOS
REFERO
RELATA
REPENTE
REQUIESCAT
RETRO
SALVE
SANCTA
SANSEACABO
SE
SEA
SEAIS
SEAMOS
SEAN

SEAS
SED
SEGUN
SEGUNDO
SEIS
SENDOS
SENTENTIAE
SEPTIMO
SERA
SERAN
SERAS
SERE
SEREIS
SEREMOS
SERIA
SERIAIS
SERIAMOS
SERIAN
SERIAS
SESENTA
SETENTA
SI
SIDO
SIEMPRE
SIENDO
SIETE
SIMILI
SIN
SINE
SINFIN
SINO
SIQUIER
SO
SOBRE
SOBRETODO
SOIS
SOLO
SOMOS
SOY
STATU
SU
SUA
SUI
SUS
SUSODICHA
SUSODICHAS
SUSODICHOS
SUSODUCHO
SUYA
SUYAS

SUYO
SUYOS
TAL
TALES
TAMBIEN
TAMPOCO
TAN
TANTA
TANTAS
TANTO
TANTOS
TARARI
TARARIRA
TARDE
TE
TEJEMANEJE
TERCERO
TI
TIEMPO
TIENE
TIENEN
TIQUIS
TIRTE
TODA
TODAS
TODAVIA
TODO
TODOS
TON
TRAS
TRAVES
TRECE
TREINTA
TRES
TRIGESIMA
TRIGESIMO
TRIGESIMOCUARTA
TRIGESIMONOVENO
TRIGESIMOPRIMERO
TRIGESIMOSEGUNDO
TRIGESIMOSEPTIMA
TRIGESIMOSEXTO
TU
TUS
TUYA
TUYAS
TUYO
TUYOS
UF
UH

ULTRA
UN
UNA
UNAS
UNDECIMA
UNDECIMO
UNO
UNOS
URBI
USTED
USTEDES
UT
VA
VADE
VECES
VEINTE
VEINTIOCHOIMEDIO
VEINTISEIS
VEINTISIETE
VEINTITANTOS
VEINTIUN
VEINTIUNA
VELAHI
VELAY
VENTICUATRO
VENTIDOS
VER
VERA
VERBI
VERDAD
VEZ
VICEVERSA
VICEVERSAS
VICTOR
VIDA
VIGESIMA
VIGESIMO
VIGESIMOCTAVA
VIGESIMOCUARTA
VIGESIMOCUARTO
VIGESIMONOVENA
VIGESIMOOCTAVO
VIGESIMOPRIMERA
VIGESIMOPRIMERO
VIGESIMOQUINTA
VIGESIMOSEGUNDA
VIGESIMOSEPTIMA
VIGESIMOSEXTA
VIGESIMOTERCERA
VISU

VITOR
VOLENTE
VOS
VOSOTRAS
VOSOTROS
VUESTRA
VUESTRAS
VUESTRO
VUESTROS
Y
YA
YO
YUY

5.2.2 Lista de vacías leve

A
ACA
ADEMAS
AHI
AHORA
AL
ALGO
ALGUN
ALGUNA
ALGUNAS
ALGUNO
ALLA
ALLI
ANTE
ANTES
APARTE
APENAS
AQUEL
AQUELLA
AQUELLAS
AQUELLO
AQUELLOS
AQUI
ASI
AUN
AUNQUE
BAJO
BASTANTE

BIEN
BUENO
CABE
CASI
CATORCE
CERCA
CIEN
CIENTO
CIENTOS
CIERTO
CINCO
CINCUENTA
COMO
CON
CONMIGO
CONSIGO
CONTIGO
CONTRA
CUAL
CUALES
CUALESQUIER
CUALESQUIERA
CUALQUIER
CUALQUIERA
CUANDO
CUANTA
CUANTAS
CUANTO
CUANTOS
CUARENTA
CUARTO
CUATRO
CUYA
CUYAS
CUYO
CUYOS
DE
DEBAJO
DECIMO
DEL
DELANTE
DEMASIADO
DENTRO
DESDE
DESPUES
DETRAS
DIECINUEVE
DIECIOCHO
DIECISEIS
DIECISIETE

DIEZ
DOCE
DONDE
DOS
DURANTE
E
EL
ELLA
ELLAS
ELLO
ELLOS
EMBARGO
EMPERO
EN
ENCIMA
ENSEGUIDA
ENTONCES
ENTRE
ESA
ESAS
ESE
ESO
ESOS
ESTA
ESTAS
ESTE
ESTO
ESTOS
FRENTE
FUERA
HACIA
HASTA
HOY
INCLUSO
JAMAS
JUNTO
JUSTO
LA
LAS
LE
LEJOS
LO
LOS
LUEGO
MALO
MAÑANA
MANERA
MAS
ME
MENOS

MI
MIA
MIENTRAS
MIL
MILES
MILLON
MILLONES
MIO
MIS
MISMA
MISMAS
MISMO
MISMOS
MOMENTO
MUCHA
MUCHAS
MUCHO
MUCHOS
MUY
NADIE
NI
NINGUNA
NINGUNAS
NINGUNO
NINGUNOS
NO
NOS
NOSOTRAS
NOSOTROS
NOVENTA
NUESTRA
NUESTRAS
NUESTRO
NUESTROS
NUEVE
NUNCA
O
OCHENTA
OCHO
OCTAVO
ONCE
ORA
OS
OTRA
OTRAS
OTRO
OTROS
PARA
PARTE
PEOR

PERO
POCAS
POCO
POCOS
POR
PORQUE
PRIMERO
PRINCIPIO
PRONTO
PUES
PUESTO
QUE
QUIEN
QUIENES
QUINCE
QUINTO
QUIZA
QUIZAS
REPENTE
SE
SEA
SEGUN
SEGUNDO
SEIS
SEPTIMO
SESENTA
SETENTA
SI
SIEMPRE
SIETE
SIN
SINO
SO
SOBRE
SOLO
SU
SUS
SUYA
SUYAS
SUYO
SUYOS
TAL
TALES
TAMBIEN
TAMPOCO
TAN
TANTA
TANTAS
TANTO
TANTOS

TARDE
TE
TERCERO
TI
TODA
TODAS
TODAVIA
TODO
TODOS
TRAS
TRAVES
TRECE
TREINTA
TRES
TU
TUS
TUYA
TUYAS
TUYO
TUYOS
UNA
UNAS
UNO
UNOS
USTED
USTEDES
VEINTE
VEZ
VOSOTRAS
VOSOTROS
VUESTRAS
VUESTRO
VUESTROS
Y
YA
YO

6. Los autómatas de estados finitos.

Para el desarrollo del lematizador, creamos una máquina de estados finitos no determinista. Antes de mostrar cómo funciona, trataremos de explicar brevemente qué es un autómata y una máquina de estados finitos, qué tipos hay de ellos, para qué sirven y cómo se han aplicado al P.L.N.

La teoría de los autómatas está basada en el trabajo de Chomsky "*Teoría de las gramáticas transformacionales*"[292], que estableció las bases de la lingüística matemática.

6.1 Definición de autómata

Un autómata es una máquina secuencial capaz de recibir información y transformándola, generar nueva información[293], esta información puede ser simplemente que la cadena es o no correcta.

Si partimos de que el lenguaje se forma con un conjunto finito de caracteres, entonces podemos representar las palabras bien letra a letra o con unidades mayores, por ejemplo sílaba a sílaba o morfema a morfema, y si podemos hacerlo así podremos tomar como unidad las palabras y representar las frases[294].

---

[292] Cf. N. CHOMSKY. *La nueva sintaxis: teoría de la rección y el ligamento*. Barcelona: Paidós, 1988.

[293] M. ALFONSECA, J. SANCHO y M. MARTINEZ ORGA. *Teoría de lenguajes, gramáticas y autómatas.* Madrid: Ed. Universidad y Cultura, D.L. 1990

[294] R. M. KAPLAN, State Finite Technology En Mathematical Method En *Survey of the State of the Art in Human Languae Technology* Oregon. National Scien Fundation 1995 p. 225-258

Los lenguajes regulares se encuentran en el nivel más bajo de las jerarquías de máquinas y lenguajes. Para cada gramática regular, siempre hay un autómata finito equivalente, que acepta todas las palabras que esa gramática genera. Por lo tanto, una palabra será generada por una gramática sí y sólo sí, si la palabra hace transitar al autómata correspondiente a sus condiciones terminales.

El autómata comienza en el estado 0 y va realizando transiciones según las entradas que se presente hasta llegar al estado final, pasando por tanto estados como considere necesario.

6.2 Definición de máquina de estados finitos

Las máquinas son modelos matemáticos que actúan sobre el problema de transformación de cadenas de símbolos: dada una sucesión de símbolos *x*, transformarla en la sucesión *y* de una forma mecánica[295]. Este problema es una abstracción de muchos problemas comunes. La máquina define las acciones que deben realizarse y para ello sólo se cuenta con número finito de estados.

El uso más común que se les da en informática, es en el ámbito de algoritmos de decisión, esto es, dada una cadena con ciertas características, la máquina decide si una cadena pertenece o no a ese conjunto. Los estados determinan la forma de relacionar los símbolos de la cadena. A ese tipo de máquinas las llamamos *reconocedores* o *autómatas*. Son usadas en el proceso de análisis léxico (autómatas) y sintáctico (para ello se necesita un autómata de pila).

[295] Máquina de estado finito. En *Lenguajes formales* [en línea] <http://www.inf.UDEC.CL/~lenform/02.htm> [Consultado el 10-3-01]

6.3 Diagrama de transiciones

Un diagrama de transiciones es una colección finita de nodos, los cuales se pueden rotular para fines de referencia, conectados por flechas que reciben el nombre de arcos. Cada arco se etiqueta con un símbolo. El diagrama de transiciones tendrá tantos nodos como elementos tiene el conjunto de estados. De ellos, el de partida es el inicial y uno o varios son finales. Una cadena de símbolos es aceptada por el diagrama de transiciones partiendo del estado inicial y recorriendo por los distintos nodos el camino correspondiente hasta llegar al final.

Los diagramas de transiciones pueden emplearse como herramientas de diseño para producir rutinas por ejemplo de análisis léxico, esta misma información se puede dar en forma de tabla (tablas de transiciones).

6.4 Tablas de transiciones

Una tabla de transiciones es una matriz bidimensional cuyos elementos proporcionan el resumen de un diagrama de transiciones. El elemento que se encuentra en la fila *m* y en la columna *n* es el estado que alcanzaría en el diagrama de transiciones al dejar el estado *m* a través del arco con la etiqueta *n*. Si no existe el arco *n* que salga del estado *m* entonces la casilla se corresponde con un error[296]. Cuando hay muchos estados y transiciones es preferible usar las tablas de transiciones a los diagramas.

[296] J. G. BROOKSHEAR *Teoría de la computación: Lenguajes formales, autómatas y complejidad*. España: Addison-Wesley Iberoamericana, 1993.

6.5 Tipos de autómatas y máquinas de estados finitos

Una máquina puede estar en un número cualquiera del conjunto finito de estados, de los cuales uno es inicial y por lo menos uno es final. A este dispositivo está unido un flujo de entrada por medio del cual llegan los símbolos de entrada, la máquina tiene la capacidad de detectar los símbolos conforme llegan y basándose en el estado en el que se encuentra y el símbolo recibido, pasar de un estado a otro, o permanecer en el estado en el que se encuentra. La determinación de este cambio o esta permanencia dependerá del mecanismo de control de la máquina.

La complejidad de un autómata está determinada por sus capacidades de transición, su dispositivo de memoria y las capacidades de inspección en su memoria.

Los autómatas los podemos clasificar en autómatas finitos, que a su vez pueden ser deterministas o no deterministas, como explicaremos más adelante, y en autómatas de pila deterministas y autómatas de pila no deterministas.

- Autómatas finitos deterministas: para cada nodo sólo hay una posible transición de cada símbolo. Es muy importante que estas máquinas no tengan ambigüedades[297].

- Autómatas finitos no deterministas: la diferencia que hay con los no deterministas es que para cada nodo puede haber varias posibilidades de camino a recorrer, por lo tanto existe la posibilidad de que se pueda aplicar más de una transición, o que no se pueda aplicar ninguna, como sucede con una máquina que no está completamente definida. Dado un autómata finito no determinista, siempre es posible construir otro

[297] Ibidem.

autómata finito determinista que acepte el mismo lenguaje que el primero[298].

6.6 Aplicaciones de los autómatas al P.L.N.

Dada su simplicidad matemática se han aplicado a tareas de representación y recuperación de la información, pero hasta hace relativamente poco tiempo no se ha empezado a difundir su uso en P.L.N[299],[300].

Actualmente se están aplicando para la construcción de analizadores léxicos[301], a la extracción de información, aunque para ello es necesario identificar previamente que el documento contiene la información relevante[302].

En la actualidad también está creciendo el interés por aplicarlos a las tareas de almacenamiento y acceso a la información de diccionarios[303], ya que las palabras procedentes se pueden compactar y representar utilizando minimizaciones técnicas[304]; también se están aplicando para reconocer y generar palabras[305], y como mostraremos más adelante también son útiles para tareas de conflación del léxico (lematizadores)

---

[298] M. ALFONSECA, J. SANCHO y M. MARTINEZ ORGA (1990) op. cit.

[299] R. M. KAPLAN (1995) op. cit.

[300] Cf. http://www.research.att.com/sw/tools/fsm/ref.html [consultado 10-03-01]

[301] http://protos.dis.ulpgc.es/docencia/seminarios/rit/index.htm [consultado el 9-03-01]

[302] R. M. KAPLAN (1995) op. cit.

[303] M. MOHRI. On Some Applications of Finite-State Automata Theory to Natural Language Processing . Natural Language Engineering, 2:1-20, 1996. [también en línea] <http://www.research.att.com/sw/tools/fsm/jnle.ps > [Consultado el 10-03-01]

[304] Cf. A. W. APPEL and C.J. JACOBSON The world's fastes scrabble program communication of the ACM, 31 (5) p. 572-578.

Cf. C.L. LUCEHESI and T. KWALTOWSKI Application of the finite automata representing large vocabularioes. Software-Practice and Experince, 23 (1) p. 15-30

[305] F. KARLSSON, L. KARTTUNEN Subsentian processing. En Languaje Analysis and understading En Survey of the State of the Art in Human Language Techology p. 111-112

7. Proceso de creación de las reglas.

Esta parte ha sido la más laboriosa e importante del programa. El objetivo era plasmar en una estructura sencilla, algo tan complicado como es parte del lenguaje natural. Las reglas contienen la información lingüística que posteriormente se implementará en la máquina de estados finitos.

Una vez que se habían elegido los sufijos, se comenzó a observar el comportamiento de las palabras que tenían el sufijo elegido. Para ello se hicieron búsquedas en el diccionario invertido del Diccionario en CD-Rom de la Real Academia de la Lengua[306] y búsquedas con truncamiento por la izquierda en el Diccionario en CD-Rom María Moliner[307]. Antes de extraer cada una de las reglas, se iban haciendo grupos con las palabras cuyos lemas habían experimentado un proceso similar. En los casos en los que el sufijo es altamente productivo se han ido haciendo muestreos aleatorios de las palabras. Cada uno de estos grupos se estudiaba por separado, intentando ir de los casos más generales, a los más concretos. Por ejemplo, los casos de los adjetivos deverbales derivados de los verbos de la segunda conjugación, adoptan distinta variante alomórfica que los de la primera conjugación, por lo tanto, el estudio se ha hecho por conjugaciones. En algunos casos también ha habido que tener en cuenta determinados contexto consonánticos, por ejemplo para formar adjetivos con el sufijo –ble, si el verbo termina en –der, habrá que tener en cuenta que el adjetivo no es **comprendible*, sino *comprensible*; si termina en bir, la forma será *descriptible* y no **describible*. Gracias a la observación de las listas de palabras, se despreciaban aquellas cuyo final coincidía con el del sufijo pero que no tenían el sufijo en sí, (*sable* aunque termina en -ble no contiene el sufijo -*ble*) Mediante este proceso se ha ido extrayendo tantas reglas, como se ha ido considerando necesario.

---

[306] Diccionario de la lengua española. 21 ed. Ed. En CD-Rom. Madrid: R.A.E, 1992

[307] M. MOLINER (1996) op. cit.

Cuando se descubría un caso en el que a pesar de tener el sufijo, por distintas causas, fonéticas, dialectales... no había evolucionado igual que la mayoría de las palabras con ese mismo sufijo, y el crear reglas suponía que solo afectara a una o dos palabras, para evitar multiplicar el número de reglas, además con la consideración de que dichas reglas suelen ser muy complicadas e introducen ruido, al solaparse con otras, se consideró a dichas palabras como excepción y se lematizaron de manera manual. Hay que tener en cuenta que estas palabras han sido muy pocas. Esto es lo que ocurre con las palabras que conservan su raíz latina como por ejemplo *"acuoso*", procede del término latino *"aqua"* y que hubiera sido necesario crear reglas para cambiar la *c* por la *g* hasta llegar al término "*agua*", para evitar esto simplemente se introdujo en la tabla palabras "*acuoso*" y como lema "*agua*".

Para el estudio de los verbos se utilizó la clasificación de Alonso Moro[308]. Se comenzó estudiando los paradigmas verbales regulares, posteriormente los irregulares. En el caso de los irregulares había que tener en cuenta varias cosas: en primer lugar que tiempo, persona y número era el que variaba y ver también si esa irregularidad afectaba a varios verbos o a uno solo, en cuyo caso no era práctico hacer una reglas, sino que simplemente se incluía la forma en la lista de palabras lematizadas manualmente.

Hay que señalar que al no tener en cuenta las tildes ortográficas el número de reglas se reduce considerablemente ya que la única manera de diferenciar algunos tiempos es mediante los acentos.

Hay que tener en cuenta que en el proceso de derivación también intervienen leyes fonéticas que son difíciles de controlar de manera automatizada. Por lo que en muchos casos es preferible incluir estas formas como excepción antes que crear una serie de reglas que solo sirvan para un caso.

---

[308] J. ALONSO MORO. *Verbos españoles* Madrid: Difusión, 1989

Con toda la información obtenida en el apartado anterior, se fueron creando los distintos estados de la máquina de estados finitos. La máquina es no determinista, ya que una vez que elige la terminación, el orden de aplicar las reglas es aleatorio. La información, se fue guardando en distintas tablas enlazadas.

Tabla número 1: en esta tabla se guarda toda la información de las reglas. La tabla tiene cinco columnas, en cada fila la información que hay es la siguiente.

- Sufijo al que se refiere el conjunto de reglas.
- El estado en el que está la máquina.
- El paso a donde llega la operación desde el nodo origen.
- La operación que se realiza; ésta puede ser sumar (+), quitar algo (-) o dejar en el estado en el que se encuentra (=)
- Texto: el texto que se añade o que se suprime.

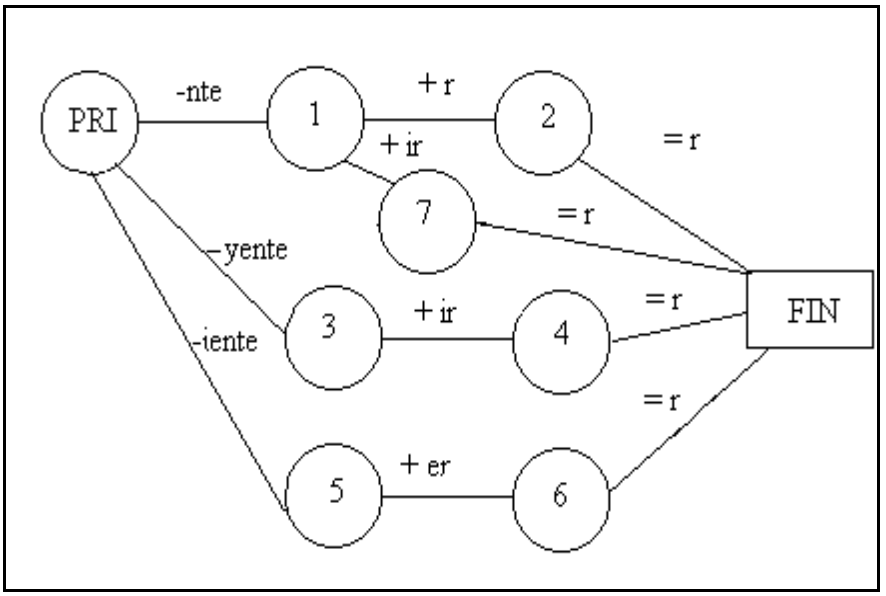
Cada fila de la tabla se considera una regla, el lematizador, como explicaremos más adelante contempla 3692 reglas para la lematización derivativa, y 2700 para la lematización simplemente flexiva.

Como ejemplo, incluimos aquí las reglas correspondientes al sufijo *–nte* con las que se forman nombres a partir de verbos.

Nombre del sufijo	**Estado de que parte**	**Estado al que llega**	**Operación**	**Texto**
nte	Pri	1	-	nte
nte	1	2	+	r
nte	2	Fin	=	ar
nte	Pri	3	-	yente
nte	3	4	+	ir
nte	4	Fin	=	r
nte	Pri	5	-	ente
nte	5	6	+	er
nte	6	Fin	=	r
nte	1	7	+	ir
nte	7	Fin	=	r

Tabla 13 Reglas de -nte

Esta misma información se puede representar con un diagrama de transiciones.



Dibujo 6 Diagrama de transiciones de la reglas de –nte.

Tabla número 2: esta tabla tiene dos campos, en uno se indica el nombre del sufijo y en el otro, (campo lógico), se indica si ese sufijo es flexivo o es derivativo. Esta información servirá para mediante consultas, seleccionar solo aquellos sufijos que pertenecen a una de las dos categorías, y establecer una lematización sólo flexiva o más completa.

En el caso de la lematización flexiva se tienen en cuenta 88 sufijos y para la lematización derivativa 230.

8. Lematización manual.

Con el fin de tener una base de datos que contuviera lemas, se eligió una base de datos con unas 2000 palabras básicas para un primer nivel de aprendizaje del español[309]. A las palabras se le fueron añadiendo los lemas de manera manual,

[309] Fruto de un trabajo de investigación dirigido por J. Manuel Bustos Gisbert destinado a definir el nivel léxico correspondiente al examen para la obtención del Diploma Básico del Español, expedido por el Ministerio de Educación y Cultura español y gestionado por la Universidad de Salamanca.

según los criterios previamente establecidos[310]. La información obtenida en esta parte se plasmó en las dos tablas siguientes.

Tabla número 3: esta tabla contiene el diccionario o tabla de búsquedas[311]. Tiene dos campos, uno para las palabras y otro para los lema correspondientes a esas palabras. La información de esta tabla se ha ido completando tanto con las palabras lematizadas manualmente como con las palabras que ha ido lematizando el programa.

En el caso de la lematización solo flexiva este diccionario no se tuvo en cuenta, pero en el caso de la lematización derivativa contiene 24414 palabras.

Tabla número 4: en esta tabla se contienen los lemas.

En el caso de la lematización derivativa esta tabla contiene 14577 lemas, muchos menos que en caso de la lematización flexiva, que se tuvieron en cuenta 79937, como explicaremos más adelante.

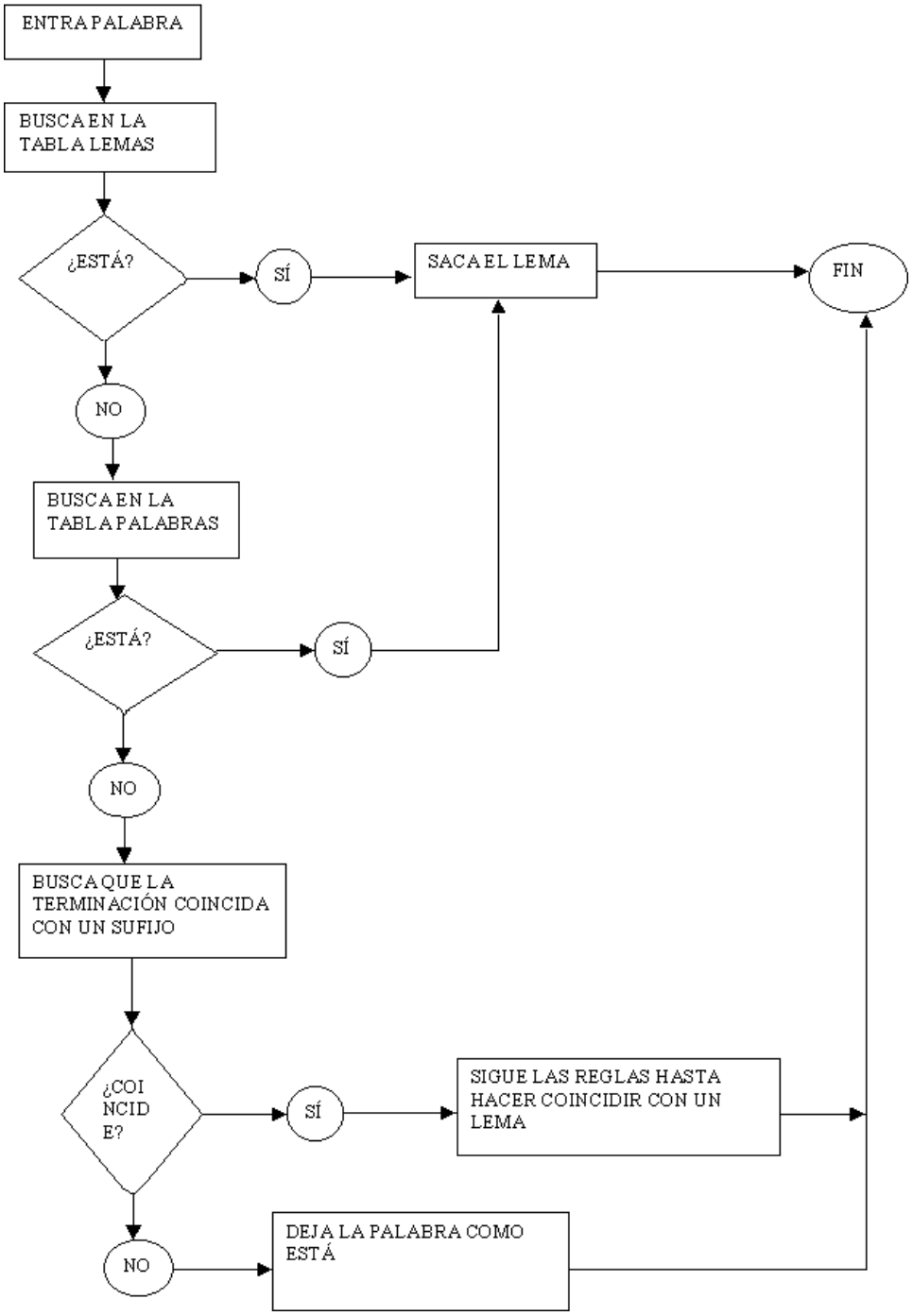
9. Funcionamiento del lematizador.

El programa comienza comprobando si el término a lematizar está incluido en la lista de lemas. En caso de que así sea el lema de dicha palabra coincidirá con el término a lematizar. Cuando la palabra a lematizar no coincide con un lema, el programa chequea el diccionario por si la palabra ya hubiera sido

[310] Ver parte correspondiente
[311] W. B. FRAKES (1992) op. cit.

lematizada, bien manualmente o por el programa en una ocasión anterior. En caso de que así sea, el lema será el que aparezca junto a la palabra, en el caso de que no sea así, el programa buscará que el final coincida con el final más largo que sea posible de la tabla donde están los sufijos. Una vez que lo encuentra irá aplicando las distintas reglas. Cuando llegue al final de la regla buscará que el resultado coincida con un lema, si no coincide con una palabra, si tampoco coincide volverá a buscar el sufijo más largo y repetirá este proceso tantas veces sea necesario hasta que resultado coincida bien con un lema, con una palabra o su final no coincida con ningún sufijo más. En este caso la palabra no se podrá lematizar y se considerará que dicha palabra es igual a su lema.

Esto lo podemos ver en el siguiente diagrama de flujos:



Dibujo 7 Diagrama de flujos del lematizador

Esto mismo lo podemos ver en el siguiente esquema:

1 ¿La palabra introducida coincide con un lema?

- Sí → Final del proceso: El lema es igual a la palabra.
- No → paso 2.

2 ¿La palabra coincide con una entrada de la tabla palabras?

- Sí → Final del proceso: El lema será el que aparezca junto a la palabra.
- No → paso 3.

3 ¿El final de la palabra coincide con algún sufijo?

- Sí → Elige el más largo y pasa al paso 4.
- No → Final del proceso: El lema será igual a la palabra.

4 Se van aplicando las reglas correspondientes al sufijo elegido, ¿llega al final de la regla?

Sí → paso 1.

No → paso 5.

5 ¿Hay otro sufijo coincidente con el final de esa palabra?

- Sí → paso 3.

- No → Final del proceso: El lema será igual a la palabra introducida por lo tanto incluimos este resultado en el diccionario del programa.

Con el fin de aclarar esto veamos un ejemplo:

Supongamos que queremos lematizar la palabra *zapateros*,

1 ¿La palabra introducida coincide con un lema?.

No coincide con ningún lema, por lo tanto pasamos al paso 2

2 ¿La palabra coincide con una entrada de la tabla palabras?

Consideramos que esta palabra no ha sido lematizada con anterioridad, por lo tanto pasamos a comprobar si su final coincide con algún sufijo.

3 ¿El final de la palabra coincide con algún sufijo?

Al comprobar esto nos damos cuenta de que coincide con los finales –s, -os, de toda la lista de sufijos. Como de éstos el más largo es –os, intentará aplicar las reglas.

4 Se van aplicando las reglas correspondientes al sufijo elegido, ¿llega al final de la regla?

Comprobamos que no puede llegar al final de la regla, puesto que la terminación os es una desinencia verbal, por lo tanto este sufijo no es adecuado, así es que habrá que buscar otro.

5 ¿Hay otro sufijo coincidente con el final de esa palabra?

En este caso elige el sufijo –s y aplica las reglas y llega a la palabra *zapatero*. Comprueba si esta palabra coincide con una entrada de la tabla lema, como no lo es, pasa a comprobar si ya ha sido lematizada alguna vez, consideramos que no lo ha sido, por lo tanto pasamos a comprobar el final más largo con el que coincide, en este caso –ero. Intentamos aplicar las reglas correspondientes a dicho sufijo y comprobamos que llega al resultado *zapato.* Comprobamos que esta palabra coincide con un lema, por lo tanto el lema de *zapateros* es *zapato.* Finalmente consideramos este resultado y lo añadimos al diccionario del programa para sucesivas lematizaciones.

10. Fases del lematizador.

El programa tuvo dos fases.

10.1 Fase uno del lematizador

10.1.1 Funcionamiento

En la primera se fueron buscando los ejemplos de manera manual. Se eligió este método ya que era necesario probar todos los sufijos sabiendo que los lemas estuvieran incluidos en la base de datos. A la vez que se iban probando las reglas y analizando los distintos fallos, para detectar donde faltaban reglas o

dónde éstas eran incorrectas. En esta primera fase se fueron añadiendo muchos lemas de manera manual a la tabla lemas.

10.2 Fase dos del lematizador

10.2.1 Funcionamiento

Cuando se consideró que el programa tenía un buen porcentaje de aciertos (aproximadamente un 80%), se pasó a probar el programa con bloques de texto, en lugar de palabra a palabra. Para ello se tomaron textos de noticias del CD del periódico El Mundo[312]. Se creó una tabla con un campo único, donde se fueron incluyendo todas las palabras: tabla "entrada", y se adaptó el programa para que fuera leyendo uno a uno los registros de la tabla entrada y escribiendo el resultado en la tabla "salida", creada previamente. Esto permitía probar con un gran número de palabras al mismo tiempo.

La única diferencia de funcionamiento, respecto de la fase anterior, es que en este caso no pregunta si el lema es correcto o no. El programa comprueba si la palabra coincide con alguna de las entradas de la tabla "lemas", si es así, el lema será igual a la palabra y pasa a la siguiente palabra. En el caso de que la palabra no coincida con ningún lema, buscará en la tabla palabras por si con anterioridad la palabra ya hubiera sido lematizada, en caso afirmativo, escribirá el lema que allí aparezca. En el caso de que la palabra no haya sido lematizada con anterioridad, buscará que su final coincida con el sufijo más largo de la lista. En el caso de que su final no coincida con ninguno de los sufijos, escribirá como lema la misma palabra. Si coincide con uno de los sufijos, aplicará las reglas diseñadas para este sufijo con el fin de encontrar el lema. El resultado lo escribirá en la tabla salida, y así sucesivamente hasta terminar con todas las palabras.

---

[312] El MUNDO. Primer semestre. Textos íntegros. Mundired Servicio electrónico. [CD] D.L. 1994 EL MUNDO ISBN 84-920059-1-2

10.2.2 Análisis de resultados

Una vez hecha la lematización se analizó la tabla salida para comprobar cuantas palabras se habían lematizado correctamente y calcular el porcentaje de aciertos y fallos.

El conjunto de artículos elegido contenía un total de 81.988 palabras, de las cuales tan solo 9144 son distintas. Del total de palabras, 76.939 se lematizaron correctamente (94% del total), es decir 9144 palabras únicas (91% del total) frente a 5.049 errores del total de palabras, lo que son 950 palabras únicas, mal lematizadas.

En la siguiente tabla vemos la distribución de estas palabras en aciertos y fallos teniendo en cuenta todas las palabras y las únicas:

	TOTAL	% TOTAL	ÚNICAS	% ÚNICAS
ACIERTOS	76939	94%	9144	91%
FALLOS	5049	6%	905	9%
TOTAL	81988		10049	

Tabla 14 Distribución de aciertos y fallos todas las palabras. Fase 1

En este primer gráfico vemos como se distribuyen los aciertos y los fallos del total de palabras , teniendo en cuenta las palabras vacías.

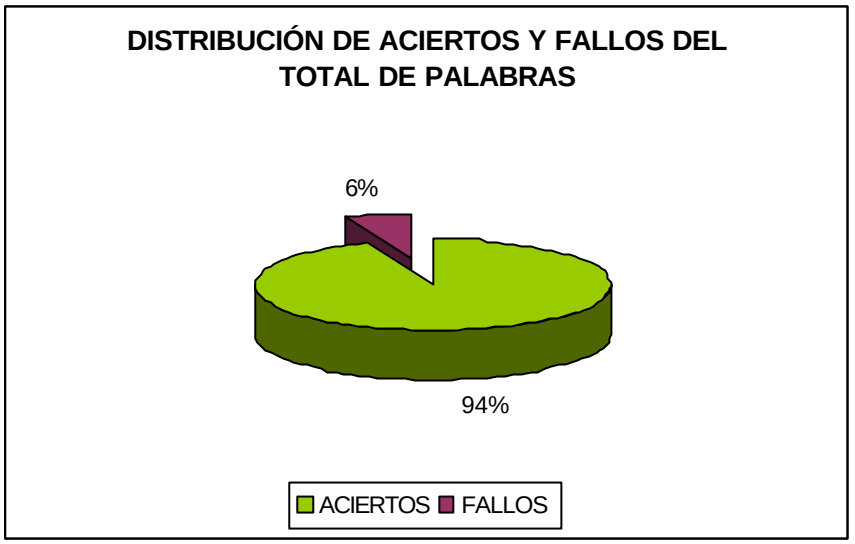


Gráfico 2 Distribución de aciertos y fallos del total de palabras.

Podemos observar como el porcentaje de las palabras mal lematizadas teniendo en cuenta el total de palabras, es muy pequeño, tan solo un 6% del total.

En este segundo gráfico vemos como se distribuyen los aciertos y los fallos del total de palabras únicas.

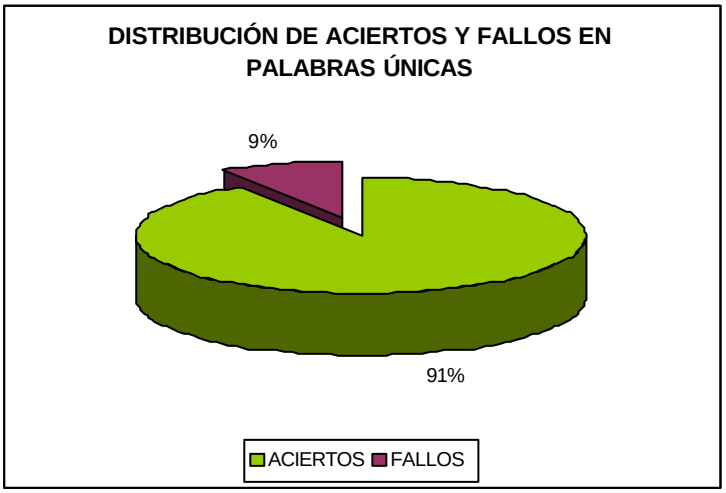


Gráfico 3 Distribución de aciertos y fallos en palabras únicas

Podemos observar como el porcentaje de los fallos aumenta un 3%; de aquí podemos deducir que la mayor parte de las palabras que no se lematizan correctamente son palabras con un índice de frecuencia muy bajo. La mayor parte, probablemente, correspondan a nombres propios y palabras extranjeras que no deben ser lematizados y por coincidencia del final se lematizan erróneamente.

Como las palabras vacías no se consideran a la hora de la recuperación, no las tuvimos en cuenta a la hora de medir estos resultados. De estos datos lo primero que vemos es que las palabras a tener en cuenta se reducen a un 54% del total. De estas palabras vemos, al igual que en el caso anterior la distribución de aciertos y fallos.

	TOTAL	**% TOTAL**	**ÚNICAS**	**% ÚNICAS**
ACIERTOS	41050	93%	8869	91%
FALLOS	3260	7%	882	9%
TOTAL	44310		9751	

Tabla 15 Distribución de aciertos y fallos sin contar las palabras vacías. Fase 1

Vemos, como al no tener en cuenta las palabras vacías, al considerar todas las palabras el porcentaje de palabras lematizadas de manera incorrecta se reduce.

Veamos esta información en gráficos.

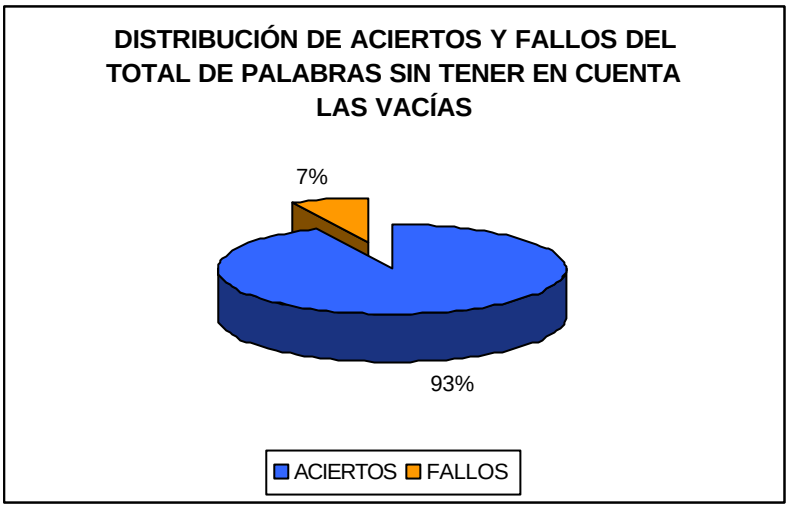


Gráfico 4 Distribución de aciertos y fallos palabras únicas y sin las vacías

Podemos observar en el siguiente gráfico como al tratarse de las palabras únicas sin tener en cuenta las vacías, los porcentajes no varían respecto a tenerlas en cuenta.

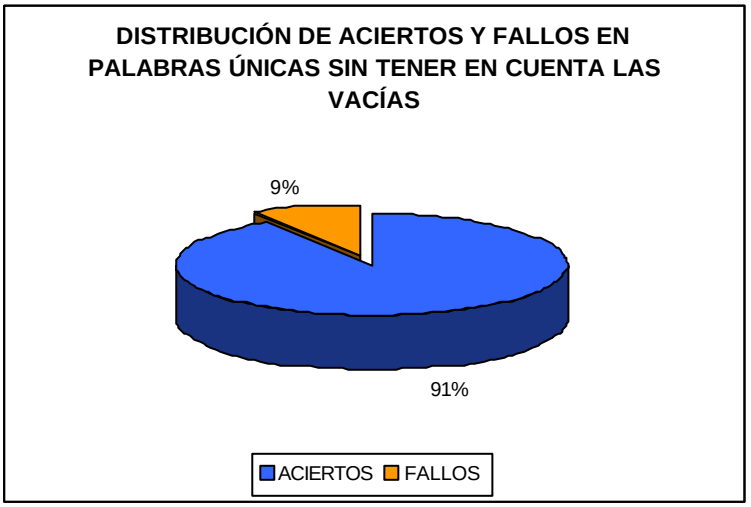


Gráfico 5 Distribución de aciertos y fallos en palabras únicas sin vacías

Una vez analizados estos porcentajes, se hizo un análisis de los fallos para detectar si los errores estaban en las reglas, los lemas o en las palabras con el fin de depurarlos. De este modo el lematizador quedó listo para su aplicación a la recuperación de información.

11. Aplicación del lematizador a la R.I.

Una vez que se consideró que la lematización tenía un porcentaje de errores aceptable, se decidió intentar mostrar si la lematización mejora la efectividad de la recuperación de la información, que es uno de los objetivos concretos de este trabajo.

Antes de realizar el experimento era necesario buscar la base de datos, crear las preguntas y hacer los juicios de relevancia.

11.1 La base de datos

La base de datos fue la misma empleada en el trabajo de Grado[313]: DATATHÈKE[314]. Esta base de datos fue creada por los Carlos García Figuerola y José Luis Alonso Berrocal, profesores del área de informática de la facultad de Traducción y Documentación de la Universidad de Salamanca. En ella hay un vaciado de artículos de revistas relacionadas con el mundo de la biblioteconomía, la documentación, la archivística, la traducción y la informática, que fue realizado por distintos alumnos, becarios y personal de dicha facultad.

Los motivos de la elección fueron los mismos que entonces: que todos los registros estuvieran en español, con un tamaño medio y que la extensión de los mismos fuera parecida. A estos motivos se añadía que era una base de datos que ya había sido depurada, quitando duplicados y registros muy cortos, para el Trabajo de Grado. La única modificación que se hizo en la base de datos respecto de la utilizada en el Trabajo de Grado fue la normalización de los números registro de los documentos.

---

[313] R. GÓMEZ (1998) op. cit.

[314] http://milano.usal.es/dtt.htm

La base de datos contiene 1117 artículos con una longitud media de 468 bites cada uno de ellos, lo que son aproximadamente unas 100 palabras por artículo.

11.2 Las preguntas y la relevancia

Al igual que en el caso anterior, nos basamos en el Trabajo de Grado[315].

Para el Trabajo de Grado ya se habían creado quince preguntas de los temas de la base de datos y la relevancia había sido calculada manualmente por dos personas, para evitar subjetividad. Para este trabajo decidimos mantener las preguntas pero la relevancia se revisó nuevamente, para depurar posibles errores. A pesar de los esfuerzos por intentar que los juicios sean lo más objetivos posible, esta tarea no siempre es fácil porque hay que considerar que algunos resúmenes son cortos, y no dan mucha información, y no todos están redactados con la misma claridad.

También hay que tener en cuenta que la relevancia se ha basado en el contenido, es decir que se consideran como relevantes los documentos que responden a las preguntas, aunque no tengan ninguno de los términos de la pregunta en el documento. Por ejemplo, como se verá a continuación la pregunta número siete hace referencia a *bibliotecas de centros educativos*, algunos artículos hablan de *bibliotecas escolares*, que aunque no es el mismo concepto, estas últimas se engloban en las primeras, y por lo tanto los documentos con estos términos son susceptibles de se relevantes a esa pregunta, pero si en la pregunta no se indica expresamente el término, la similaridad en la respuesta, en el caso de que el documento se recupere por los otros términos, será menor. Esto hace que

[315] R. GÓMEZ (1998) op. cit.

algunos de los documentos no se puedan recuperar sin utilizar un mecanismo de realimentación de relevancia.

A continuación indicamos las preguntas con el número de documento relevantes para ellas.

1. - *La formación continuada de los profesionales de archivos, bibliotecas y museos.*

Para esta pregunta encontramos un total de 16 documentos relevantes.

2.- *La gestión de los recursos humanos, administrativos, económicos y financieros de los centros de documentación, archivos y bibliotecas.*

24 documentos relevantes.

3.- *Las actividades de animación a la lectura y de expansión de la biblioteca infantil y juvenil.*

27 documentos relevantes.

4.- *¿En qué consiste la gestión de calidad y cómo influye en los proyectos de planificación y gestión de las bibliotecas, archivos y museos españoles?.*

14 documentos relevantes.

5.- *¿Qué funciones son propias del bibliotecario y cuales del auxiliar. Qué función tiene cada uno de ellos?*

7 documentos relevantes.

6.- *¿Qué salidas ofrece el mercado de trabajo para los licenciados y diplomados en documentación?*

12 documentos relevantes.

7.- *Colaboraciones entre bibliotecas públicas y bibliotecas de centros de educación primaria y secundaria.*

14 documentos relevantes.

8.- *Cómo se puede llevar a cabo la formación de usuarios y qué experiencias hay en bibliotecas, centros de documentación y de información especializados.*

14 documentos relevantes.

9.- *¿Cuál es el perfil de los profesionales en el mundo de la información: bibliotecarios, archiveros y documentalistas?.*

15 documentos relevantes.

10.- *Fuentes para la selección de obras de literatura infantil y juvenil.*

15 documentos relevantes.

11.- *Las fuentes de información en soporte óptico.*

13 documentos relevantes.

12.- *¿Qué es el marketing?. Aplicación en bibliotecas, archivos, museos, centros de documentación e información.*

8 documentos relevantes.

13.- *La edición electrónica*

16 documentos relevantes.

14.- *La industria editorial en España.*

26 documentos relevantes.

15.- *El impacto de Internet en el mundo de la difusión de la información.*

12 documentos relevantes.

11.3. El sistema de recuperación

El sistema de recuperación tiene tres pasos: el proceso de lematización, el de indización y el de recuperación.

11.3.1 Proceso de lematización

Lo primero es lematizar por una parte los documentos, y por otra las preguntas. Para ello se introducen las palabras indicando para cada palabra el número de documento al que pertenecen y el campo en el que está. Se lematizan las palabras según lo visto en el apartado anterior y se vuelven a agrupar las palabras por documentos o por preguntas. En lugar de las palabras, tenemos los documentos y las preguntas formados por los lemas de las palabras, con las palabras vacías o sin ellas, dependiendo del experimento que se trate, según indicaremos más adelante. Con las preguntas se hace lo mismo.

11.3.2 Proceso de indización

En el caso de la lematización, se toman los lemas y sobre ellos se establece la indización. Para el proceso de indización es necesario establecer una serie de cálculos.

- Frecuencia del término en cada documento: la información que tiene esta tabla es el número de documento al que pertenece la palabra, la palabra y el número de veces que aparece dicha en ese documento.

- Suma de los pesos de los términos en cada documento: para asignar los pesos, se multiplica la frecuencia del término en el documento (tf), por el número de veces que ese término aparece en la base de datos.

- Indice de frecuencia inversa (IDF) para cada documento Se calcula mediante la siguiente ecuación:

$$IDF = Ln\, N/ndoc + 1$$

Ecuación 17 Cálculo del idf

Donde N representa el número de veces que aparece un término en toda la base de datos y ndoc el número de documentos

- Frecuencia del término de consulta: en esta tabla se guarda la misma información que en la frecuencia de término en cada documento, pero en este caso de las preguntas.

- El sumatorio de los pesos de cada término en la consulta. Esta tabla contiene la información del peso de cada término de la pregunta.

El mismo proceso que se sigue en la indización de los documentos se emplea para las preguntas. Sobre esta indización, se establece la recuperación.

11.3.3 Proceso de recuperación

Para la recuperación se establece la comparación entre los documentos y las preguntas. El programa ofrecerá una salida ordenada de los documentos en función de la similaridad. Recordemos que la ecuación de la similaridad es:

$$Similaridad(d_j q_k) = \frac{\sum_{i=1}^{n} (td_{ij} \bullet tq_{ik})}{\sqrt{\sum_{i=n}^{n} td_{ij}^{\ 2} \bullet \sum_{i=1}^{n} tq_{ik}^{\ 2}}}$$

Ecuación 18 Similaridad Harman[316]

12. Los experimentos.

Con el fin de mostrar qué tipo de lematización era la más efectiva, si la derivativa o la flexiva, en qué momento era más eficaz la supresión de palabras vacías, si antes de hacer la lematización o después y qué lista era la más adecuada, si una con categorías gramaticales vacías de contenido y algunos verbos, o también incluyendo palabras con alta frecuencia de aparición en textos en español, hemos realizado una serie de experimentos que hemos agrupado en tres grandes bloques: sin lematizar, lematización derivativa y lematización flexiva.

[316] D. HARMAN (1992) op. cit.

12.1 Sin lematizar

De esta manera hemos denominado a los experimentos que no aplican ningún tipo de lematización. El único tratamiento que aplican es la normalización de las palabras. Dicha normalización consiste en suprimir los signos de acentuación, y puntuación, y poner todo el texto en mayúsculas. Dentro de estos experimentos hemos probado con la supresión de las palabras vacías según las dos listas antes especificadas. De este modo dentro de este grupo tenemos tres experimentos:

- Con vacías: es decir sin suprimir ninguna de las listas de palabras vacías
- Sin vacías leve: suprimiendo la lista de palabras vacías basada en las categorías gramaticales vacías de contenido
- Vacías fuerte: suprimiendo la lista de palabras vacías que además de las categorías gramaticales suprime también aquellas que son más frecuentes en español.

La principal misión de estos experimentos es la de servir de control con respecto a los otros dos bloques de experimentos.

12.2 Lematización derivativa

Este tipo de lematización es la más agresiva, con ella se pretende que la tasa de exhaustividad sea más alta respecto a aquellos experimentos que no aplican lematización y también respecto de aquellos otros que aplican lematizaciones más leves. Para poder realizar estos experimentos ha sido necesaria la construcción del lematizador según hemos explicado con anterioridad. Este lematizador utiliza la lista más larga de sufijos 230, con 3692 reglas asociadas ellos. En cuanto a la tabla lemas, tiene 14577 elementos. La tabla

palabras 24414 que ya habían sido introducidas bien de manera manual, o automáticamente con las pruebas previas de lematizador.

12.3 Lematización flexiva

Este tipo de lematización no es tan radical como la anterior. Está basada en los experimentos de Harman[317] y de Kroventz[318], con ella pretendíamos mostrar si al igual que ocurre con el inglés es mejor una lematización menos agresiva que la que se consigue con un lematizador derivativo. La lematización flexiva tan solo normaliza los plurales y el género y reduce los tiempos verbales a la raíz. Este lematizador es mucho más sencillo que el que hace lematización derivativa, su lista de sufijos es más corta, sólo contempla 88 sufijos[319]. El número de reglas que necesita también es menor, 2700. En cambio necesita una tabla de lemas mucho más amplia que el otro lematizador. En este caso, como lista de lemas empleamos todas las entradas únicas del diccionario de la Real Academia[320], salvo las que se refieren a los afijos, expresiones, en total se tuvieron en cuenta 79937 lemas.

En el caso de la lematización flexiva no empleamos la tabla de palabras puesto que el lematizador flexivo no lo habíamos probado anteriormente, ya que las reglas de los verbos son iguales para los dos lematizadores y la regla de cambio del género femenino al masculino y del plural al singular también había sido probada con anterioridad.

---

[317] D. HARMAN (1991) op. cit.

[318] KROVENTZ (1993) op. cit,

[319] Ver lista de sufijos

[320] En algunas ocasiones las palabras con distintas acepciones viene más de una vez, en estos casos solo se introdujo una.

13. La evaluación de los resultados.

Según vimos anteriormente, la evaluación hay que hacerla teniendo en cuenta: la corrección de la lematización, la ejecución de la compresión, la velocidad y la efectividad de la recuperación de la lematización. Los dos primeros parámetros afectan solamente al lematizador, el último sólo a la aplicación del lematizador a la recuperación. La velocidad afecta a ambos ya que puede medirse el tiempo empleado en hacer la lematización, la indización y en obtener la respuesta. En este punto no vamos a entrar ya que la capacidad y la velocidad de procesamiento de los ordenadores está creciendo a una velocidad vertiginosa y por lo tanto creemos que este parámetro carece de importancia.

13.1 Corrección de la lematización

La corrección de la lematización ya ha sido analizada anteriormente, por ello ahora nos vamos a centrar en la compresión y en la evaluación de los resultados del lematizador aplicados a la recuperación de información.

13.2 Compresión

Una de las ventajas importantes que tiene la lematización, como indicamos anteriormente, es que nos permite comprimir la información. Para mostrar la efectividad de la compresión hicimos los siguientes cálculos: comprobamos el número de palabras únicas a lematizar,suprimiendo previamente las palabras vacías según la lista vacías fuerte, y comprobamos a cuantos lemas se reducían, esta cantidad la dividimos entre el número total de palabras.

	Palabras	Lemas	Compresión
Derivativa sin vacías leve	9841	4400	44,71 %
Flexiva sin vacías leve	9841	6560	66,66 %
Derivativa sin vacías fuerte	9752	4419	45,31%
Flexiva sin vacías fuerte	9752	6607	67,75%

Tabla 16 Tasas de compresión

Si comparamos los datos de nuestra lematización, tanto los de la derivativa como los de la flexiva, con los calculados por Harman para su S'Stemmer, el lemazador de Porter y el de Lovins, salvando las distancias de las diferencias de los idiomas, vemos que nuestra tasa de compresión es menor que en el caso del S'Stemmer de Harman, donde la tasa de compresión es del 88% (7489/8460). Se aproxima a la del algoritmo de Porter, que es de un 69% (6028/8460), y es ligeramente superior a de de Lovins que es de un 61% (5226/8460)[321]. En cuanto a la lematización flexiva, como es de esperar la compresión es menor.

13.3 Evaluación de la recuperación

Para evaluar la aplicación del lematizador a la R.I., hemos seguido la propuesta de Salton, es decir, hemos ido calculando la precisión y la exhaustividad de los experimentos, para cada experimento, pregunta a pregunta, después hemos calculado las medias de los resultados de cada uno de los

[321] Los datos están tomados de D. Harman (1991) op. cit.

experimentos, como mostraremos a continuación en las siguientes tablas y gráficos[322].

Para el análisis de los resultados, hemos seguido la misma clasificación que utilizamos en la explicación de los mismos: los que no aplican ningún tipo de lematización, (*sin lematizar*) los correspondientes a la lematización flexiva y derivativa, (*lematización derivativa*) y los que sólo aplican lematización flexiva (*lematización flexiva*). Dentro de cada grupo se ha probado sin suprimir las palabras vacías y suprimiendo según la lista de *vacías leve* y *vacías fuerte* antes y después de lematizar en el caso de la lematización derivativa y flexiva.

En primer lugar comprobaremos la precisión media de cada experimento, después la exhaustividad y finalmente pondremos en relación las dos medidas con el fin de determinar qué experimento es el que obtiene mejores resultados.

13.3.1 Precisión

Recordemos que con esta medida pretendemos evaluar la correlación de la pregunta con la base de datos[323]. La ecuación que hemos utilizado es[324] :

$$\text{Precisión} = \frac{\text{Documentos relevantes recuperados}}{\text{Documentos recuperados}}$$

Ecuación 19 Precisión.

Después de cada grupo de resultados haremos la representación gráfica de los mismos marcando en el eje de las *x* el número de documentos y en el de las

[322] Para detalles referidos a cada consulta ver el apéndice de resultados.

[323] KOWALSKI (1987) op. cit.

[324] G. SALTON (1983) op. cit.

y los valores de precisión con el fin de comparar los resultados por las curvas descritas por los resultados.

13.3.1.1 Precisión media sin lematizar

En estos experimentos simplemente se han normalizado los términos, para ello se han suprimido los signos de acentuación y puntuación. No se ha aplicado ningún tipo de lematización.

En la siguiente tabla se muestran las medias de los resultados obtenidos con los 10, 20, 30, 50 y 100 primeros documentos recuperados, simplemente normalizando las palabras y suprimiendo las palabras vacías según la lista *vacías leve* y *vacías fuerte.*

Documentos recuperados	**Con vacías**	**Sin vacías leve**	**Sin vacías fuerte**
10	0,4733	0,4933	0,4866
20	0,3466	0,3466	0,3466
30	0,2622	0,2666	0,2688
50	0,2	0,2	0,2013
100	0,1186	0,1211	0,1222

Tabla 17 Precisión de los experimentos *sin lematizar.*

Como puede apreciarse tanto en la tabla anterior, como en el gráfico siguiente, tan solo se encuentran diferencias significativas hasta aproximadamente los primeros 20 documentos recuperados. A partir de este punto, las líneas prácticamente se solapan.

En estos 20 primeros documentos recuperados, vemos como los peores resultados son aquellos que no suprimen las palabras vacías y los mejores aquellos que sí las suprimen según la lista vacías leve. Por lo que según este gráfico podemos afirmar que cuando no se realiza lematización al suprimir las palabras vacías obtenemos una precisión mayor que si no las suprimimos.

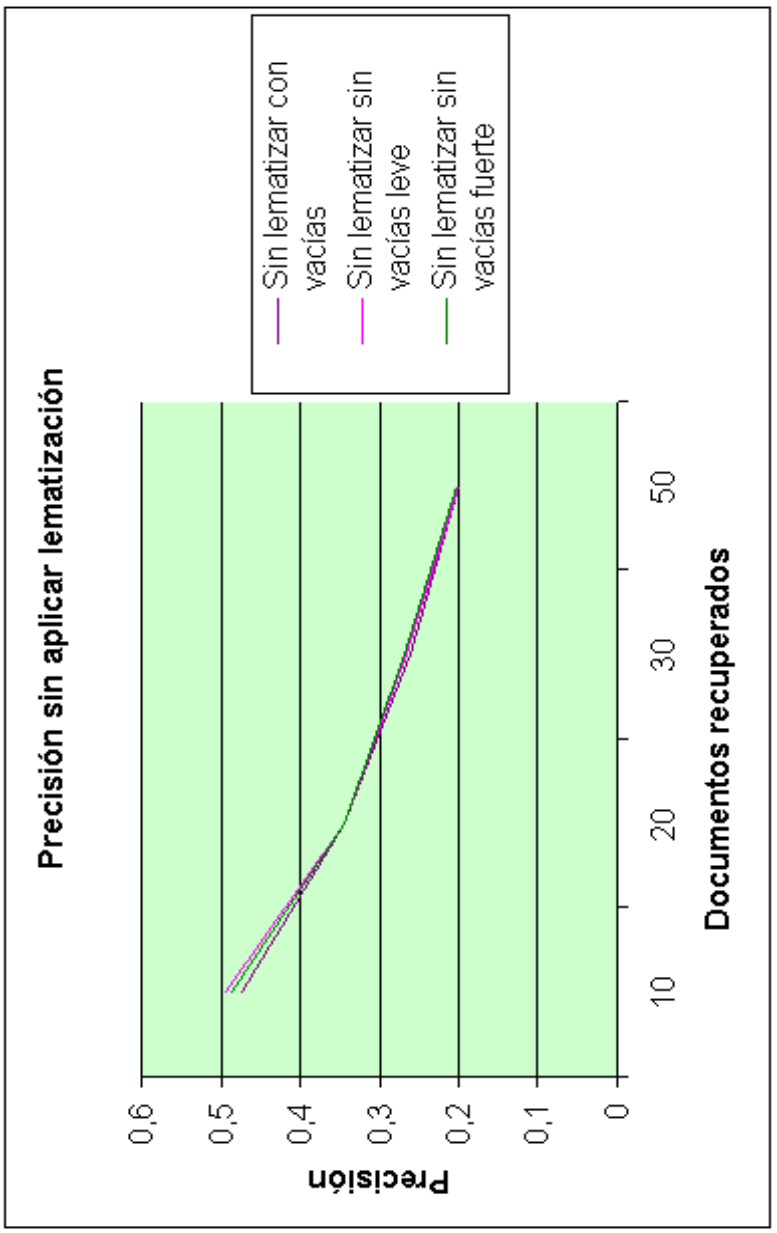


Gráfico 6 Precisión *sin lematizar*

En los dos grupos de experimentos siguientes, hemos probado suprimiendo las palabras vacías antes y después de lematizar con el fin de mostrar si el momento de la supresión de las palabras vacías afecta o no en la precisión de los documentos recuperados.

13.3.1.2 Precisión de la lematización derivativa

En la siguiente tabla se muestran las medias de los resultados obtenidos con los 10, 20, 30, 50 y 100 primeros documentos recuperados, aplicando lematización derivativa, y combinando esta, con la supresión de palabras vacías según la lista *vacías leve* y *vacías fuerte,* antes y después de lematizar, y manteniendo dichas palabras vacías.

Documentos recuperados	**Con vacías**	**Sin vacías leve después**	**Sin vacías fuerte después**	**Sin vacías leve antes**	**Sin vacías fuerte antes**
10	0,4666	0,4866	0,4933	0,48	0,5
20	0,36	0,3566	0,3533	0,36	0,3533
30	0,2777	0,2755	0,2777	0,27	0,2822
50	0,204	0,2013	0,204	0,2	0,2053
100	0,1253	0,1253	0,1246	0,1246	0,124

Tabla 18 Precisión *lematización derivativa*.

Como puede observarse por la tabla de resultados, en lo que se refiere a la precisión, a penas hay diferencias entre suprimir las palabras vacías antes y después de lematizar, tanto para la lista vacías leve como vacías fuerte. Por este

motivo, y con el fin de que los gráficos resulten más claros, no tendremos en cuenta los experimentos que suprimen las palabras vacías después de lematizar.

Veamos el siguiente gráfico.

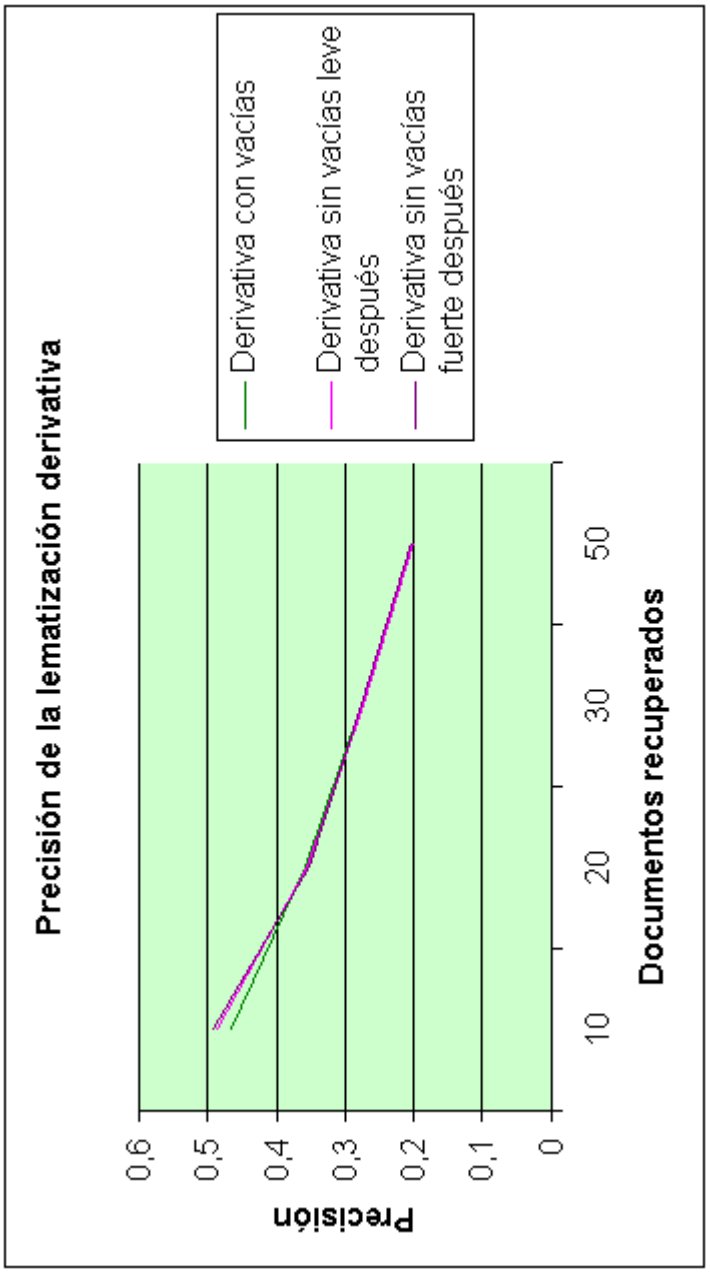


Gráfico 7 Precisión *lematización derivativa*

Como podemos ver por el gráfico, en los primeros documentos recuperados, es donde hay diferencia, pero a partir de los 15 documentos recuperados, las líneas prácticamente se solapan, al igual que ocurría en los que no aplican lematización.

En lo que a la lematización derivativa se refiere podemos afirmar, a la vista de los resultados que la supresión de palabras vacías mejora los resultados respecto a los que no las suprimen. En cuanto a que si es mejor hacerlo según la lista de *vacías leve* o de *vacías fuerte* apenas hay diferencia, aunque son ligeramente mejores los de las *vacías fuerte.*

13.3.1.3 Precisión lematización flexiva

En la siguiente tabla se muestran las medias de los resultados obtenidos con los 10, 20, 30, 50 y 100 primeros documentos recuperados, aplicando lematización flexiva, y combinando ésta, con la supresión de palabras vacías según la lista *vacías leve* y *vacías fuerte,* antes y después de lematizar, y manteniendo dichas palabras vacías.

Documentos recuperados	Con vacías	Sin vacías leve después	Sin vacías fuerte después	Sin vacías leve antes	Sin vacías fuerte antes
10	0,5466	0,54	0,54	0,54	0,5466
20	0,3933	0,3933	0,3866	0,39	0,39
30	0,2977	0,3	0,3	0,3	0,2977
50	0,2133	0,2146	0,2146	0,2146	0,216
100	0,1266	0,1285	0,1285	0,1285	0,1285

Tabla 19 Precisión *lematización flexiva*.

Al igual que en el caso anterior, en lo que se refiere a los gráficos no vamos a tener en cuenta los resultados de la supresión de las palabras vacías después de lematizar, ya que como se muestra en la tabla apenas hay diferencias entre los resultados.

Como podemos apreciar tanto por la tabla de resultados como por el gráfico, las diferencias entre suprimir las palabras vacías y no suprimirlas, son muy pequeñas y lo mismo ocurre entre las dos listas de palabras vacías por lo que podemos afirmar que con la lematización flexiva la supresión de palabras vacías apenas incide en la precisión de la recuperación.

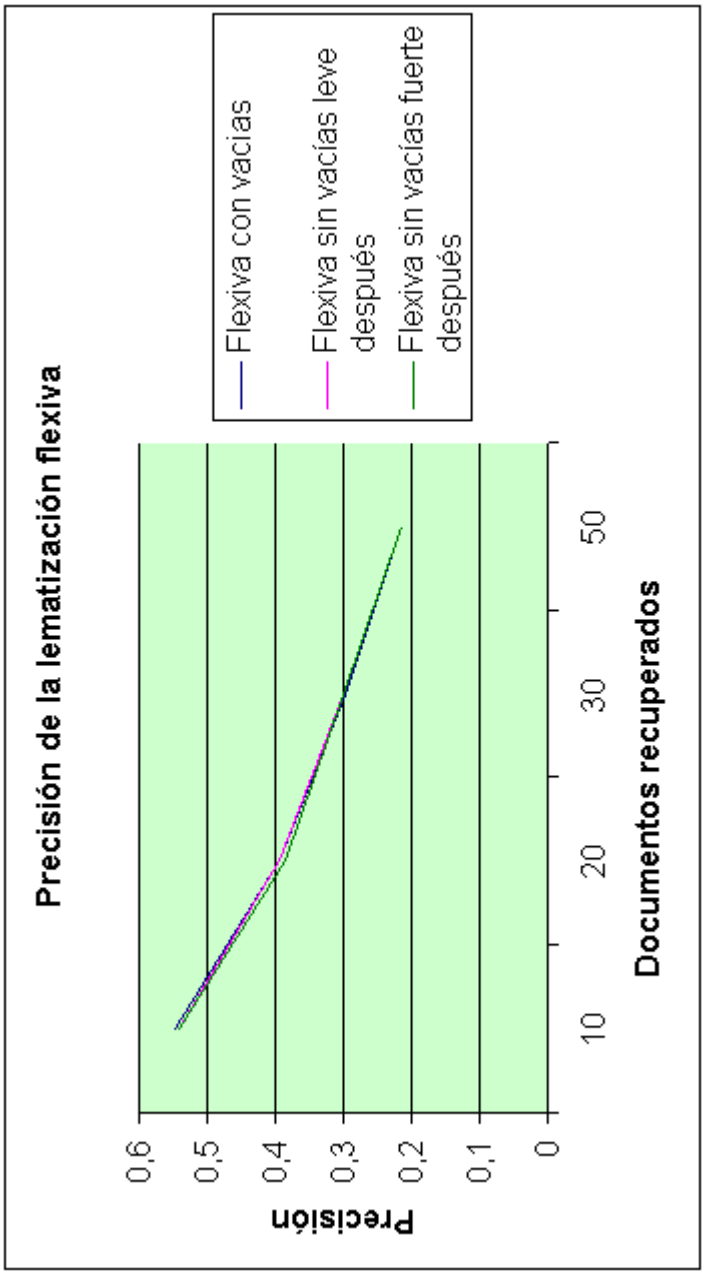


Gráfico 8 Precisión *lematización flexiva*

Con el fin de mostrar cuál de todos los experimentos realizados tiene una precisión mejor, hemos elegido de cada grupo el mejor de ellos para compararlos en el siguiente gráfico

Como podemos observar por el gráfico siguiente la lematización flexiva es la que obtiene unos resultados de precisión más altos. La mayor diferencia se ve en los primeros documentos recuperados. En segundo lugar está la precisión correspondiente al experimento de la lematización derivativa y finalmente, el peor resultado es el de los que no aplican lematización.

Entre la precisión correspondiente a la lematización derivativa y la no lematización, vemos como para los primeros documentos recuperados apenas hay diferencias, pero a partir de los 20 documentos recuperados son mejores los de la recuperación derivativa.

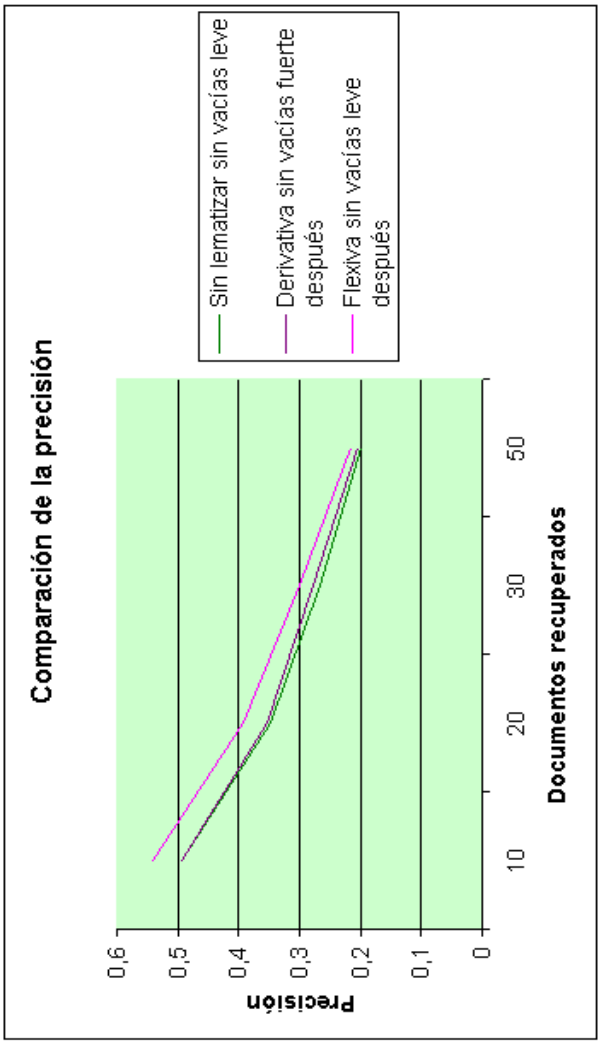


Gráfico 9 Comparación de la precisión

13.3.2. Exhaustividad

Para la exhaustividad haremos lo mismo que hemos hecho con la precisión, agruparemos los experimentos y después compararemos los que obtengan mejores resultados de cada grupo.

Con esta medida pretendemos ver si el sistema es capaz de recuperar solo aquellos documentos que son relevantes a la pregunta, recordemos que esta medida está directamente relacionada con la precisión.

La ecuación empleada es la que aparece en el libro de Salton[325].

$$\text{Exhaustividad} = \frac{\text{documentos relevantes recuperados}}{\text{documentos relevantes}}$$

Ecuación 20 Exhaustividad

Para su representación gráfica marcamos los valores de exhaustividad en le eje de las *y* y los documentos recuperados en el de las *x*. Mediante estos gráficos será más fácil la comparación de los experimentos.

13.3.2.1 Exhaustividad sin lematizar

En la siguiente tabla se muestran los resultados de exhaustividad a los 10, 20, 30 y 50 primeros documentos recuperados, sin aplicar lematización, con palabras vacías, y sin ellas según las dos listas: *vacías leve* y *vacías fuerte.*

[325] G. SALTON (1983) op. cit.

Documentos recuperados	Con vacías	Sin vacías leve	Sin vacías fuerte
10	0,3047	0,3175	0,3133
20	0,4463	0,4463	0,4463
30	0,5064	0,5150	0,5193
50	0,6437	0,6437	0,6480
100	0,7639	0,7639	0,7725

Tabla 20 Exhaustividad *sin lematizar*

Como podemos observar tanto en la tabla de resultados como en el gráfico siguiente, la exhaustividad de los tres experimentos es muy parecida, pero ligeramente inferior en el caso de no suprimir las palabras vacías. Entre suprimir una y otra lista de palabras vacías para los primeros documentos recuperados son mejores lo que suprimen la lista de palabras *vacías leve*, aunque hay que señalar que las diferencias son muy pequeñas.

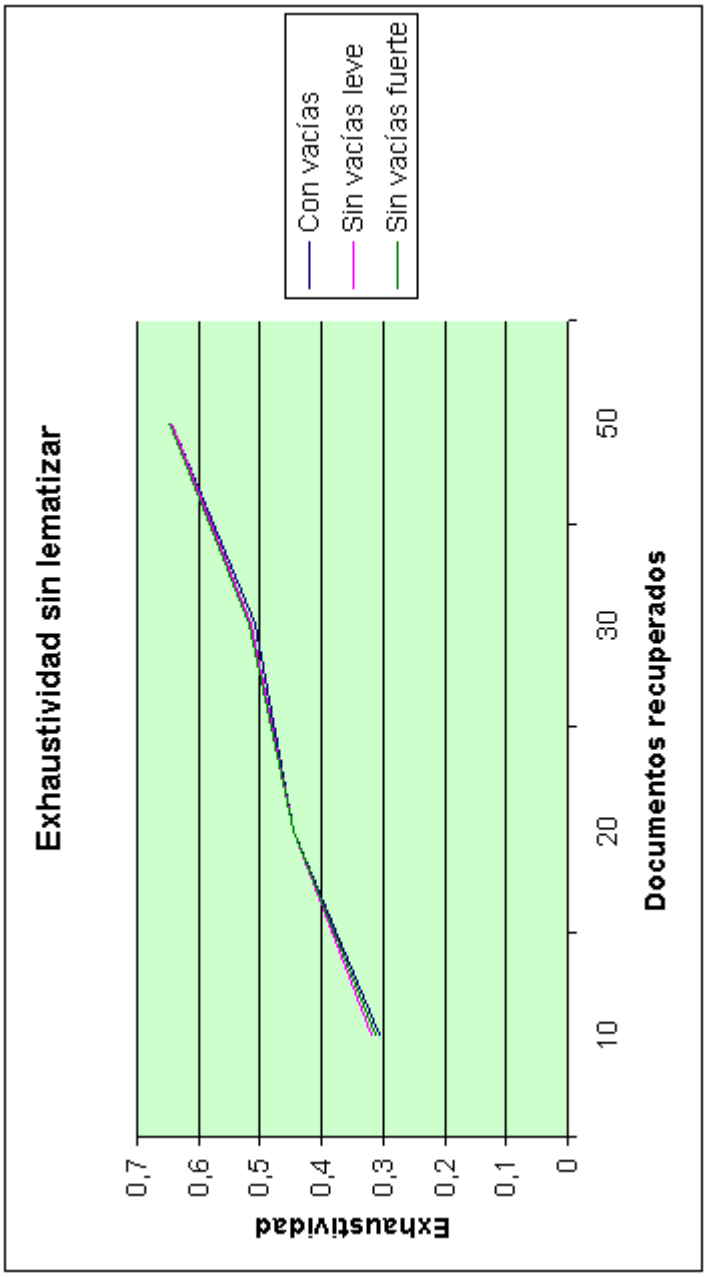


Gráfico 10 Exhaustividad *sin lematizar*

En lo que se refiere a los valores de exhaustividad, tanto para lematización derivativa como para la flexiva simplemente hemos calculado los datos suprimiendo las palabras vacías antes de la lematizar, puesto que si en la precisión no hay diferencias, entendimos que en la exhaustividad tampoco las habría.

13.3.2.2 Exhaustividad lematización derivativa

En la siguiente tabla se muestran los resultados de exhaustividad a los 10, 20, 30 y 50 primeros documentos recuperados, aplicando lematización derivativa, con palabras vacías, y sin ellas según las dos listas: *vacías leve* y *vacías fuerte.*

Documentos recuperados	**Derivativa sin vacías leve antes**	**Derivativa sin vacías fuerte antes**
10	0,309	0,3218
20	0,4635	0,4549
30	0,5321	0,545
50	0,6437	0,6609
100	0,8025	0,7982

Tabla 21 Exhaustividad *lematización derivativa*

Como vemos en la tabla anterior y se muestra en el gráfico siguiente, los resultados son muy parecidos, aunque ligeramente mejores para los valores iniciales y finales el experimento que suprime la lista de palabras vacías fuerte. En el caso de la palabras vacías leve, los resultados son ligeramente superiores en los valores centrales.

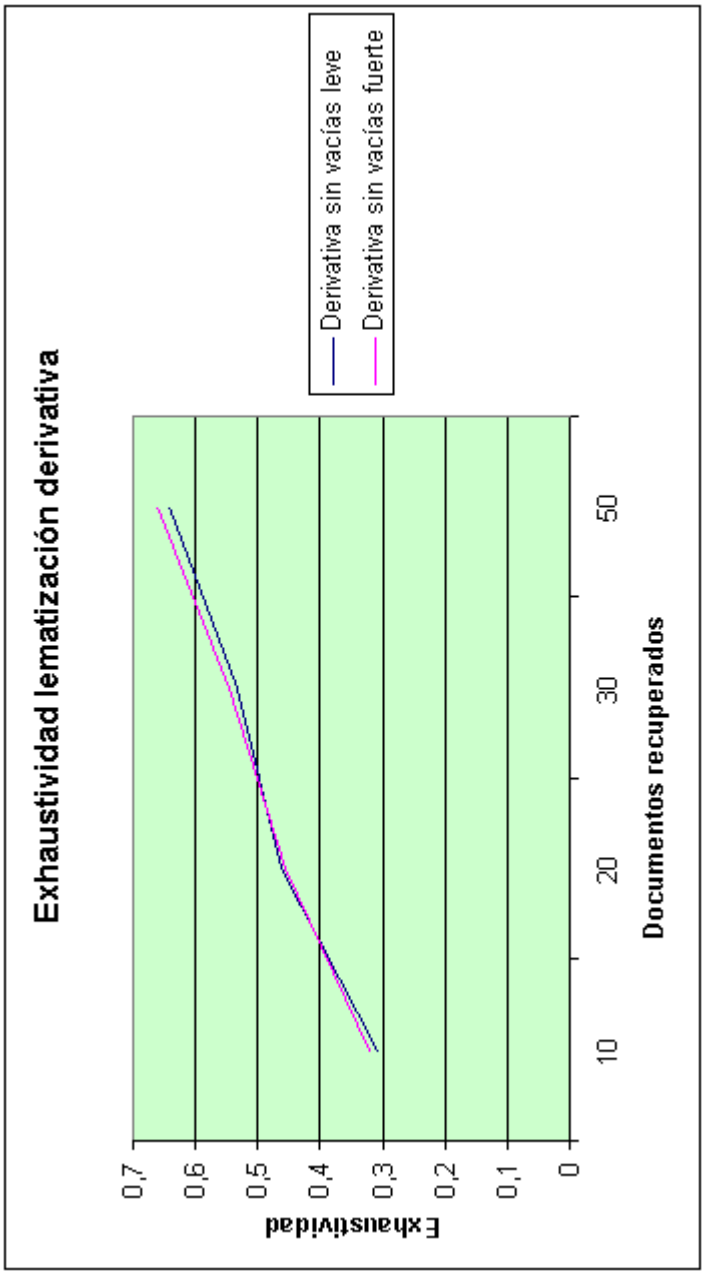


Gráfico 11 Exhaustividad *lematización derivativa*.

13.3.2.3.Exhaustividad lematización flexiva

En la siguiente tabla se muestran los resultados de exhaustividad a los 10, 20, 30 y 50 primeros documentos recuperados, aplicando lematización flexiva, con palabras vacías, y sin ellas según las dos listas: *vacías leve* y *vacías fuerte.*

Documentos recuperados	Flexiva sin vacías leve antes	Flexiva sin vacías fuerte antes
10	0,3476	0,3519
20	0,5021	0,5021
30	0,5793	0,5751
50	0,6909	0,6952
100	0,8154	0,8154

Tabla 22 Exhaustividad *lematización flexiva*.

En este caso como se muestra en la tabla y en gráfico siguiente los resultados prácticamente se solapan. Por lo que podemos afirmar que en la lematización flexiva la exhaustividad no se ve afectada por la supresión de una u otra listas de palabras vacías.

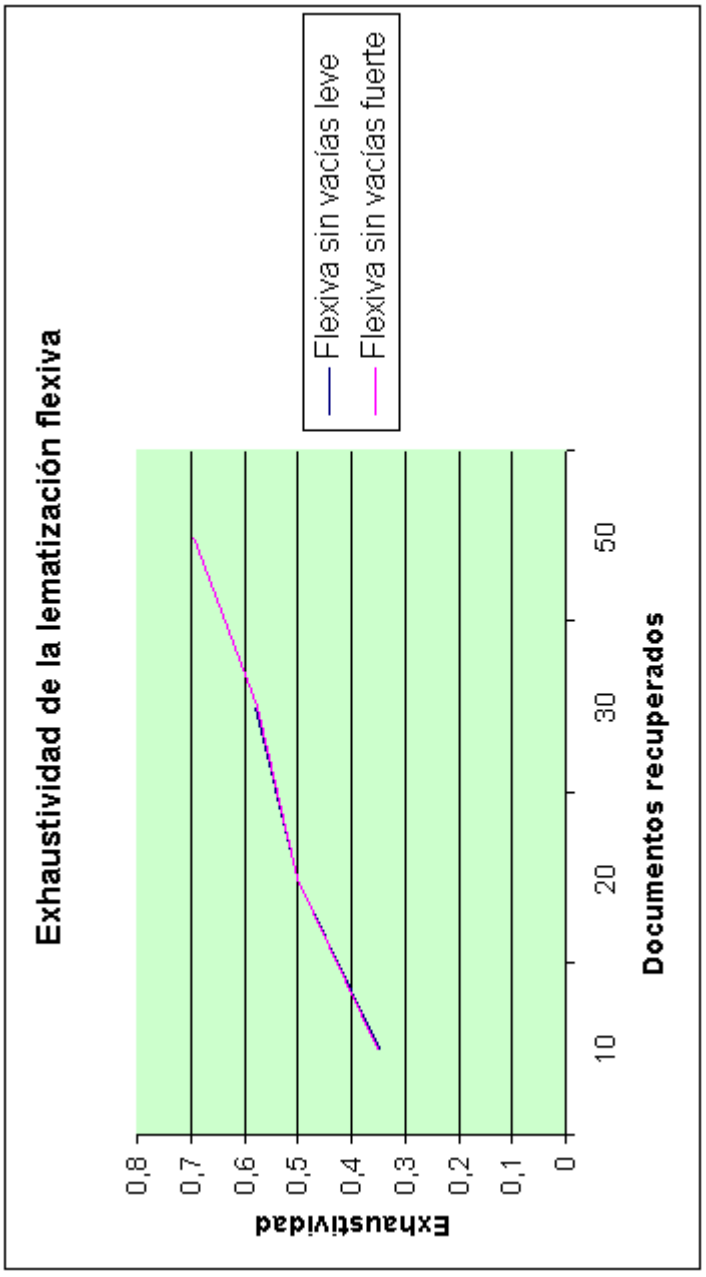


Gráfico 12 Exhaustividad *lematización flexiva*

Como en el caso anterior vamos a comparar el experimento con mejores resultados de cada grupo. Con el fin de determinar cual de ellos es el que tiene una exhaustividad mayor.

Como puede observarse en el gráfico siguiente, la exhaustividad de la lematización flexiva es superior a la lematización derivativa y a la no aplicación de la lematización. Entre la lematización derivativa y la no lematización, son mejores los resultados de la lematización derivativa con una diferencia mayor a partir de los 20 primeros documentos recuperados.

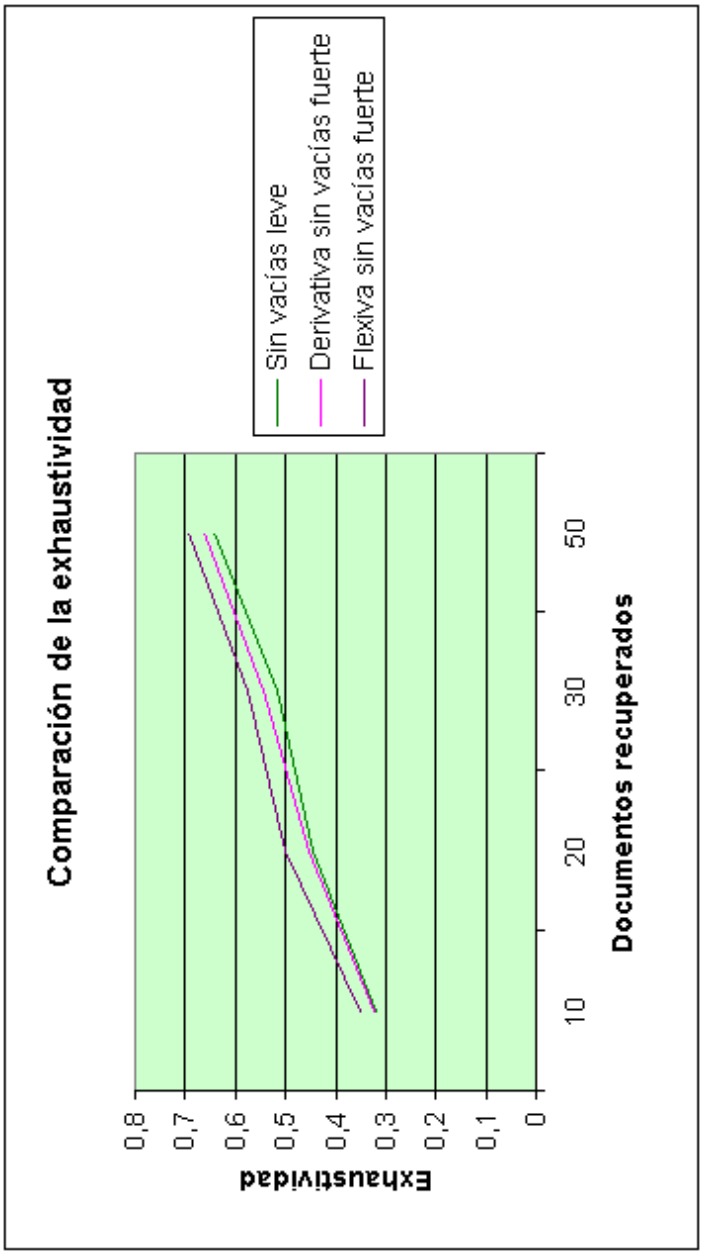


Gráfico 13 Comparación de la exhaustividad

13.3.3 Precisión-exhaustividad

Finalmente vamos a poner en relación los resultados de la precisión y la exhaustividad con el fin de mostrar como se relacionan estas dos medidas en cada experimento, ya que como vimos antes[326], en ocasiones una precisión alta puede hacer que la exhaustividad sea baja y es necesario ver que estas medias se compensan puesto que si un experimento tuviera una precisión alta y una exhaustividad muy baja, o viceversa, no podríamos considerar que los resultados son buenos.

Para poner en relación las dos mediadas, hemos calculado para cada valor de exhaustividad entre 0 y 1, el valor de precisión que le corresponde. Para su representación gráfica hemos marcado en el eje de las *x* la exhaustividad y en el de las *y* la precisión correspondiente.

Al igual que hemos ido haciendo con la precisión y la exhaustividad, hemos agrupado los experimentos en función del tipo de lematización y después compararemos los mejores resultados en un gráfico.

13.3.3.1 Precisión-exhaustividad sin lematizar

En la siguiente tabla vemos la precisión correspondiente a cada valor de exhaustividad para los experimentos que no aplican ningún tipo de lematización. Al igual que en los casos anteriores esto lo hemos combinado con la supresión de palabras vacías según las listas antes utilizadas y manteniendo las palabras vacías.

[326] Ver parte de medidas de evaluación.

EXHAUSTIVIDAD	Con vacías	Sin vacías leve	Sin vacías fuerte
0,1	0,6179	0,6	0,6462
0,2	0,547	0,564	0,6482
0,3	0,4925	0,4961	0,6133
0,4	0,3956	0,4	0,4615
0,5	0,2725	0,274	0,3816
0,6	0,2246	0,225	0,2636
0,7	0,1447	0,144	0,199

Tabla 23 Precisión-Exhaustividad *sin lematizar.*

En este primer gráfico se muestra como la supresión de palabras vacías indice positivamente en la recuperación de la información en los sistemas que no aplican lematización. En el gráfico vemos que son mejores resultados los que suprimen las vacías fuerte con una diferencia significativa respecto a los que no suprimen las palabras vacías y respecto a los que suprimen solamente las vacías leve.

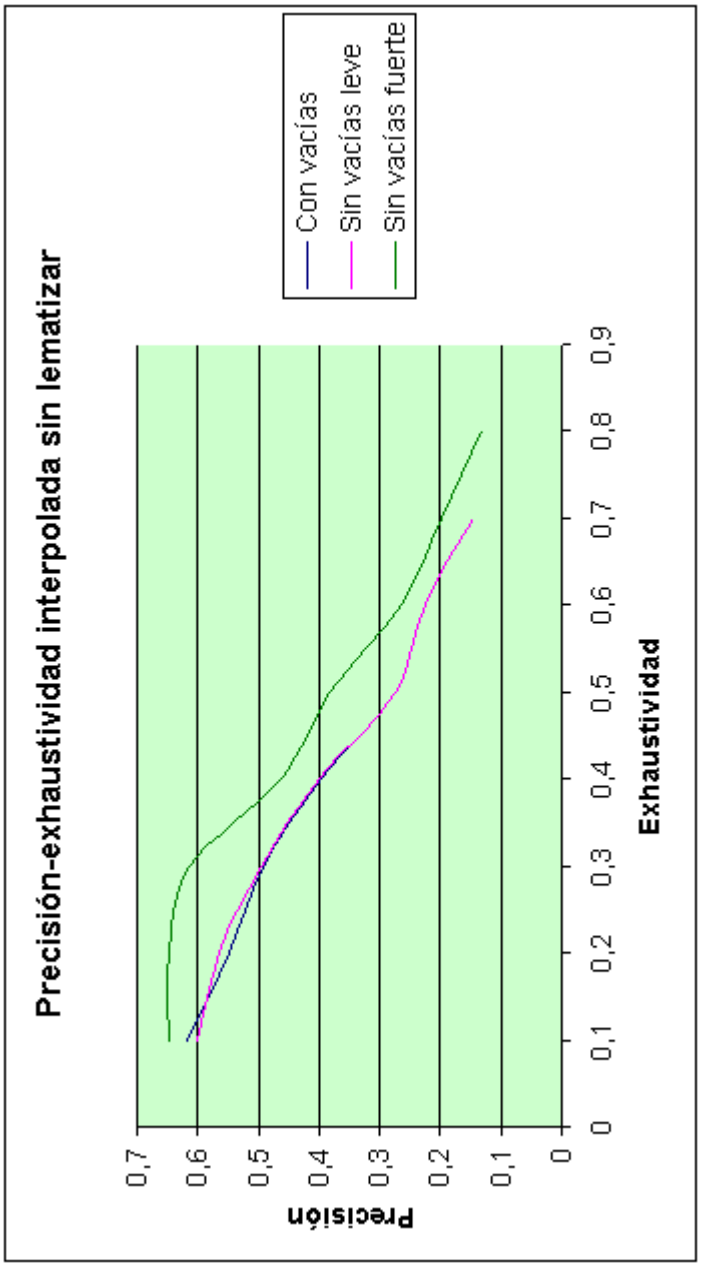


Gráfico 14 Precisión-Exhaustividad *sin lematizar*

13.3.3.2 Precisión-exhaustividad lematización derivativa.

En la siguiente tabla vemos la precisión correspondiente a cada valor de exhaustividad para los experimentos que aplican lematización derivativa, y suprimen las palabras vacías según la lista *vacías fuerte* y *vacías leve.*

EXHAUSTIVIDAD	Con vacías	Vacías leve después	Vacías fuerte
0,1	0,6275	0,6405	0,6179
0,2	0,5646	0,5632	0,574
0,3	0,4883	0,4931	0,4965
0,4	0,4277	0,4475	0,4515
0,5	0,3042	0,3042	0,305
0,6	0,2494	0,2494	0,2494
0,7	0,1824	0,1803	0,1824
0,8	0,1329	0,1274	0,1274

Tabla 24 Precisión-Exhaustividad *lematización derivativa*.

Como se muestra en la tabla y en el gráfico, en la lematización derivativa los resultados son muy parecidos, aunque ligeramente mejores los que suprimen las palabras vacías, de cualquiera de las dos listas, frente a los que no eliminan este tipo de palabras vacías.

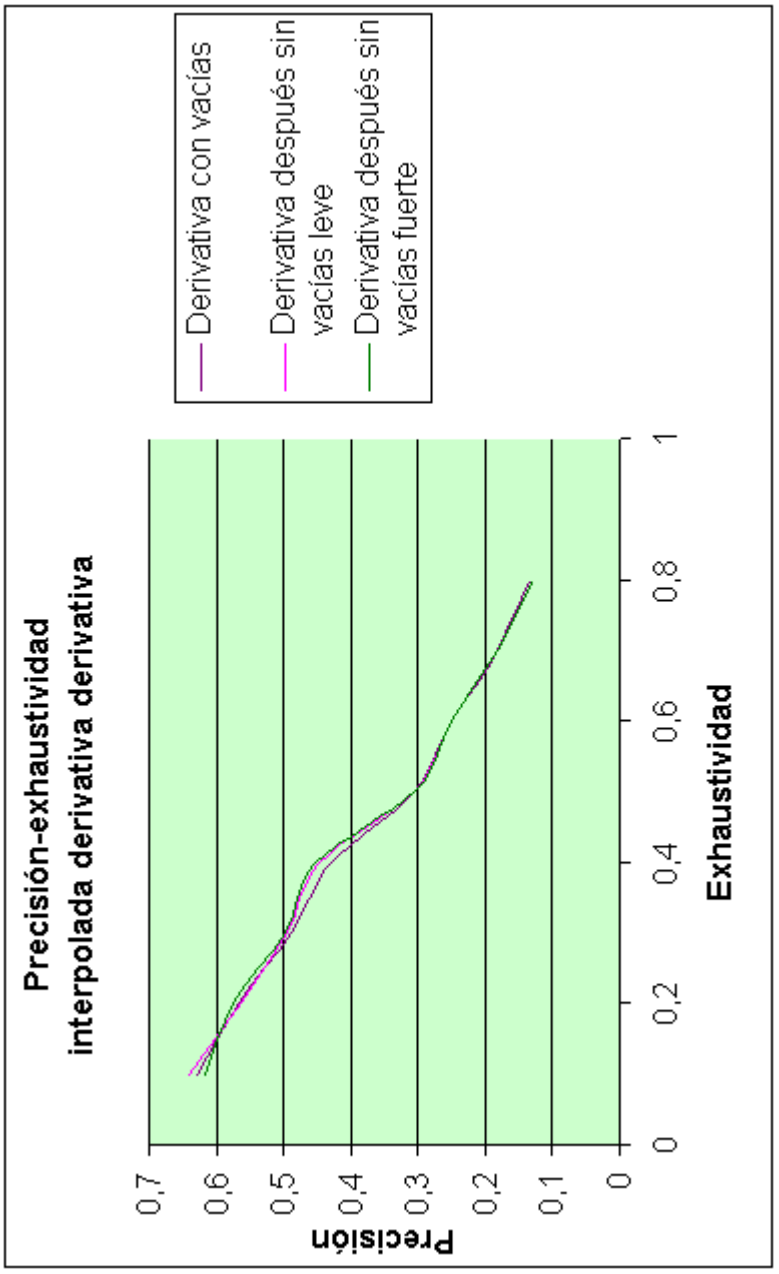


Gráfico 15 Precisión-exhaustividad *lematización derivativa*.

13.3.3.3 Precisión-exhaustividad lematización flexiva

En la siguiente tabla vemos la precisión correspondiente a cada valor de exhaustividad para los experimentos que aplican lematización flexiva, y suprimen las palabras vacías según la lista *vacías fuerte* y *vacías leve.*

EXHAUSTIVIDAD	Con vacías	Sin vacías leve	Sin vacías fuerte
0,1	0,6477	0,6462	0,6462
0,2	0,6216	0,6488	0,6482
0,3	0,6067	0,6133	0,6133
0,4	0,4884	0,4922	0,4615
0,5	0,3967	0,3984	0,3816
0,6	0,2624	0,2563	0,2636
0,7	0,2027	0,1994	0,199
0,8	0,1388	0,1334	0,1317

Tabla 25 Precisión-Exhaustividad *lematización flexiva*.

Como podemos apreciar tanto por los datos de la tabla como por el gráfico siguiente, en lo que se refiere a la exhaustividad, los mejores resultados son los que suprimen las palabras vacías. Respecto a las dos listas, apenas hay diferencias.

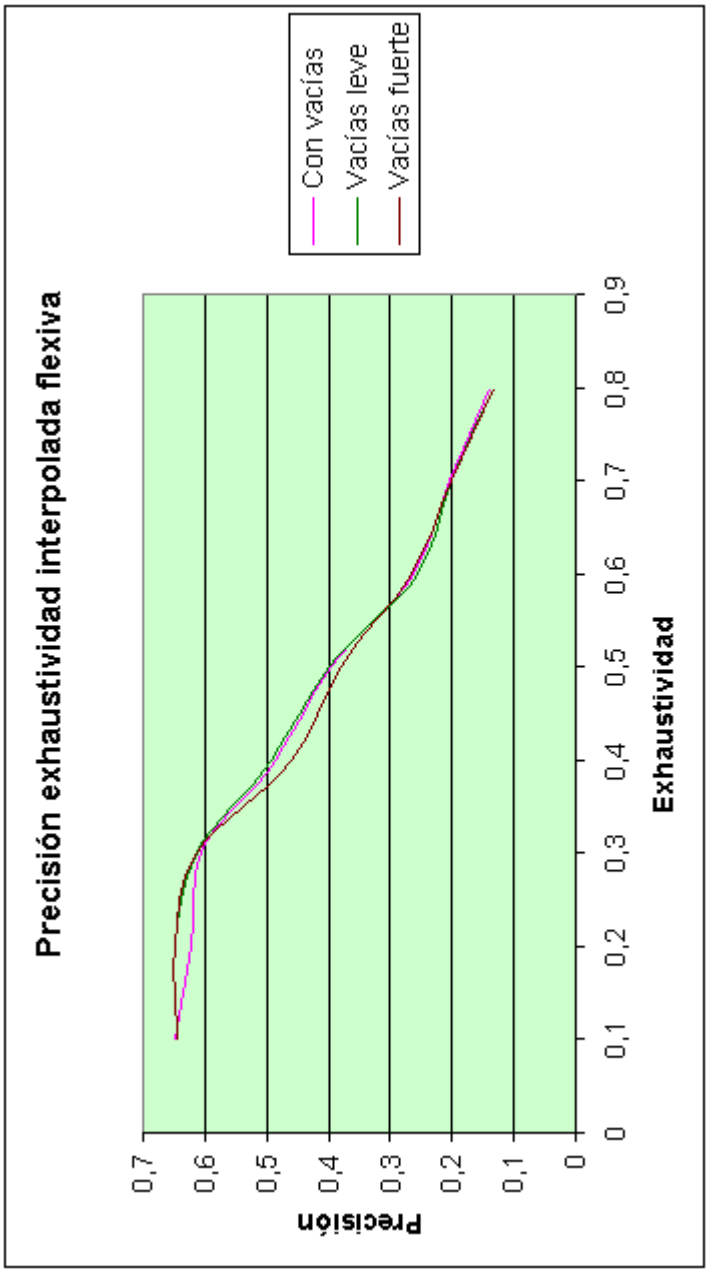


Gráfico 16 Precisión-exhaustividad *lematización flexiva*.

A continuación, tal y como hicimos con los experimentos anteriores hemos elegido los mejores resultados de cada grupo de experimentos, con el fin de compararlos para mostrar cual de todos los experimentos es el que obtiene mejores resultados. En los tres casos hemos elegido la supresión de la lista de vacías leve.

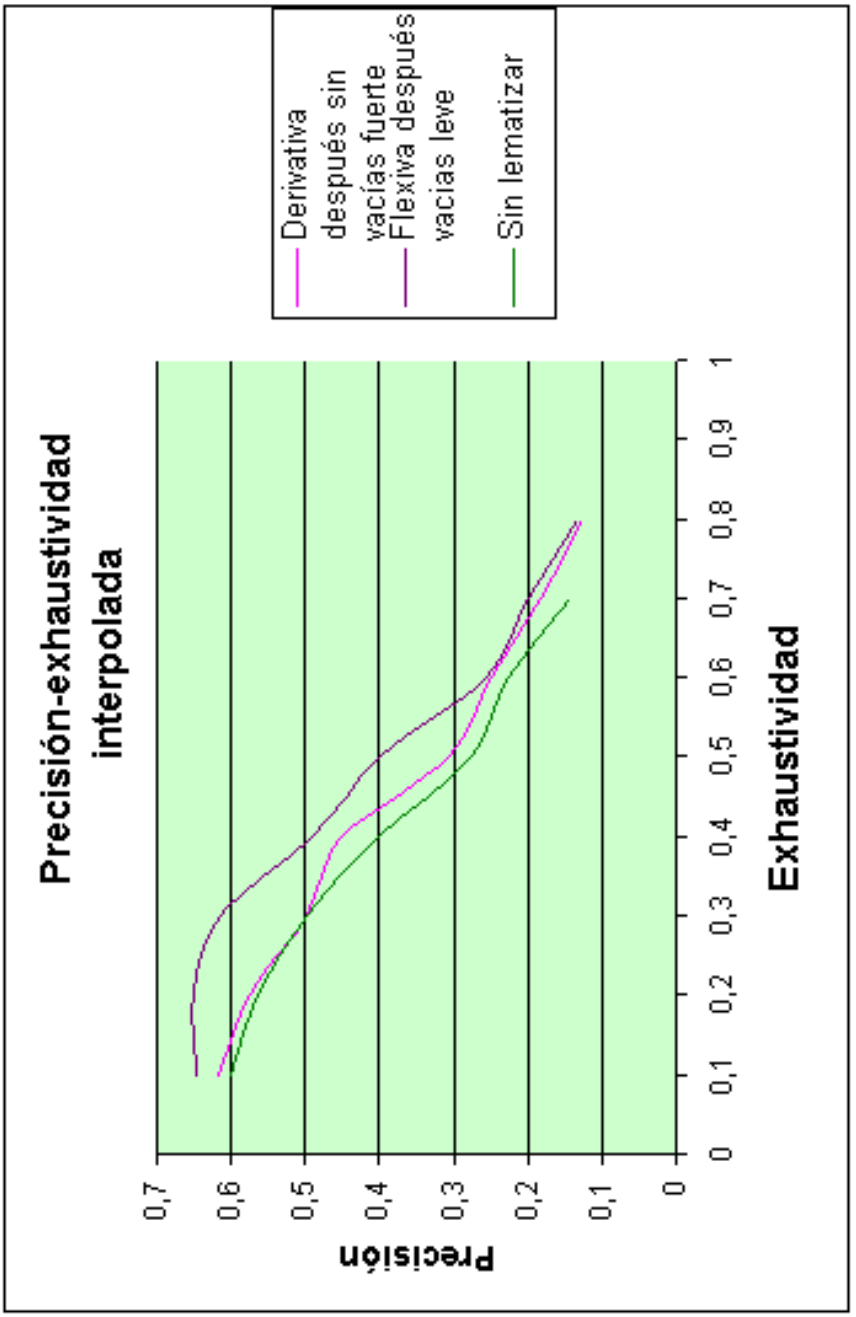


Gráfico 17 Comparación precisión exhaustividad

En este gráfico, se ve claramente como la lematización flexiva es la que obtiene una mejor relación de precisión-exhaustividad, alcanzando valores próximos a 0,65, le sigue el experimento de lematización derivativa, y finalmente

los peores resultados son los obtenidos por los experimentos que no aplican lematización.

En este gráfico también podemos ver cómo en el caso de la lematización flexiva sin vacías leve, en los primeros documentos se mantienen los valores de precisión a medida que la exhaustividad crce, en cambio con los otros dos experimentos, esto no se mantiene y a medida que la exhaustividad aumenta, la precisión decae.

14. Conclusiones

Una vez analizados los experimentos uno a uno, vistas las medias de los resultados y los gráficos correspondientes a cada experimento, hemos llegado a las siguientes conclusiones. Con el fin de que la exposición sea más clara, las hemos agrupado en tres grandes bloques:

- palabras vacías
- lematización derivativa
- lematización flexiva.

14.1 Palabras vacías

En cuanto a las palabras vacías, los resultados del Trabajo de Grado[327], ya mostraban que la supresión de palabras vacías, afectaba positivamente tanto a

[327] R. GÓMEZ (1998) op. cit.

la precisión como a la exhaustividad. En este sentido, los experimentos realizados aquí lo confirman, tanto si aplicamos lematización, sea del tipo que sea, como si no aplicamos ningún tipo de lematización[328].

Con estos experimentos hemos demostrado que la supresión de las palabras vacías mejora la precisión y la exhaustividad de la recuperación, tanto si consideramos estos valores de manera independiente como si relacionamos ambas medidas, tanto si aplicamos lematización como si no la aplicamos.

Respecto al tipo de lista que es más adecuado utilizar, si las que simplemente suprimen palabras vacías de contenido, por la categoría gramatical a la que pertenecen, o aquellas con frecuencia de aparición alta, en nuestros experimentos no hemos encontrado grandes diferencias, salvo que no se aplique ningún tipo de lematización, en este caso es preferible suprimir la lista de palabras vacías más amplia. Por lo tanto consideramos que es mejor tener en cuenta simplemente las palabras vacías según las categoría gramaticales, lo que en este trabajo hemos denominado *vacías leve*, ya que estas listas, son fácilmente adaptables a cualquier base de datos.

El hecho de que no haya grandes diferencias en los resultados entre el uso de la lista más corta y la más larga, puede deberse a que parte de estas palabras no se encuentren entre el texto. También hay que tener el cuenta, que aunque para hacer la lista se empleó un diccionario de frecuencias, esas palabras no son necesariamente las más frecuentes en la base de datos que empleamos.

En lo que se refiere al momento de la supresión de las palabras vacías, según los resultados de los experimentos, hemos visto, que apenas variaban, por lo que es preferible eliminarlas antes de lematizar, ya que de este modo el número de palabras a tener en cuenta se reduce, lo que hace reducir el espacio en disco, y

[328] Ver gráficos de precisión.

el tiempo que necesitamos invertir en el proceso de lematización y en el de indización. Y el único inconveniente que a priori puede haber al eliminarlas antes, es que hay que tener en cuenta todas formas de una misma palabra (femenino/masculino, singular/plural, tiempos verbales para ser, estar, haber), y si se hiciera después, esto no habría que tenerlo en cuenta en la confección de la lista.

14.2 Lematización derivativa

Respecto a los resultados en la lematización derivativa, los resultados esperados eran que estos experimentos tuvieran la tasa más baja de precisión, por detrás de la lematización flexiva y la no lematización, en cambio como hemos podido observar en el gráfico que compara los tres tipos de experimentos, vemos que los resultados de lematización derivativa están por encima de la no lematización.

En lo que se refiere a los resultados de exhaustividad, era de esperar que el orden de los experimentos era el inverso al de precisión, es decir, el más exhaustivo debería haber sido la lematización derivativa, por debajo, la lematización flexiva, y los peores resultados, los de la no lematización. En cambio, aunque sí están por encima de los obtenido con la no lematización, no lo están de la lematización flexiva, probablemente porque la lematización derivativa produce un factor alto de ruido documental, se recuperan muchos documentos pero no son los adecuados.

Al poner en relación los resultados de precisión y exhaustividad en este tipo de lematización, vemos como los resultados son mejores que la no lematización pero peores que la lematización flexiva.

14.3 Lematización flexiva

Con este tipo de lematización menos radical que la derivativa se esperaba obtener una tasa de precisión menor que sin lematizar, y mayor que aplicando la lematización derivativa y una tasa de exhaustividad menor que en la derivativa y mayor que con la no lematización. Pero como se ha podido observar por los resultados tanto en valores de precisión, como de exhaustividad en este tipo de lematización es la que obtiene las tasas mayores y al relacionarlos, vemos como la precisión y la exhaustividad están compensadas.

Después de todo este trabajo podemos concluir que para el español es más beneficiosa la lematización flexiva suprimiendo las palabras vacías, según la lista vacías leves, antes de proceder a la lematización.

15. Comparación de nuestro lematizador con otros, utilizados en otros idiomas

A continuación hemos comparado nuestros lematizadores con los algoritmos de lematización clásicos para el inglés. También hemos incluido otro cuadro con los algoritmos para idiomas distintos del inglés. En ambos caso habrá que tener en cuenta las diferencias entre los idiomas.

	LOVINS	SALTON	DAWSON	PÒRTER	KROVENTZ	ESPAÑOL	
						DER.	FLEXIVO
AÑO	1968	1968	1974	1980	1993	2001	
FINALES	260 finales	260 finales	55 conjuntos de finales	60 finales	3 finales	230	88
Nº REGLAS	34	No especifica	No especifica	No especifica	No especifica	3692	2700
PRINCIPIO	Suprime el sufijo más largo	No especifica	Suprime el sufijo más largo	Suprime plurales y después el sufijo más largo	De plural a singular, de participio y gerundio a infinitivo	Sufijo más largo	
LONGITUD MÍNIMA	Juega con diversos tamaños	2 caracteres	2 caracteres	Al menos un conjunto formado por una vocal y una consonante	No	Sí, no menor de 3 caracteres	
TIPO DE CONTEXTO	Sensible	No especifica	Sensible	Sensible: restricciones cuantitativas y cualitativas	No	Sensible	
PASOS QUE SIGUE	Dos: suprime primero los plurales y después el resto de las reglas.	No especifica	No especifica	Suprime plurales y después el sufijo más largo	NO	Comprueba que la entrada no esté en la tabla lemas o palabras y después aplica las reglas del sufijo más largo, según la lista de finales	
DICCIONARIO	No	No especifica	Sí	Sí	No	Sí	
ITERATIVO	SI			Sí	Sí	Sí	

Tabla 26 Comparación de los algoritmos para inglés y el español

	HOLANDÉS	ESLOVENO	FRANCÉS	MALAYO	ARABE	LATIN	GRIEGO	ESPAÑOL	
								DER	FLE
AÑO	1994	1992	1993/1999	1993	1992/1994/1995	1996	1995	2001	
BASADO	Porter (Frakes)	Porter	No especifica	No especifica	No especifica	Porter (Frakes)	SMART (Salton)	Porter	
TRATA	Prefijos Sufijos infijos	Sufijos	Prefijos Sufijos	Prefijos Sufijos Prefijos&Sufijos infijos	Sufijos	Sufijos Sufijos enditicos	Sufijos	Sufijos	
NÚMERO DE SUFIJOS	No especifica	5276	No especifica	No especifica	No especifica	90 de nombres y adjetivos 26 de verbos	5 flexivos	230 flex y der	88 flex
NÚMERO DE REGLAS	98	No especifica	35	121	No especifica	No especifica	No especifica	3692	2700
EVALUACIÓN	Según Paice. Evalúa el algoritmo pero no su aplicación a la R.I.	Salton. Evalúa la aplicación a la recuperación	Salton. Evalúa la aplicación a la recuperación	Según Paice. Evalúa el algoritmo pero no su aplicación a la R.I.	Salton. Evalúa la aplicación a la recuperación	No especifica	Salton. Evalúa la R.I.	Salton Evalua la aplicación a la recuperación	
TIPO DE SUFIJOS	No especifica	No especifica	Flexivos y derivativos	No especifica	No especifica	Flexivos, derivativos y enditicos	Flexivos y derivativos	Flexivos y derivativos	Flexivos
MODO DE ACTUAR	No especifica	Reglas de contexto sensible	1° elimina los sufijos flexivos 2° derivativos por las categorías gramaticales	Las reglas se aplican en orden alfabético.	No especifica	6 pasos	Primero los flexivos y después los derivativos	Aplica las reglas correspondientes al sufijo más largo	

Tabla 27 Comparación de los algoritmos para idiomas distintos del inglés, y el español.

16. Otras aplicaciones del lematizador

Aparte de la aplicación del lematizador a la recuperación de la información, consideramos que puede ser útil para otras investigaciones futuras. Aquí simplemente las indicaremos:

Recuperación multilingüe: aunque en este trabajo la aplicación ha sido diseñada pensando en la recuperación en español, en un futuro se podrá aplicar a recuperación multilingüe. En ese caso será necesario hacer un lematizador similar para otros idiomas o utilizar alguno de los ya existentes. Habrá que establecer las equivalencias de los lemas en los distintos idiomas y a partir de aquí hacer la recuperación.

Estudio del léxico en determinados ámbitos: gracias a la reducción de todas las formas variantes de una palabra a su lema será más fácil estudiar el léxico, ya que se reduce el número de formas a analizar.

IV REVISIÓN DE OBJETIVOS Y CONCLUSIONES.

Antes de terminar con el trabajo, vamos a ir mostrando en qué medida hemos sido capaces de ir cumpliendo los objetivos marcados en la introducción.

El primer objetivo, era ver cuál es el estado de la cuestión de la recuperación de la información: modelos más importantes, medidas de evaluación, experimentos más significativos hechos con el español. Para cumplir este objetivo hicimos un análisis de los distintos modelos según la clasificación de Belkin, y otras clasificaciones complementarias a la de este autor, así mismo hemos revisado las medidas de evaluación. A pesar de que indicábamos en la introducción que era difícil encontrar referentes para nuestro trabajo por la escasez de investigación de esta área, hemos analizado los experimentos que se realizaron en las TREC en los años 1994, 1995 y 1996 para el español. A pesar de que dichos trabajos apenas dan detalles de cómo han ido realizando los distintos experimentos han sido una fuente importante de estudio para nosotros, y nos han orientado en algunas cuestiones que más tarde tuvimos que aplicar.

El segundo objetivo pretendía mostrar si es eficaz un modelo de recuperación basado en información no estructurada en campos. Como se ha podido ver en la parte de los experimentos de lematización, la lematización permite hacer las búsquedas aunque la información esté en un único campo, el tamaño de estos campos no está limitado, por lo que podemos afirmar que sí es posible crear un modelo de recuperación eficaz basado en la información no estructurada en campos.

El tercero de los objetivos era hacer un estudio más detallado de la lematización y de los algoritmos para llevar a cabo la misma, tanto de los

elaborados para el inglés como los realizados para otros idiomas. Ver las distintas clasificaciones que hay al respecto. En el capítulo segundo, de este trabajo, se dedica a la revisión bibliográfica de los distintos algoritmos de lematización, se comparan primero entre sí y en el capítulo tercero con el que hemos realizado para el español. Respecto a las clasificaciones de los algoritmos, como novedad, aportamos una clasificación más, la basada en el conocimiento lingüístico que aplican dichos algoritmos.

El cuarto y el quinto objetivo están muy relacionados, por lo que los vamos a revisar de manera conjunta. El cuarto consistía en ver si es posible la creación para el español de un lematizador flexivo y otro derivativo mediante una máquina de estados finitos. El quinto, en ver si se podía aplicar a la recuperación de información y si ello producía mejoras en términos de precisión y exhaustividad en las búsquedas. También pretendíamos establecer qué tipo de lematización es más ventajosa para la recuperación de información en español, si la flexiva o la derivativa. Como se ha mostrado en el capítulo tercero, no solo es posible la creación de un algoritmo de lematización, que en función del conocimiento lingüístico que le añadamos será flexivo o derivativo, sino que su aplicación a la R.I. aporta mejoras en precisión y exhaustividad tanto por separado como al relacionar estas medidas respecto a los sistemas que no aplican ningún sistema de lematización. En lo que se refiere a cuál de los dos lematizadores[329] es mejor de cara a la recuperación, como vimos en el capítulo anterior, la lematización flexiva obtiene mejores rendimiento que la derivativa, al igual que ocurre con el S' Stemer de Harman y el K' Stemer de Kroventz para el inglés.

[329] Aunque hablamos de dos lematizadores técnicamente el algoritmo es el mismo, simplemente hay que variar el conocimiento lingüístico, pero por claridad en la exposición hablamos de lematizadores.

Finalmente, el último objetivo pretendía mostrar cómo incide la eliminación de palabras vacías en la recuperación, qué criterios se deben elegir a la hora de crear dichas listas. Mostrar si hay diferencias significativas entre los distintos tipos de listas. Respecto a esto, hemos corroborado en este trabajo el hecho de que la supresión de las palabras vacías incide positivamente en la R.I. en lo que se refiere al criterio de elaboración de las listas, salvo en el caso de no aplicar lematización, donde además de suprimir las categorías gramaticales vacías de contenido influye el quitar las palabras con alta frecuencia de aparición, consideramos que es mejor basarse simplemente en categorías gramaticales: preposiciones, conjunciones, artículos, pronombres, alguno verbos sin contenido (ser, haber, estar) pero no incluir palabras con alta frecuencia de aparición ya que eliminar en estas palabras puede ser adecuado para determinar bases de datos pero no en otras. Utilizar una lista no basada en frecuencias altas de aparición hace que la lista sea más fácil de aplicar en distintos contextos.

BIBLIOGRAFÍA

ABU-SALEM, H. ; AL-OMARI, M. ; EVENS, M. Stemming Methodologies Over Individual Queri Words for an Arabic Information Retrieval System *Journal of the American Society for Information Science* 1999 50 (6) p. 524-529.

AGUADO, P. M. los sistemas expertos y la recuperación documental: ejemplos de aplicación. *Scire* julio-diciembre 1995 1 (2) p. 21-32.

AHMAD, F.; YUSOFF, M.; SEMBOK, T. Experiments with a Stemming Algorithm for Malay Word *Journal of the American Society for Information Science* 1996 47 (12) p. 909-918.

ALAMEDA, J. R. CUETOS, F. *Diccionario de frecuencias de las unidades lingüísticas*. Vol I-II. Oviedo: Universidad, 1995.

ALARCOS LLORACH, E. *Gramática de la lengua española*. 7ª reimp. Madrid: Espasa Calpe, 1995

ALCINA FRANCH, J. BLEQUA, J. M. *Gramática española*. 8ª ed. Barcelona: Ariel, 1991

ALFONSECA, M. SANCHO, J. MARTÍNEZ ORGA, M. *Teoría de los lenguajes, gramáticas y autómatas*. Madrid: Ed. Universidad y Cultura, 1990.

ALFONSO MORO, J. *Verbos españoles.* Madrid: Difusión, 1989

ALLAN, J. BALLESTEROS, L. CALLAN, J. P CROFT, W. B. LU, Z. Recent Experiments with INQUERY 1995 [en línea] <http://trec.nist.gov/pubs/trec4/t4_proceedings.html> [Consultado el 24/07/00]

ALLAN, J. BALLESTEROS, L. CALLAN, J. P. CROFT, W. B. LU, Z. INQUERY at TREC-5 1996 [en línea] <http://trec.nist.gov/pubs/trec5/t5_proceedings.html> [Consultado el 24/07/00]

Almacenamiento y recuperación de la información textual [en línea] <http://protos.dis.ulpgc.es/docencia/seminarios/rit/analisis-lexico [consultado el 10-03-01]

ALMELA PÉREZ, R. *Procedimientos de formación de palabras en español.* Barcelona: Ariel, 1999.

AMAT I NOGUERA, N. *Documentación científica y nuevas tecnologías de la información*. Madrid: Pirámide, 1987.

APPEL, A. W. and JACOBSON, C. J. The world's fastes scrabble program communication of the ACM, 31 (5) p. 572-578.

ARENAS ALEGRÍA, L. *Efectividad y dinamismo en la recuperación documental mediante Análisis Cluster*. [Microforma] Tesis doctoral. Bilbao: Departamento de Publicaciones, Universidad de Deusto, 1992.

BARRY, C. L. User–defined relevance criteria: An exploratory sudy *Journal of the American Society for Information Science* 1994 45 (3) p. 149-159.

BELKIN, N. J., BRUCE, C. W. Retrieval Techniques *Annual of Information Science and Tecnology.* 1987, vol 22. p 109-145.

BELL, C. and JONES K. P.. Toward everyday languaje information retrieval system via minicomputer. J*ournal of the American Society for Information Science* 1979 30 p. 334-338.

BELTRÁN, C. Modelo informático de recuperación documental. [en línea]<http://www.ucm.es/info/multidoc/revista/cuadern5/ceseda.htm [consultado el 17/06/99]

BELY, N. [et al.] *Procedures d' analyse sémantique appliquées a la documentation scientifique*. Paris: Gauthier Villar, 1970.

BLAIR D. C. and MARON M. E. An evaluation of retrieval effectiveness for a full-text document retreival systems. *Comunication to ACM* March 1985 28 (3) p. 289-299.

BLAIR D. C. Searching bases in large interactive document retrieval systems *Journal of the American Society for Information Science* 1980 (31) 4 p. 271-277

BORLUM, P. and INGWERSEN, P. The development of method for evaluation of interactive Information Retrieval System. *Journal of Documentation* 1997 53 (3) p. 225-250

BOSQUE, I. DEMONTE, V. (dir) *Gramática descriptiva de la lengua española*. Madrid: Espasa 1999

BOSQUE, I. La morfología en ABAD, F. Y GARCÍA BERRIO, A. *Introducción a la lingüística*. Madrid: Alambra, 1983

BOYCE B. Beyond Topically: A two storage view of relevance and retrieval process *Information procesing and Management* 1992 18 p. 105-109.

BROGLIO, J., CALLAN, J. P, CROFT, W. B, NACHBAR, D. W.. Document Retrieval and Routing Using the INQUERY System 1994 [en línea] <http://trec.nist.gov/pubs/trec3/t3_proceedings.html> [Consultado el 24/07/00]

BROOOKSHEAR, J. G. *Teoría de la computación lenguajes formales, autómatas y complejidad*. España: Addison-Wesley Iberoamericana, 1993

BUCKLAND M. K. Relatedness, Relevance and Responsiveness in Retrieval Systems, Information *Processing and Management* 1983 19 p. 237-241.

BUCKLEY, C. SALTON, G. ALLAN, J. SINGHAL, A. Automatic Query Expansion Using SMART: TREC 3 1994 [en línea] <http://trec.nist.gov/pubs/trec3/t3_proceedings.html> [Consultado el 24/07/00]

BUCKLEY, C. SINGHAL, A. MITRA, M. New Retrieval Approaches Using SMART: TREC 4 1995 [en línea]<http://trec.nist.gov/pubs/trec4/t4_proceedings.html> [Consultado el 24/07/00]

BUCKLEY, C. SINGHAL, A. MITRA, M. Using Query Zoning and Correlation within SMART: TREC 5 1996 [en línea]<http://trec.nist.gov/pubs/trec5/t5_proceedings.html> [Consultado el 24/07/00]

BUSTOS GISBERT, J. M. *Definición de glosarios léxicos del español: niveles inicial e intermedio*. [en prensa]

CARRETERO, J. RODRIGUEZ, S. Building lexical tools to manage Information written in Spanish. *Journal of Information Science* 1996 22 (5) p. 391-399

CASADO VELARDE, M. Otros procedimientos morfológicos: acortamientos, formación de siglas y acrónimos EN BOSQUE, I. DEMONTE, V. *Gramática descriptiva de la lengua española*. Madrid: Espasa, 1999

CAVNAR, W. N-Gram-Based Text Filtering For TREC-2 In HARMAN, D. (Ed) Proceedings of TREC-2: Text Retrieval Pearce & Miller 25 Conference 2, Gaithersburg, MD, 1993. National Institute of Standards and Technology. [También en línea] http://trec.nist.gov/pubs/trec2/t2_proceedings.html [Consultado el 24/07/00].

CAVNAR, W. Using An N-Gram-Based Document Representation With A Vector Processing Retrieval Model 1994 [en línea] <http://trec.nist.gov/pubs/trec3/t3_proceedings.html> [Consultado el 24/07/00]

CHEN, H. and HOUSTON, A. L. Internet browsing and searching: User evaluations of category map and Concep space techniques. *Journal of the American Society for Information Science* 1988 49 (7) p. 582-603.

CHEN, H. and KIN GANNET, J. A machine learning approach to document retrieval. *Journal oj Management Information Systems* 1995, 11 p 7-41.

CHEN, H. Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms. *Journal of the American Society for Information Science* 1995, 46 (3). p 194-216.

CHOMSKY, N. *La nueva sintaxis: teoría de la rección y el ligamento.* Barcelona: Paidós, 1988.

CLEVERDON, C. W. Design and Evaluation of Information System. *Annual review of information Science and Tecnology*. 1971, n 6, p. 42-73

CLEVERDON, C. W. MILLS, J. KEEN, M. Factors determining the perfomance indexing systems. ASLIB Cranfield proyect, 1966

CODINA, L. Teoría de recuperación de información: modelos fundamentales y aplicaciones a la gestión documental. *Information World en español*. Octubre 1995, n 38 p. 18-22

Collins Spanish-English English -Spanish Diccionario español- inglés inglés-español. Colins Smith ed. Barcelon: Grijalbo, 1979.

CORDON, O. MOYA, F. DE, ZARCO, M. C. Breve estudio sobre la aplicación de algoritmos genéticos a la recuperación de Información En LÓPEZ HUERTAS, M. J. Y FERNÁNDEZ MOLINA, J. C. *La representación y la organización del conocimientos en sus distintas perspectiva, su influencia en la Recuperación de la Información: Actas del IV Congreso ISKO-España EOCONSID'99 22-24 de abril de 1999*. Granada, 1999 p. 179-185

COOPER, S. The Paradoxical role of unesamied documes in the evaluation of retrieval effectiveness. Information Processing and Management 1976 12 p. 367-375

COOPER, W. S Expected Search Lenght: A single neasure of retrieval effectiveness based on the weak ordering action retrieval systems. *American Documentation* 1968 19 p. 30-41.

COOPER, W. S. On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science* 1973 24 p. 87-100

CROFT, W. B., JINXI, X. Corpus-Specific Stemming using word From Co-ocurrence [en línea] http://cobar.cs.umass.edu.info/psfiles/irpubs [Consultado el 16-6-2000]

CRYSTAL, D. *Enciclopedia del lenguaje de la Universidad de Cambridge*. Madrid: Taurus, 1994

CUADRA, A. C. and KATTER, R. V. Opening the black box of "relevance". *Journal of documentation* 1967 23 (4) p 291-303.

DARMÓSTENES, A. *Traité de la formation des mots composés*. Paris: Ed. Bouillón, 1974

Datathèke: http//milano.usal.es/dtt.hml

DAVIS, M. DUNNING, T. A TREC Evaluation of Query Translation Methods For Multi-Lingual Text Retrieval 1995 [en línea] <http://trec.nist.gov/pubs/trec4/t4_proceedings.html> [Consultado el 24/07/00]

DAVIS, M. New Experiments In Cross-Language Text Retrieval At NMSU's Computing Research Lab 1996 [en línea] <http://trec.nist.gov/pubs/trec5/t5_proceedings.html> [Consultado el 24/07/00]

DAWSON, J. Suffix Removal and Word Conflation *Assoc. Liter and linguistic Computing Bulletin*, 1974 p. 33-46

DELGADO LÓPEZ CÓZAR, E. Diagnóstico de la investigación en Biblioteconomía y Documentación en España (1976-1996): Estado embrionario. EN *Journal of Spanish Research on Information Science/Revista de Investigación Iberoamericana en Ciencia de la Información y Documenta*ción. Vol 1 (1) 2000 p. 79-93

Diccionario de la lengua española 21ª ed. en CD-Rom Madrid: R.A.E., 1992

DRISCOLL, J. D. Multi-lingual Text Filtering Using Semantic Modeling, (Praxis Technologies), S. Abbott, K. Hu, M. Miller, G. Thesis (University of Central Florida)

El Mundo. Primer semestre. Textos íntegros. Mundired. Servicio electrónico, [CD] 1994 El mundo ISBN 84-920059-1-2

EL-HAMDOUCHI, WILLETT, A P. Comparison of hierachic agglomerative clustering methods of document retrieval. *The Computer Journal*. 1989 32 (3) p. 220-227.

ELLIS, D. *New Horizons in Information Retrieval*. London: Library Association, 1990.

FIGUEROLA, C. G. GÓMEZ, R. LOPEZ DE SAN ROMÁN, E. Stemming and n-grams in Spanish: an evaluation of their impact on information retrieval. *Journal of Information Science* 2000 26 (6) p. 461-467

FRAKES, W. B. Stemming Algorithms EN FRAKES, W. B. and BAEZA-YATES (ed) *Information Retrieval : Data Structures and Algorithms.* México: Prencite-Hall Hispanoamericana, 1992. p. 131-161 (b)

FRAKES, W. B. Term Conflation for information retrieval C.J. van Rijsbergen (ed.). *Research and development in information retrieval*. Cambridge: C.U.P., 1984 p. 383-390

FRAKES, W.B. , BAEZA YATES, R. (ed) *Information Retrieval: Data Structures and Algorithms.* Mexico: Prentice-Hall, 1992.

FUENTE REDONDO, P. de la. Bibliotecas digitales [Conferencia pronunciada en Valladolid el 16 de marzo de 1998] EN *Nuevas tendencias en gestión de la información*. Valladolid 12 al 18 de marzo de 1998]

GEY, F. C. CHEN, A. HE, J. XU, L. and MEGGS, J. Term importance boolean conjunct training, negative term, and foreing language retrieval: probabilistic algorithm at TREC-5 1996 [en línea] <http://trec.nist.gov/pubs/trec4/t4_proceedings.html> [Consultado el 24/07/00]

GEY, F. C. CHEN, J. A. HE, M. and JASON Logistic Regression at TREC4: Probabilistic Retrieval from Full Text Document Collections 1995 [en línea] <http://trec.nist.gov/pubs/trec4/t4_proceedings.html> [Consultado el 24/07/00]

GIL LEIVA, I. RODRÍGUEZ MUÑOZ, J. V. El procesamiento del lenguaje natural aplicado al análisis de contenido de los documentos. *Revista General de Información y Documentación*. 1996. 6 (2) p. 205-218.

GOFFMAN and NEWILL *Methodology for test and evaluation of information retrieval systems*. Information Storage and Retrieval 3 p 19-25

GÓMEZ DÍAZ, R. *La Recuperación de la Información en español: evaluación del efecto de sus peculiaridades lingüísticas*. Universidad de Salamanca. Trabajo de Grado, 1998. [trabajo no publicado].

GONZALEZ COLLAR, A.L., GOÑI MENOYO, J. M. GONZÁLEZ CRISTOBAL, J. C. Un analizador morfológico basado en chart [consultado en línea] http://www.mat.upm.es/~arias/paper.html [consultado el 5-03-99]

GOÑI MENOYO, J. M. GONZÁLEZ J. C. A Framework for lexical representation [consultado en línea] http://www.mat.upm.es/~arias/paper.html [consultado el 5-03-99]

GREEN, B. F. [et al.] Baseball: An automatic question answerer. En FEIGENBAUM, E. A. and FELDMAN, J. (eds) *Computer and Thought*, 1963 p. 207-216.

GROSSMAN, D. A. HOLMES, D. O. FRIEDER, O. NGUYEN, M. D. and KINGSBURY, C. E. Improving Accuracy and Run-Time Performance for TREC-

4 1995 [en línea] http://trec.nist.gov/pubs/trec4/t4_proceedings.html [Consultado el 24/07/00]

GROSSMAN, D. A. REICHART, J. CHOWDHURY, A. LUNDQUIST, C. HOLMES, D. FRIEDER, O. Using Relevance Feedback within the Relational Model for TREC-5 1996 [en línea] http://trec.nist.gov/pubs/trec5/t5_proceedings.html [Consultado el 24/07/00]

GROSSMAN, D. A. Using Relevance Feedback within the Retrieval Model for Trec-5 1996 [en línea] <http://trec.nist.gov/pubs/trec5/t5_proceedings.html> [Consultado el 24/07/00]

GRUMBACH S. and TAHI F. A new challenge for compression algorithms: genetic sequences. *Journal Information Processing and Management* 1994 30 (6) p. 875-886.

GUERRIE, B. Online Information System: Use and operating Characteristics, Limitations and Desing Alternatives. *Information Resources Pres*, 1983.

GUTIERREZ CUADRADO, J. PASCUAL, J. A. De cómo el castellano se convirtió en español En GARCÍA SIMÓN (ed) *Historia de una cultura. La singularidad de Castilla.* Valladolid: Junta de Castilla y León, 1995 p. 319-368.

HAFER, M., and WEIS, S. Word segmentation by letter succesor varities *Information Storage and Retrieval*, 1974 10 p. 371-385.

HALVORSEN, P. K. Overview EN ZAENEN, A (ed.) Document Processing EN *Survey of the State of the Art in Human Language technology.* Oregon: National Science Foundation, 1995 pp 255-258.

HARMAN, D. How Effective is Suffixing*?* *Journal of the American Society for Information Science* 1991 42 (1) p. 7-15.

HARMAN, D. Overview of the Fourth Text Retrieval Conference (TREC-4) 1995 [en línea] <http://trec.nist.gov/pubs/trec4/t4_proceedings.html>[Consultado el 24/07/00]

HARMAN, D. Overview of the Third Text Retrieval Conference (TREC-3) EN Pro*ceedings of the 3rd Text Retrieval Conference* (TREC-3), 1995 p. 1-19 [También en línea] <trec.nist.gov/pubs/trec3/t3_proceedings.html > [consultado el 10-01-01].

HARMAN, D. Ranking Algorithms EN FRAKES, W. B. and BAEZA-YATES (ed) *Information Retrieval : Data Structures and Algorithms.* México: Prencite-Hall Hispanoamericana, 1992. p. 363-392

HARMAN, D. Relevance Feedback and Other Queri Modifications Techniques EN FRAKES, W. B. and BAEZA-YATES (ed) Information Retrieval : Data Structures and Algorithms. México: Prencite-Hall Hispanoamericana, 1992 p. 241-263.

HARMAN, D. SCHAÜBLE, P. and SMEATON, A. Document Retrieval. EN *Survey of the State of the Art in Human Language Technology*. Oregon: National Science Foundation, 1995 p. 259-262

HARTER, S. P. *Online Information Retrieval: concepts, principles and Techniques*. San Diego: Academic Press, 1986.

HARTER, S. P. Variations in Relevance Assessment and Measurement of Retrieval Effectiveness. *Journal of the American Society for Information Science* 1996 47 (1) p. 37-49.

HARTER, S. P., HERT, C. A. Evaluation of Information Retrieval Systems: Approaches, Iusues, and Methods. *Annual Review of Information Science and Technology (ARIST)* vol 32, 1997 p. 3-94.

HARTER, S.P. Psycological Relevance and Inmortation Science. *Journal of the American Society for Information Science* 1992 43 (9) p. 602-615.

HAYWOOD, J. A. *Nueva gramática árabe: claves y ejercicios.* Madrid: Coloquio, 1993.

HEARST, M., PEDERSEN, J., PIROLLI, P., SCHUTZE, H., GREFENSTETTE, G. HULL, D. A.. Xerox Site Report: Four TREC-4 Tracks 1995 [en línea] <http://trec.nist.gov/pubs/trec4/t4_proceedings.html> [Consultado el 24/07/00]

HERT, C. A. Understanding information retrieval interaction: theoretical and practical implications. Greenwich: Ablex Publishing Corporation, 1997.

http://protos.dis.ulpgc.es/docencia/seminarios/rit/index.htm [consultado el 9-03-01]

http://www.research.att.com/sw/tools/fsm/ref.html [consultado 10-03-01]

HUFFMAN, S. Acquaintance: Language-Independent Document Categorization by N-Grams 1995 [en línea] <http://trec.nist.gov/pubs/trec4/t4_proceedings.html> [Consultado el 24/07/00]

HULL, D. A. GREFENSTETTE, G. HEARST, M. SCHUTZE, H . PEDERSEN, J. PIROLLI, P. Xerox TREC-5 Site Report: Routing, Filtering, NLP and Spanish Tracks 1995 [en línea] <http://trec.nist.gov/pubs/trec5/t5_proceedings.html> [Consultado el 24/07/00]

HULL, D. A. Stemmings Algorithms: A Case Study for Detailed Evaluation *Journal of the American Society for Information Science* 1996 47 (1) p. 70-84.

INGWERSEN, P. *Information Retrieval interaction*. London: Taylor Graham, 1992.

JACOB, P. Text Interpretation: Extracting Information En *Survey of the State of the Art in Human Language Technology*. Oregon: National Science Foundation, 1995 p 263-265.

JACQUEMIN C. and TZOUKERMAN E.. NLP for term variant extraction: synergy between morphology, lexicon, and syntax. EN STZALKOWSKI, T. (ed) *Natural Languaje Information Retrieval*. Dordrcht: Klwer Academic Publisher, 1999 p. 25-74

JAIN, A. K , DUBES, R. C *Algorithms for Clustering Data.* New Jersey: Prentice Hall, 1988.

JAKOPIN, F. ¿ Quieres hablar esloveno?. Ljubljana: Slovenska izslejeniska matica, 1962

JARDINE, N., RIJSBERGEN, K. V. The Use of hierarchic clustering in information retrieval*. Information Storage and Retrieval,* 1971, 7 p. 217-240.

JONHSON, A and FOTOUHI. Adaptative indexing in very large databases. *Journal of database Management* 1995 6 (6) p 4-12

KALAMBOUKIS, T. Z. Suffix stripping with moderm Greek. *Program* 1995 29 (3) p. 313-321.

KAPLAN, R. M. Finite State Technology. EN Mathematical Method En *Survey of the State of the Art in Human Language Technology*. Oregon: National Science Fundation, 1995 p. 419-422

KARLSSON, F. KARTTUNEN, L. Subsentian processing. En Survey of the State of the Art in Human Language Techology p. 111-112

KAY, M. and. MARTINS, G.R. The MIND System: the Morpholocical. Analysis. *Program*, US Air Force Proyect Rand Report R.M. 6265/2 PR April (1970).

KEEN, E. M. *Evaluation parameters*. Report ISR-13 to the Natinal Science Foundation, Section IIK, Cornell Univrsity Departament of Computer Science, 1967

KEEN, E. M. Evaluation parameters. En SALTON, G. (ed), *The SMART retrieval system Experimentes in automatic document processing*. Englewood Cliffs, New York: Prentice-Hall.1971 p. 74-111

KEEN, E. M. Measures and Averaging Methods Used in Performance Testing Indexing System. Crandfield, Eng.,Aslib Crandfield Project 1966.

KELLEDY, F and SMEATON A. F. TREC-5 Experiments at Dublin City University: Query Space Reduction Spanish and Character Shaape Encoding [en línea]<http://trec.nist.gov/pubs/trec5/t5_proceedings.html> [Consultado el 24/07/00]

KENT, A. [et al.] Machine literature searching. VIII. Operational Criteria for Designing Information Retrieval Systems American Documentation 6 (2) Abril 1955 p. 93-101

KORFHAGE, R. R. *Information Storage and Retrieval*. New York: John Wiley and Sons, 1997.

KOWALSKI, G. *Information Retrieval Sytstems: Theory and Implementarion*. 2nd prin. Boston: Kluwer Academic Publisher, 1997.

KRAAIJ, W. and POHLMANN, R. Evaluation of Dutch stemming algorithm [on-line] <http:// rayela.ieec.uned.es/~ircourse/doc/uplift/ > [consultado el 25-11-99].

KRAAIJ, W. and POLMANN, R. Viewing Stemming as Recall Enhacement [en línea] <http:// rayuela.ieec.uned.es/~ircourse/doc/uplift/sigir96revised.ps> [consultado el 25-11-99]. Edición revisada de la de SIGIR'96.

KRAAIJ, W., and POLMANN, R. *Porter´s Steming algorithm for Dutch*. In L. Noordman and W. De Uroomen (eds) Informatiewetenschap 1994: Wertenschapplijke bijdragen aan de derde STINFON Conferentie p. 167-180 43 [También en línea] <http://rayuela.ieec.uned.es/~ircourse/doc/uplift/> [Consultado el 25-11-1999]

KROVENTZ Viewing morphology as inference process. *Proceedings of the 16 th ACM/SIGIR* Conference. New York: Association for Computing Machinery 1993 p. 191-202. [en línea] ftp://ftp.cs.umass.edu/pub/techrept/techreport/1993/UM-CS-1993-036.ps [Consultado el 20-12-1999].

KULKARNI, J. and PARSII, H.R. Information resource matrix for production and intelligent manufacturing using genetic algorithms techniques *Computer and Industrial Engeniering* 1992 23 p 483-485

La loi de Zipf. Linguistique et Statistique de *L' Enciclopedie Universalis Version 3.0 Sur CD-ROM* [en línea] <http://users.info.unicaen.fr./~guguet/java/zipfeu.html [consultado el 13-01-01]

LANCASTER W. F. and WARNER, A. J. *Information Retrieval Today.* Arlington: Information Resources Press, 1993.

LANCASTER, F. W. *Information Retrieval Systems: Characteristic, Testing and Evaluation*, 2nd ed. New York: Wiley, 1979.

LANCASTER, F. W. *Vocabulary Control for Information Retrieval*. 2nd ed. Arlington: Information Resource Press, 1986.

LANG, M. F. *Formación de palabras en español: morfología derivativa productiva en el léxico moderno.* Madrid: Cátedra, 1992

LAPESA, R. *Historia de la lengua española*. 9ª ed. Cor y aum. Madrid: Gredos, 1988

LARGE, A. TEDD, L. A. HARTLEY, R. J. Information Seeking in the Online Age: Principles and Practice. London: Bowker, 1998

LÁZARO CARRETER, F *Diccionario de términos filológicos* 3ª ed. Madrid: Gredos, 1987

LÁZARO MORA, F. A. Sobre la parasíntesis en español *DICENDA Cuadernos de filología hispánica,* 5 Madrid: Ed. de la Universidad Complutense, 1986 p. 221-225

LELU, A. From data analysis to neural networks: new propects for efficient browsing through databases. *Journal of Documentation Science*, 1991 17 p, 1-2

LENON M., [et al.] An evaluation of same conflation algorithms for Information Retrieval *Journal of Information Science* 1981 3 p.177-188.

LESPINASSE, K. TREC: une conférence pour l´évaluation des systèmes de recherche d´information. *Documentaliste Sciences de l´information*. 1997, 34 (2) p. 74-81

LI, W. Zipf Law and the Structure and Evolutioon of Languages" (letter to Editors) *Complexiti* 1997, 2 (5) 12-13. [en línea] <http://linkage.rokfeller/edu/wli/pub/comp98_zipf.html> [Consultado el 13-01-01]

LIDDY, E. D. Enhanced Text Retrieval Using Natural Language Processing. [en línea] http://www.asis.org/Bulletin/Apr-98/liddy.html [consultado el 19/09/00]

LÓPEZ HUERTAS, M. J. Y FERNÁNDEZ MOLINA, J. C. (ed.) *La representación y la organización del conocimientos en sus distintas perspectiva, su influencia en la Recuperación de la Información: Actas del IV Congreso ISKO-España EOCONSID'99 22-24 de abril de 1999*. Granada, 1999.

LOSEE, R. M. Comparig Boolean and Probalistic Information Retrieval Systems across Queries and Disciplines. *Journal of The American Society for Information Science* 1997 48 (2) p. 143-156.

LOSEE, R. M. Learning syntactic rules and tags with genetic algorithms for information retrieval and filering: an empirical basis for gramatical rules. *Infomation processing and Management* 1996 32(2) p. 185-197.

LOVINS, J. B. Development of a Stemming Algorithm *Mechanichal translations and Computational Linguistics*. 1968 11 (1-2) p. 22-31

LUCEHESI C. L. and KWALTOWSKI, T. Application of the finite automata representing large vocabularioes. *Software-Practice and Experince*, 23 (1) p. 15-30

LUNH,.H. P. A Statistical approach to mechanized encoding and searching of literary information *IBM Journal of Research and Development* 1957, 1 (4) p. 309-313

Máquina de estado finito. En *Lenguajes formales* [en línea] <http://www.inf.UDEC.CL/~lenform/02.htm> [Consultado el 10-3-01]

MARCHIONINI, G. and SHNEIDERMAN, B. Finding fatcs vs. Browsing Knowledge in hypertext systems. *IEEE Computer*, 1988 21 (3) p. 70-79

MARONS, M.E. and KUHNS, J. L. On Relevance, Probabilistic Indexing and Information Retrieval. *Journal of the ACM*, 7 (3) p. 216-244

MARTÍN VEGA, A.. Las redes de neuronas artificiales en la recuperación de información: Algunas fuentes para su estudio. EN *Los profesionales ante el reto del siglo XXI: integración y calidad. IV Jornadas españolas de documentación automatizada. Documat 94* (Gijón, 6,7,y 8 de octubre de 1994) Actas. Oviedo: Universidad, 1994. p 403-410.

MENÉNDEZ PIDAL, R. *Los orígenes del español*. 4ª ed. Madrid: Espasa-Calpe, 1956

MENÉNDEZ PIDAL, R. *Manual de gramática histórica española*. 20ª ed. Madrid: Espasa-Calpe, 1989

MIRANDA, J. A. *La formación de palabras en español*. Salamanca: Ed. del Colegio de España, 1994.

MOHRI, M. On Some Applications for Finite State Automata Theory to Natural Language Processing Natural Engieering 1996 2 p. 1-20 [También en línea] http://www.search.att.com/sw/tools/fsm/jnle.ps [Consultado 20-01-01]

MOLINER, Mª. *Diccionario del uso del español*. Ed en CD-Rom. Madrid: Gredos, 1996

MONGE, F. Aspectos de la sufijación en español. *Revista española de lingüística*. 1996 26 (1) p 46-56

MORENO BORONAT, L. [et al.] *Introducción al procesamiento del lenguaje natural*. Alicante: Universidad de Alicante, 1999

MORENO, A. GOÑI MENOYO, J. M. GRAMPAL: A Morphological Processor for Spanish implemented in Prolog [consultado en línea] http://www.mat.upm.es/~arias/paper.html [consultado el 5-03-99]

MOYA ANEGÓN, Félix. La investigación española en Recuperación de Información (R.I.): análisis bibliométrico (1984-1999). EN *Revista de investigación Iberoamericana en Ciencia de la Información y documentación*. 2000 1 (1) p. 117-123

NICOLAIDIS, S., KALAMBOUKIS T.Z. Evaluation of stemming algorithms with modern greek 54 (en prensa)

NIEDERMAN, G. T. THURMAIR, G. & BÜTTEL, J. MARS: a retrieval tool on the basis morphological analysis En C.J. van Rijsbergen (ed.). *Research and development in information retrieval*. Cambridge: C.U.P., 1985

OARD, D. W. Alignment of Spanish and English TREC Topic Descriptions 1996 [en línea] <http://trec.nist.gov/pubs/trec5/t5_proceedings.html> [Consultado el 24/07/00]

OLVERA LOBO, Mª. D. Evaluación de los sistemas de recuperación de información: aproximaciones y nuevas tendencias. *El profesional de la información*. 1999 8 (11) p. 4-14

OLVERA LOBO, Mª. D. Métodos y técnicas para la indización y la recuperación de recursos de la World Wide Web. *Boletín de la Asociación Andaluza de Bibliotecarios.* 1999 . 57 p. 11-22

PAICE, C. D. Another Stemmer *ACM SIGIR Forum*, 1990 24 (3) p. 56-61

PAICE, C. D. *Information Retrieval and the Computer*. London: McDonal and Janes, 1977

PAICE, C. D. Method for Evaluation of Stemming Algorithms Based on Error Counting. J*ournal of the American Society for Information Science* 1966 47 (8) p. 32-649

PENA, J. La formación de verbos en español: la sufijación verbal. En VARELA, S. *La formación de palabras*. Madrid: Taurus, 1993

PENA, J. La palabra: estructura y proceso morfológicos. *Verba*. 1991 18 p. 69-128

PÉREZ ÁLVAREZ –OSSORIO, J. R. *Introducción a la información y documentación científica*. Madrid: Alhambra, 1990

PINTO MOLINA, M. *Análisis documental. Fundamentos y procedimientos*. 2ª ed. rev y aum. Madrid: EUDEMA, 1993.

PINTO MOLINA, Mª *El resumen documental: principios y métodos*. Salamanca: Fundación Germán Sánchez Ruipérez, 1992

POLLOCK, J. J. and ZAMORA, A. System Design for Detection and Correction of Spelling error in Scientific Scholary Text. *Journal of the American Society for Information Science* 1984 35(2) p. 104-109

POPOVIC M. and WILLETT, P The effectiveness of stemming for natural-language access to Slovene textual data *Journal of the American Society for Information Science* 1992 43 (5) p. 384-390

PORTER, M. F. An algorithm for Suffix Stripping *Program*, 1980 14 (3) p 130-137

RAPHAEL, B. SIR: A computer program form semantic information retrieval. *Semantic Information Procesing*, 1968 p. 33-145.

RESNIKOFF and DOLVI The Nature of Affixing in Written English Part I *Mechanical Translation*, 8 nº 3. 1965 y Part II Mechanical Translation and Computational Linguistic vol 9 n. 2 (1966)

RICH, E. and KNIGHT, K. Natural Language Processing En *Artificial Intelligence*. 2nd ed. New York: Mc Graw-Hill, 1991

RIJSBERGEN, K. V. *Information Retrieval*. 2nd Ed. London: Butterworths, 1979.

RILOFF, E. and LEHNER, W. Information Extraction as a Basis for High-Precisión Text Clasification. *ACM Transactions on Information Systems* 1996 12 (3) p. 296-333

RILOFF, E. and LORENZEN, J. Extraction-based text Categorization: Generating domain-Specific Role relationship automatically. En STRZALKOWSKI, T. *Natural language Information Retrieval*. Dordrecht: Kluwer Academic Publisher, 1999 p. 167-196

RÍOS HILARIO, A. B.. Metodología, técnicas y estrategias de investigación en las Jornadas Españolas de Documentación Automatizada (1981-1996*)* IV Jornadas Españolas de Documentación. En *Los sistemas de información al servicio de la sociedad: Actas de las jornadas*. VI Jornadas españolas de documentación. Valencia: FESABID, 1998.

ROBERTSON and SPARK JONES, K.. Relevance Weighting on Seach Term. *Journal of the American Society for Information Science* 1976 27 (3) p. 129-146

ROBERTSON, A. M. and WILLET, P. Applications of n-grams in textual information system. *Journal of Documentation* 1998 54 (1) p. 48-69

ROBERTSON, S. E. The Parametric Description of Retrieval Test Part 2 Overall measures. *Journal of Documentation*, 1969.25 (2) p. 93-107

ROLSTON, D. W. *Principios de inteligencia artificial y sistemas expertos*. Bogotá: McGraw-Hill, 1990

SALTON, G. McGRILL, M. *Introduction to Moderm Information Retrieval*. New York: McGraw Hill, 1983

SALTON, G. *Automatic Information Organization and Retrieval*. New York: McGraw-Hill, 1968.

SALTON. G. *Automatic text Procesing: the transformation, analysis and retrieval of information by computer*. Massachusset: Addison-Wesley, 1989

SALTON. G. On the relationship between theoretical Retrieval Models. EGGHE, L. ROUSSEAU, R. (ed) *Infometrics* 87/88. Amsterdam: Elsevier, 1988 p. 263-270.

SANDERSON, M. *Word Sense Disambiguation and Information Retrieval, PhD Thesis, Technical Report (TR-1997-7) of the Department of Computing Science at the University of Glasgow*, Glasgow G12 8QQ, UK.

SANTANA, O. [et al.] Flaver: Flexionador y lematizador automático de formas verbales [consultado en línea] http://protos.dis.ulpgc.es/art_ps/art28.ps [consultado el 26-10-99]

SARACEVIC, T. Relevance: A review of a framework for the thinking on the notion in Information Science. *Journal of the American Society for Information Science* 1975 26 (6) p. 321-343

SARACEVIC, T., et al. A study of information seeking and retrieving, background and metodology. *Journal of the American Society for Information Science* , 39 (3) p. 161-176.

SAVOY, J. A Stemming Procedure and Stopword List for General French Corpora. J*ournal of the American Society for Information Science* 1999 50 (10) p. 944-952.

SAVOY, J. Stemming of French Words Based on Gramatical Categories *Journal of the American Society for Information Science* 1993 44 (1). p 1-9

SCHAMBER, L. Relevance and Information behaviours. *Annual Review of Information Science and Technology (ARIST)* 1994 29 p. 3-48

SCHAMBERG, L. EINSEMERG, M.B. and. NILO M. S A re-examination of relevance: toward a dynamic, situational definition *Information Procesing and Management* 1990, 26 (6), p. 755-776.

SCHINKE, R., GREENGRAS, M. ROBERTSON, M. A., WILLETT, P. A Stemming Algorithm for Latin text databases. *Journal of Documentation* 1996 52 (2) p.172-187

SMEATON and VAN RIJSBERGEN, C.J. Experiments on Incorporating Syntactic Processing of User Queries into a Document Retrieval Strategy In Proceedings of the 11 th International ACM-SIGIR Conference on Research & Development in Information Retrieval, ed. Y. Chiaramella, Grenoble, June 1988. p. 31-51

SCHOLTES, J. C. *Artificial neural netwoks for information retrieval in Libraries context*. Brussels: Office for official Publucations of the European Commnueties, 1995

SMEATON, A. F. Using NLP or NLP resources for Information Retrieval En STRZALKOWSKI, T (ed). *Natural language Information Retrieval*. Dordrecht: Kluwer Academic Publisher, 1999 [también en línea] <http://rayela.ieccc.uned.es/~ircourse/doc/smeaton/NLP-Ir-BOOK.ps>

SMEATON, A. F., O'DONNELL, R. KELLEDY, F Thresholding Posting Lists, Query Expansion with WordNet and POS Tagging of Spanish" 1995 [en línea] http://trec.nist.gov/pubs/trec4/t4_proceedings.html [Consultado el 24/07/00]

SMEATON, A. F., O'DONNELL, R. KELLEDY, F. Indexing Structures Derived from Syntax in TREC-3: System Description 1994 [en línea] <http://trec.nist.gov/pubs/trec3/t3_proceedings.html> [Consultado el 24/07/00]

SMITH M.P. and SMITH M. The use of genetic programming to build Boolean queries for text retrieval throught relevance feedback. *Journal of documentation* 1997 23 (6) p. 423-431

SPARK JONES, K. A Statistical interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation* 1972 28 (1) p.11-20

SPARK JONES, K. and TAIT, J.L. Automatic Search Term Variant Generation *Journal of Documentation* 1984 40 (1) p 50-66.

SPARK JONES, K. Experiment in Relevance weiting on Search Term. *Information Processing and Management* 15 (3) p. 133-144.

SPARK JONES, K. The role of P.L.N. in Text Retrieval En T. STRZALKOWSKI *Natural language Information Retrieval*. Dordrecht: Kluwer Academic Publisher, 1999

SPINK, A. and LOSEE, R. M. Feedback in Information Retrieval. *Annual Review of Information Science and Technology* 1996, vol 31. p. 31-81.

STRZALKOWSKI, T. (ed) *Natural Languaje Information Retrieval*. Dordrecht: Klwer Academic Publisher, 1999

SU, L.T. The Relevance of Recall and Precision in User Evaluation. *Journal of the American Society for Information Science* 1994 45 (3) p. 207-217

SUN, Q. SHAW, D. And DAVID, C.H. A model or Estimating of de Occurrencie of Same frecuenciy Words and the Boundary between High and Low Frequency Words in Text *Journal of The American Society for Information Science* 1999 50 (3) p. 280-286.

Survey of the State of the Art in Human Language Technology. Oregon: National Science Fundation, 1995

SWANSON, D. R. Subjetive versus objetive relevance in bibliographic retrieval system. *Library Quartely* 1986 56, p. 389-398

SWETS, J. A. Information-retrieval Systems *Science,* 141 (3577): July 1963 p. 245-250

TAGUE-STUCLIFFE, J. M. Some Perspectives on the Evaluation of Information Retrieval Systems *Journal of the American Society for Information Science* 1996, 47 (1) p. 1-3

TAYLOR, A. G. *The organization of information*. Englewood: Libraries Unlimited Inc, 1999.

The Porter Stemming Algorithms [on-line] <http://www.muscat.com/~martin/stem.html> [Consultado el 20/07/00] 40

VARELA ORTEGA, S. *Fundamentos de morfología*. Madrid: Sínteis, 1992

VARELA ORTEGA, S. *La formación de palabras*. Madrid: Taurus, 1993

WALKER, G. JANES, J. Online Retrieval: a Dialoge of Theory and Practice. Englewood: Libraries Unlimited, Inc, 1993.

WILKINSON, R. ZOBEL, J. SACKS-DAVIS, R. Similarity Measures for Short Queries 1995 [en línea] <http://trec.nist.gov/pubs/trec4/t4_proceedings.html> [Consultado el 24/07/00]

WILSON, P. Situacional relevance Information *Storage and Retrieval* 1973.9 p. 457-469

WONG, B. K. BODNOVICH, T. A. A Bibliography of genetic algorithm bussiness application research: 1988-1996. *Expert Systems* 1998 15 (2) p. 75-82

WONG, B. K. BODNOVICH, T. A. SELVY, Y. A bibliography of neural network business applications research: 1988 september 1994. *Expert Systems* 1995 vol 12 (3) p. 253-261

WORNG, S. K RAGHAVA, M. *Vector Space model of information retrieval*. Research & Development in Information Retrieval. Cambridge: University Pres, 1984.

WRITHT, R. *Latín tardío romance temprano en España y la Francia Carolingia*. Madrid: Gredos, 1989

YANG, J., KORFHAGE, R. R. Adaptative information retrieval system in vector model. *Symposium on Document Analysis and Information Retrieval*. Las Vegas p. 134-150

YAO , J. J. Measuring Effectiveness Bases on User Preference of Documents. *Journal of the American Society for Information Science* 1995 46 (2) p. 133-145

YOUNG, H. (ed) *Glosario ALA de bibliotecología y ciencias de la Información*. Madrid: Díaz de Santos, 1988.

YU, K-I., SCHEIBE, P. , NORDBY, F. The FDF Query Generation Workbench. *Trec-3* [en línea] http://trec.nist.gov/pubs/trec3/t3_proceedings.html [consultado el 10-01-01]

ZIPF, H. P. *Human behaviour and the Principle or least Effort* Addison Wesley. Massachusett: Cambridge, 1949.

GLOSARIO DE TÉRMINOS.

Activación de la diseminación: véase **Spreading disemination**.

Adjetivación: mecanismo por el cual se forman adjetivos tomando como base nombres o verbos.

Adverbialización: mecanismo por el cual se forman adverbios tomando como base adjetivos, u otros adverbios.

Algebra booleana: sistema matemático de funciones lógicas que relaciona entre sí los términos y por lo tanto los datos, únicamente por medio de los operadores Y (and) , O (or) , NO (not).

Algoritmo: proceso definido o conjunto de reglas secuenciales y preestablecidas para la resolución de un problema específico.

Algoritmo de lematización: algoritmo que pretende reducir las variantes de inflexión y derivación de las palabras a formas comunes.

Algoritmo genético: programas de ordenador cuyo fin es imitar el proceso de selección natural según la teoría de Darwing, por el cual las poblaciones crecen por el cambio de genes, perviven las mejores soluciones y se descartan las peores.

Alomorfo: variante de los sufijos con un mismo origen y forma parecida con el mismo sentido que la forma principal.

Autómata de estados finitos: Es un mecanismo teórico que comienza en un estado particular y cambia de estado cuando se dan determinadas condiciones. La red básica consiste en un grupo de nodos conectados por arcos. Cada nodo representa un estado en máquina de estados finitos y los arcos muestran las transiciones desde un estado a otro. Un autómata comprende un conjunto de

estados, una serie de categorías léxicas; las funciones de transición, un estado inicial y un conjunto de estados finales. La cadena es aceptada siempre que la sucesión de arcos llegue desde el principio hasta el fin.

Autómata finito determinista: máquina analizadora de cadenas que acepta aquellas aceptadas por su diagrama de transiciones interno y rechaza las demás.

Autómata finito no determinista: máquina analizadora de cadenas que acepta una cadena siempre que es posible que su análisis deje la máquina en un estado de aceptación.

Browsing: consulta rápida, lectura por encima, lectura superficial. Lectura de un documento o consulta de fondos sin prisa.

Centroide: documento que representa al conjunto de documentos que pertenecen a la misma clase.

Clasificación: proceso de decidir la categoría apropiada para un documento.

Cluster: es un grupo de documentos con contenido similar. En un modelo de espacio vectorial, la recuperación se puede hacer comprando el vector de la pregunta con los centroides de los cluster.

Complemento de fallout: véase **índice de irrelevancia**.

Complemento del índice de irrelevancia: documentos no relevantes no recuperados partido los documentos no relevantes.

Complemento del ratio de exhaustividad: documentos relevantes no recuperados partido los relevantes.

Complemento del ratio de precisión: documentos no relevantes recuperados partidos los documentos recuperados.

Composición: combinación de lexemas independientes para formar un término nuevo.

Conflación: acto de poner juntas dos o más palabras de manera que puedan ser tratadas de manera conjunta.

Coocurrencia: relación de aparición conjunta de entre términos.

Derivación: procedimiento de formación de palabras mediante la adicción, supresión o intercambio de afijos.

Descriptor: en indización cada una de las palabras clave significativas que expresan y representan el contenido de un documento.

Diagrama de estados: Véase **diagrama de transiciones**.

Diagrama de transiciones: colección finita de círculos los cuales se puede rotular conectado por flechas que reciben el nombre de arcos.

Discard: véase **Indice de irrelevancia**.

Efectividad de un sistema: medida por la que se relaciona la satisfacción del usuario con la salida porporcionada por el sistema.

Eficacia de un sistema: capacidad que el sistema tiene de satisfacer las necesidades de información de los usuarios, recuperando el material solicitado y rechazando el no deseado.

Esfuerzo de exhaustividad: es el ratio del número de documentos relevantes deseados partido por el número de documentos examinados para encontrar el número de documentos relevantes deseados.

Especificidad: véase **complemento del índice de irrelevancia**.

Exhaustividad de recuperación: proporción de material relevante que se ha recuperado.

Exhaustividad relativa: ratio de documentos relevantes recuperados, examinados por el usuario, partido por el número de documentos que el usuario quiere examinar.

Factor de exhaustividad: Véase **exhaustividad.**

Factor de pertinencia: Véase **pertinencia.**

Factor de ruido: Véase **ruido.**

Fallout: véase **Indice de irrelevancia**

Fichero invertido: colección de índice donde cada término aparece ordenado en una lista con indicación de aquellos documentos donde aparece.

Frecuencia del término: véase **tf**.

Generalidad: valor absoluto de los documentos recuperados partido el número de documentos.

IDF: medida de la frecuencia particular de la aparición de los términos en la colección.

Indice de irrelevancia: se obtiene de dividir los documentos recuperados e irrelevantes a la pregunta entre el total de los documentos contenidos en la colección.

Indización: operación documental dirigida a representar por medio de un lenguaje documental o natural los datos resultantes del análisis de contenido de un documento o una demanda de información.

Infralematización: problema que se da cuando quitamos un sufijo más pequeño del que debemos. Provoca silencio informativo.

Interfijo: afijo que va en posición intermedia.

Inverso de la frecuencia de aparición de un término en un documento: véase **IDF.**

Lema: etiqueta utilizada en informática para hacer referencia a la base sobre la que actúan las distintas formas flexivas y derivativas

Lematización: acción de extraer la esencia, es decir, el tema o lema de una palabra.

Lenguaje controlado: léxico construido con la ayuda de un conjunto de reglas que trata de reperesetar unívocamente el contenido de los documentos y las demandas.

Lenguaje natural: lenguaje no manipulado.

Longitud de búsqueda esperada: que es el número de documentos no buscados que el usuario puede esperar examinar antes de encontrar el número de documentos deseados.

Medida de F: media que relaciona la precisión y la exhaustividad

Modelo probabilístico: modelo que considera la probabilidad de que un termino aparezca en un documento o la probabilidad de que un documento satisfaga la necesidad de información**.**

N-grama: es una ventana que contiene n caracteres que se va desplanzando a lo largo de un texto.

Nominalización: mecanismo por el cual se forman nombres, tomando como base adjetivos, verbos, u otros nombres

Número de generalidad: que relaciona el número de documentos que son relevantes para una pregunta concreta entre el número total de documentos en la colección.

P.N.L. véase **Procesamiento del lenguaje natural**.

Palabra clave: palabra o grupo de palabras seleccionadas, de cualquier parte del documento o de la demanda de información para representar el contenido bien de la pregunta o del documento.

Palabra vacía: palabra que de manera autónoma no tiene significado, suelen pertenecer a este grupo las preposiciones, conjunciones, algunos adverbios...

Parasísntesis: proceso de derivación en el cual un prefijo y un sufijo actúan simultáneamente sobre una base sin que exista la forma sin prefijar.

Part-of speach-taggin: etiquetador de textos

Pertinencia: es la medida de cómo un documento se ajusta a una necesidad informativa.

Pesado: valor numérico que expresa la capacidad de describir un texto.

Polling: método de cálculo de la relevancia donde sobre una base de datos se trata de extraer los documentos relevantes a una pregunta o a un conjunto de ellas mediante diversos sistemas. El conjunto resultante se analiza para extraer los que se consideran realmente relevantes.

Post-taggin: etiquetado de textos.

Precisión: proporción del material recuperado realmente relevante del total de documentos recuperados.

Prefijo: afijo que precede a la base

Probabilidad condicional de bajada falsa: véase **Indice de irrelevancia.**

Probabilidad condicional de una búsqueda: véase **Complemento del ratio de exhaustividad.**

Probabilidad condicional de una correcta respuesta negativa: véase **complemento de fallout**.

Probabilidad condicional de una pérdida: véase complemento del ratio de exhaustividad.

Procesamiento del lenguaje natural: estudio y análisis de los aspectos lingüísticos de un texto a través de los programas informáticos.

Ratio de aceptación: Véase **precisión**

Ratio de cobertura: es la proporción de documentos relevantes conocidos por el usuario que son actualmente recuperados.

Ratio de esnobismo: véase **Complemento del ratio de exhaustividad**.

Ratio de novedad: proporción de documentos relevantes recuperados que previamente son conocidos por el usuario.

Ratio de precisión: véase **factor de ruido.**

Realimentación de consultas: véase **realimentación de la relevancia**.

Realimentación de la relevancia: proceso por el cual una pregunta se expande utilizando los términos de los primeros documentos recuperados, el número varía en función de los sistemas, para obtener mayor exhaustividad en la recuperación.

Recall**:** véase **exhaustividad**.

Recuperación de la información: conjunto de operaciones encaminadas a la interpretación de necesidades de información y la recuperación de la misma mediante medios automáticos.

Red de transiciones: véase **diagrama de transiciones.**

Redes de neuronas artificiales: véase **redes neuronales.**

Redes neuronales: modelos informáticos inspirados en la estructura de bajo nivel del cerebro

Relevancia: es la medida de cómo una pregunta se ajusta a un documento.

Rellamada: Véase **exhaustividad**.

Retroalimentación de la relevancia: véase **realimentación de la relevancia**.

Ruido: datos obtenidos en la recuperación que sobrepasan en profundidad, superficialidad o extensión lo estrictamente solicitado.

Sensibilidad: Véase **exhaustividad**.

Silencio: datos solicitados al sistema en un módulo de búsqueda pero no obtenidos, aún existiendo, debido a la distorsión de un proceso.

Similaridad: medida con la que se indica el grado de coincidencia en la comparación de una pregunta con el contenido de una base de datos.

Sistemas basados en el conocimiento: véase **sistema experto.**

Sistemas inteligentes: véase **sistema experto.**

Sistema experto: programas que ofrecen soluciones a problemas simulando el proceso de razonamiento humano mediante la aplicación específica del conocimiento y la inferencia.

Stop list: véase **palabras vacías**.

Diccionario negativo: véase **palabras vacías**.

Sobrelematización: problema que se da cuando quitamos un sufijo más largo del que debemos. Provoca ruido informativo.

Spreading disemination: técnica de R.I. que pertenece a la categoría de red, en la cual mediante una pregunta se activan las partes de la red que describen el contenido del documento al que se hace referencia en la pregunta.

Sufijo: afijo que se pospone a la base.

Tabla de transiciones: matriz bidimensional cuyos elementos proporcionan el resumen del diagrama de transiciones correspondiente.

tf: frecuencia de aparición del término t en un determinado documento.

Umbral de futilidad: punto a partir del cual se considera que al usuario ya no le interesan los documentos.

Variante alomórfica: véase **alomorfo.**

Vector: lista finita de elementos en un orden determinado. Cada elemento hace referencia en función de su posición.

Verbalización: mecanismo por el cual se forman verbos tomando como base adjetivos, nombres, adverbios, pronombres u otros verbos.

APÉNDICE

En el siguiente apéndice, se muestran los resultados de preción para la cada el experimento *sin lematizar* con palabras vacías, y suprimiendo según ambas listas. En cuanto a la lematización los resultados son los de la lematización quitando previamente las palabras vacías según las listas *vacías leve* y *vacías fuerte.*

1. - **La formación continuada de los profesionales de archivos, bibliotecas y museos**

16 documentos relevantes.

Sin lematizar

	Con vacías	**Sin vacías leve**	**Sin vacías fuerte**
10	0,6	0,6	0,5
20	0,35	0,35	0,35
30	0,2666	0,2666	0,2666
50	0,2	0,2	0,2
100	0,13	0,13	0,13

Lematización derivativa

	Sin vacías leve	**Sin vacías fuerte**
10	0,4	0,4
20	0,35	0,35
30	0,233	0,3
50	0,18	0,2
100	0,15	0,15

Lematización flexiva

	Sin vacías leve	**Sin vacías fuerte**
10	0,5	0,5
20	0,35	0,35
30	0,3	0,3
50	0,22	0,22
100	0,13	0,13

2.- **La gestión de los recursos humanos, administrativos, económicos y financieros de los centros de documentación, archivos y bibliotecas**

24 documentos relevantes.

Sin lematizar

	Con vacías	**Sin vacías leve**	**Sin vacías fuerte**
10	0,5	0,5	0,5
20	0,25	0,25	0,25
30	0,166	0,166	0,166
50	0,18	0,18	0,18
100	0,15	0,15	0,15

Lematización derivativa

	Sin vacías leve	**Sin vacías fuerte**
10	0,6	0,6
20	0,35	0,35
30	0,266	0,266
50	0,18	0,18
100	0,15	0,14

Lematización flexiva

	Sin vacías leve	**Sin vacías fuerte**
10	0,6	0,6
20	0,35	0,35
30	0,3	0,3
50	0,2	0,2
100	0,15	0,15

3.- **Las actividades de animación a la lectura y de expansión de la biblioteca infantil y juvenil.**

27 documentos relevantes

Sin lematizar

	Con vacías	**Sin vacías leve**	**Sin vacías fuerte**
10	0,7	0,7	0,7
20	0,45	0,45	0,45
30	0,4	0,4	0,4
50	0,34	0,34	0,34
100	0,2	0,2	0,2

Lematización derivativa

	Sin vacías leve	**Sin vacías fuerte**
10	0,6	0,6
20	0,4	0,4
30	0,366	0,366
50	0,28	0,28
100	0,19	0,19

Lematización flexiva

	Sin vacías leve	**Sin vacías fuerte**
10	0,6	0,6
20	0,5	0,5
30	0,466	0,466
50	0,38	0,38
100	0,21	0,21

4.- **¿En qué consiste la gestión de calidad y cómo influye en los proyectos de planificación y gestión de las bibliotecas, archivos y museos españoles?.**

14 documentos relevantes

Sin lematizar

	Con vacías	**Sin vacías leve**	**Sin vacías fuerte**
10	0,8	0,8	0,8
20	0,55	0,55	0,55
30	0,4	0,4	0,4
50	0,26	0,26	0,26
100	0,13	0,13	0,13

Lematización derivativa

	Sin vacías leve	**Sin vacías fuerte**
10	0,6	0,6
20	0,55	0,55
30	0,366	0,366
50	0,28	0,28
100	0,14	0,14

Lematización flexiva

	Sin vacías leve	**Sin vacías fuerte**
10	0,9	0,9
20	0,5	0,5
30	0,4	0,4
50	0,26	0,26
100	0,13	0,13

5.- **¿Qué funciones son propias del bibliotecario y cuales del auxiliar. Qué función tiene cada uno de ellos?**

7 documentos relevantes

Sin lematizar

	Con vacías	**Sin vacías leve**	**Sin vacías fuerte**
10	0,1	0,3	0,3
20	0,15	0,15	0,15
30	0,1	0,133	0,1666
50	0,08	0,1	0,1
100	0,05	0,06	0,06

Lematización derivativa

	Sin vacías leve	**Sin vacías fuerte**
10	0,3	0,2
20	0,15	0,15
30	0,1	0,1
50	0,08	0,08
100	0,04	0,04

Lematización flexiva

	Sin vacías leve	**Sin vacías fuerte**
10	0,2	0,2
20	0,2	0,15
30	0,133	0,133
50	0,08	0,08
100	0,05	0,05

6.- **¿Qué salidas ofrece el mercado de trabajo para los licenciados y diplomados en documentación?**

12 documentos relevantes

Sin lematizar

	Con vacías	**Sin vacías leve**	**Sin vacías fuerte**
10	0,6	0,6	0,6
20	0,4	0,4	0,4
30	0,3	0,3	0,3
50	0,22	0,22	0,22
100	0,11	0,11	0,11

Lematización derivativa

	Sin vacías leve	**Sin vacías fuerte**
10	0,6	0,6
20	0,45	0,45
30	0,333	0,333
50	0,22	0,22
100	0,11	0,11

Lematización flexiva

	Sin vacías leve	**Sin vacías fuerte**
10	0,7	0,7
20	0,4	0,4
30	0,266	0,266
50	0,22	0,22
100	0,11	0,11

7.- **Colaboraciones entre bibliotecas públicas y bibliotecas de centros de educación primaria y secundaria**

14 documentos relevantes

Sin lematizar

	Con vacías	**Sin vacías leve**	**Sin vacías fuerte**
10	0,6	0,6	0,6
20	0,4	0,4	0,4
30	0,2666	0,2666	0,2666
50	0,2	0,18	0,2
100	0,11	0,11	0,11

Lematización derivativa

	Sin vacías leve	**Sin vacías fuerte**
10	0,7	0,7
20	0,45	0,45
30	0,3	0,3
50	0,18	0,18
100	0,1	0,1

Lematización flexiva

	Sin vacías leve	Sin vacías fuerte
10	0,8	0,8
20	0,45	0,45
30	0,3	0,3
50	0,18	0,18
100	0,12	0,12

8.- **Cómo se puede llevar a cabo la formación de usuarios y qué experiencias hay en bibliotecas, centros de documentación y de información especializados**

14 documentos relevantes

Sin lematizar

	Con vacías	**Sin vacías leve**	**Sin vacías fuerte**
10	0,4	0,4	0,4
20	0,3	0,25	0,3
30	0,233	0,233	0,233
50	0,18	0,18	0,18
100	0,12	0,12	0,12

Lematización derivativa

	Sin vacías leve	**Sin vacías fuerte**
10	0,4	0,5
20	0,25	0,25
30	0,166	0,233
50	0,16	0,2
100	0,11	0,11

Lematización flexiva

	Sin vacías leve	**Sin vacías fuerte**
10	0,5	0,5
20	0,4	0,35
30	0,3	0,3
50	0,2	0,2
100	0,11	0,12

9.- **¿Cúal es el perfil de los profesionales en el mundo de la información: bibliotecarios, archiveros y documentalistas?**

15 documentos relevantes

Sin lematizar

	Con vacías	**Sin vacías leve**	**Sin vacías fuerte**
10	0,4	0,4	0,4
20	0,4	0,4	0,4
30	0,266	0,3	0,333
50	0,2	0,2	0,2
100	0,11	0,11	0,12

Lematización derivativa

	Sin vacías leve	**Sin vacías fuerte**
10	0,5	0,6
20	0,4	0,35
30	0,3	0,333
50	0,24	0,24
100	0,12	0,12

Lematización flexiva

	Sin vacías leve	**Sin vacías fuerte**
10	0,7	0,7
20	0,55	0,6
30	0,4	0,4
50	0,28	0,28
100	0,14	0,14

10.- **Fuentes para la selección de obras de literatura infantil y juvenil.**

15 documentos relevantes

Sin lematizar

	Con vacías	**Sin vacías leve**	**Sin vacías fuerte**
10	0,5	0,5	0,5
20	0,45	0,45	0,5
30	0,366	0,366	0,366
50	0,24	0,24	0,24
100	0,15	0,15	0,15

Lematización derivativa

	Sin vacías leve	**Sin vacías fuerte**
10	0,4	0,5
20	0,5	0,5
30	0,366	0,366
50	0,24	0,24
100	0,15	0,15

Lematización flexiva

	Sin vacías leve	**Sin vacías fuerte**
10	0,5	0,5
20	0,55	0,5
30	0,366	0,366
50	0,26	0,26
100	0,15	0,15

11.- **Las fuentes de información en soporte óptico**

13 documentos relevantes

Sin lematizar

	Con vacías	**Sin vacías leve**	**Sin vacías fuerte**
10	0,1	0,1	0,1
20	0,15	0,15	0,15
30	0,1	0,1	0,1
50	0,08	0,08	0,08
100	0,05	0,05	0,05

Lematización derivativa

	Sin vacías leve	**Sin vacías fuerte**
10	0,1	0,1
20	0,1	0,1
30	0,1	0,1
50	0,12	0,12
100	0,06	0,06

Lematización flexiva

	Sin vacías leve	Sin vacías fuerte
10	0,2	0,2
20	0,1	0,1
30	0,1	0,1
50	0,08	0,08
100	0,05	0,05

12.- **¿Qué es el marketing?. Aplicación en bibliotecas, archivos, museos, centros de documentación e información.**

8 documentos relevantes

Sin lematizar

	Con vacías	**Sin vacías leve**	**Sin vacías fuerte**
10	0,3	0,3	0,3
20	0,2	0,25	0,25
30	0,166	0,166	0,166
50	0,1	0,1	0,1
100	0,06	0,05	0,05

Lematización derivativa

	Sin vacías leve	**Sin vacías fuerte**
10	0,3	0,3
20	0,25	0,2
30	0,166	0,133
50	0,1	0,1
100	0,06	0,06

Lematización flexiva

	Sin vacías leve	**Sin vacías fuerte**
10	0,3	0,2
20	0,2	0,25
30	0,133	0,166
50	0,1	0,1
100	0,06	0,06

13.- **La edición electrónica**

16 documentos relevantes

Sin lematizar

	Con vacías	**Sin vacías leve**	**Sin vacías fuerte**
10	0,5	0,5	0,5
20	0,4	0,4	0,4
30	0,266	0,266	0,266
50	0,24	0,24	0,24
100	0,13	0,13	0,13

Lematización derivativa

	Sin vacías leve	Sin vacías fuerte
10	0,6	0,6
20	0,4	0,4
30	0,366	0,366
50	0,26	0,26
100	0,15	0,15

Lematización flexiva

	Sin vacías leve	Sin vacías fuerte
10	0,5	0,5
20	0,5	0,5
30	0,33	0,33
50	0,22	0,22

14.- **La industria editorial en España**

26 documentos relevantes

Sin lematizar

	Con vacías	**Sin vacías leve**	**Sin vacías fuerte**
10	0,5	0,6	0,6
20	0,35	0,35	0,35
30	0,333	0,333	0,333
50	0,28	0,28	0,28
100	0,16	0,16	0,16

Lematización derivativa

	Sin vacías leve	**Sin vacías fuerte**
10	0,5	0,6
20	0,45	0,45
30	0,4	0,4
50	0,3	0,28
100	0,13	0,13

Lematización flexiva

	Sin vacías leve	Sin vacías fuerte
10	0,6	0,6
20	0,45	0,45
30	0,366	0,366
50	0,34	0,34
100	0,24	0,24

15.- **El impacto de Internet en el mundo de la difusión de la información**

12 documentos relevantes

Sin lematizar

	Con vacías	Sin vacías leve	Sin vacías fuerte
10	0,5	0,5	0,5
20	0,4	0,4	0,35
30	0,3	0,3	0,3
50	0,2	0,2	0,22
100	0,12	0,12	0,12

Lematización derivativa

	Sin vacías leve	**Sin vacías fuerte**
10	0,6	0,6
20	0,35	0,35
30	0,3	0,266
50	0,18	0,22
100	0,12	0,12

Lematización flexiva

	Sin vacías leve	**Sin vacías fuerte**
10	0,6	0,6
20	0,35	0,4
30	0,3	0,3
50	0,22	0,2
100	0,12	0,12