ISSN: 1579-9794

La recuperación de información y la traducción: hitos en el proceso de globalización

(Information recovery and translation: landmark in the globalization process)

MANUEL MARCOS ALDÓN
(Universidad de Córdoba)

Fecha de recepción: 18 de mayo de 2010
Fecha de aceptación: 14 de junio de 2010

Resumen: La información en sus diversas modalidades tanto la científica como la humanística y, en concreto, el proceso intercultural de la traducción, son analizados como hitos que integrados en un modelo de análisis permiten clarificar el proceso mundial de la globalización. No es posible pensar en el desarrollo global sin considerar la recepción y producción por las sociedades occidentales del proceso informativo y traductológico. Ítem más las traducciones como bien indicó Itama Ben-Zohar son indicadores naturales del proceso de recepción e interconexión entre diversas culturas; y de ahí la necesidad de generar modelos de recuperación especializados en lugar de motores de búsquedas no especializados.

Palabras clave: Globalización, Información, Documentación aplicada, Traducción, Recuperación de información.

Abstract: Information in its various modalities, namely the scientific and the humanistic and, more specifically, the intercultural process of translation, are all analysed as hallmarks which integrated in a model of analysis allow us to clarify the world process of globalization. It is not possible to think of a global development without considering the reception and production of the informative and translational process on the part of western societies. Furthermore, translations, as Ben-Zohar already noted. Hence, the need to generate models of specialized retrieval instead of non specialized search engines.

Key words: Globalization, Information, Applied Documentation, Translation, Information Retrieval.

INTRODUCCIÓN

Cualquiera que se haya aproximado a un motor de búsqueda en alguna ocasión, es decir, cualquier usuario, especializado o no, podría considerar su primera búsqueda en ellos como una metáfora borgiana.

Nuestra primera búsqueda produce comúnmente millones de resultados cuando no decenas o cientos de millones de resultados. Revisar, aunque sea someramente, estos resultados ocuparía toda nuestra vida a la manera de el inagotable Libro de arena o la visión unitaria universal de El Aleph.

Partiendo de esta premisa han existido desde la publicación de la emblemática obra de Shannon y Weaver una necesidad de ampliar los procesos de cálculo para resolver la entropía informativa. En su pionera obra ambos autores fijaron en su cuarto teorema un proceso matemático de comunicación en el que no se consideraba el valor semántico. En el caso español los trabajos de S. Montes y R. Pérez-Amat analizando este teorema e integrando una modificación con la entropía de Luca-Termini, posibilitando un desarrollo de un proceso de búsqueda con carácter significativo sobre el valor semántico de los mensajes. Esto permitía mejorar los procesos de análisis heurístico para búsquedas especializadas.

Ambos modelos no dejan de ser modelos de análisis para la recepción y envío de los mensajes en un sistema de información global. Dado este carácter último es conveniente comprender y analizar el proceso global de información para comprender y construir mejores sistemas de comunicación y, en nuestro caso, de recuperación documental, puesto que la aceleración o ralentización del proceso de globalización está indudablemente marcada actualmente por el proceso de recuperación de información. Para lograr este objetivo se han utilizado como muestras de los hitos de traducción y de información las fechas propuestas por obras y entidades internacionales como Anthony Pym (2010): *Attempt at a chronology of Hispanic translation history*, en Index Translationum, Unesco: http://www.tinet.cat/~apym/on-line/chronology/8-11.html [22 de octubre de 2010]; para el mundo Árabe: MuslimHeritage.com, *The impact of translations of Muslim Sciences on the West*, en Index Translationum, Unesco:
http://www.muslimheritage.com/topics/default.cfm?TaxonomyTypeID=22&TaxonomySubTypeID=114&TaxonomyThirdLevelID=-1&ArticleID=344; [22 de octubre de 2010],
para América Latina, hitos obtenidos de HISTAL, *Historia de la Traducción en América Latina*, en Index Translationum, Unesco: http://www.histal.umontreal.ca/espanol/versionsp.htm [22 de octubre de 2010] y de forma general de *Los Traductores en la Historia*, Editorial Universidad de Antioquia, 2005, bajo el patrocinio de la Unesco

Puede definirse como globalización el proceso asociado al desarrollo social mediante el cual se produce un incremento en la dependencia de todo tipo entre los individuos, entidades y naciones. A partir de la definición anterior y del análisis del desarrollo histórico de la Sociedad, el modelo para caracterizar el proceso de globalización de la sociedad

propuesto en este trabajo, se resume en los siguientes postulados:

Es una Ley Universal del desarrollo social que conduce a la progresiva mayor interdependencia entre los individuos, entidades y naciones, que ha estado presente a lo largo de la historia, y por tanto, no es un fenómeno nuevo ni creado por la actividad del hombre.

El desarrollo científico–técnico es la causa de la manifestación de su existencia.

Su caracterización requiere de la identificación de eventos de gran impacto cuya duración pueda ser estimada con un nivel de certeza, por lo cual lo más razonable es recurrir a los procesos de comunicación.

Por último, es conveniente señalar que la creación de sistemas de comunicación globales (véase la Teoría General de los Sistemas de Ludwig Von Berttalanfy) en la Conferencia de Breton Wood (1944), constituye un antecedente de globalización consciente, aunque con diferentes enfoques y objetivos, al igual que la actual tendencia de formación de bloques regionales como la Unión Europea, el Tratado de Libre Comercio de México, USA y Canadá, el MERCOSUR, etc.

1. Modelo propuesto

De todo lo expuesto se aprecia que, para establecer un modelo para el objeto de estudio, es necesario identificar procesos de relevancia mundial que exhiban regularidades en su comportamiento que puedan expresarse cuantitativamente. Como último aspecto antes de pasar a la formulación del modelo, es necesario señalar que el carácter social de los fenómenos y procesos que se abordarán les confiere el atributo de fronteras temporales borrosas, al resultar muy difícil, y en ocasiones imposible, determinar la fecha exacta de comienzo y terminación del evento.

Considerando todo lo expuesto, para la elaboración del modelo se utilizaron como eventos de gran impacto dos procesos de tipo histórico: el auge y desarrollo de la imprenta, con sus sucesivas mejoras y perfeccionamientos con una duración de 406 años aproximadamente y el proceso de desarrollo de la Traducción en la Edad Contemporánea, cuya duración es de 201 años, iniciado por los procesos revolucionarios sociales y económicos como la Revolución Francesa y los movimientos de liberación en América, que por pertenecer a la Historia Moderna y Contemporánea, permite disponer de fechas concretas de inicio y fin:

El comportamiento antes descrito, puede expresarse matemáticamente a través de las dos ecuaciones siguientes, donde los símbolos utilizados tienen el significado que se indica:

$$\Delta t_j = \frac{400}{2^j}$$

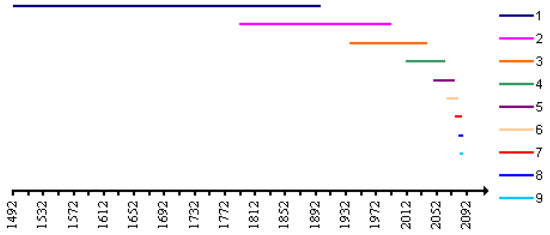
$$I_n = 1492 + \sum_{j=0}^{n} 0.74\Delta t_j$$

Δt_j: duración del evento j, considerando como referencia (j=0) el proceso de hitos de traducción especializada y de publicación de la información.

I_n: expresa la fecha (año) en la cual se inicia el evento n.

Gráficamente, la magnitud I_n puede representarse como se indica en la figura, donde se aprecia que a medida que transcurren los años la duración de los procesos de gran impacto se reduce, hasta que a partir de un momento es prácticamente cero.

Figura 2. Presencia de eventos de gran impacto



La interpretación de este pronóstico es que, a medida que el proceso de globalización avanza los fenómenos de gran impacto duran menos debido a que el nivel de la globalización los facilita. Por tanto, el límite de la magnitud I_n representa en términos prácticos que la sociedad ha alcanzado un estado de globalización total, es decir, el modelo propuesto pronostica que el proceso de globalización de toda la sociedad se alcanzará en el 2084.

Sin embargo, en el modelo anterior, aunque coherente, se aprecia que permanecen fuera del modelo eventos importantes anteriores a la Revolución Francesa por lo que podría sufrir claramente una desviación el proceso de demostración. Esta variables es por lo tanto, necesario cuantificarla y así poder retroalimentar la inconsistencia del modelo.

Esta posible inconsistencia del modelo original puede interpretarse, al menos, desde los dos puntos de vista siguientes: el modelo es un fracaso y se renuncia a la coherencia de todo lo que es capaz de predecir, en particular lo relacionado con la globalización total de la sociedad en el año 2084 o considerar que el modelo es válido, pero requiere de una reformulación, que contribuya a su universalidad. Utilizando este último enfoque, un análisis de los eventos que no recoge el modelo inicial

evidencia que éstos anteceden al inicio del proceso globalizador en un 12% de la duración del evento precedente, lo cual es consistente con el modelo original en los aspectos siguientes: su ocurrencia está cuantificada en relación con el evento anterior: 12% de su duración.

Esta magnitud representa aproximadamente la mitad de la magnitud que indica la diferencia de tiempo entre el final de un proceso en ejecución y el inicio de uno nuevo. Nótese que la reducción de la duración de un período a otro es también la mitad, luego esta relación está presente en el modelo original y no resulta descabellado que dicho comportamiento se repita en otra magnitud característica.

Históricamente, la modificación al modelo original puede interpretarse desde este enfoque como la existencia de eventos que sirven de índice de la próxima ocurrencia de procesos relevantes, es decir, constituyen un avance de lo que sucederá en el futuro próximo cuando un proceso de gran impacto se presente.

Asumiendo como válido este comportamiento, los procesos de difusión y desarrollo de la imprenta, por ejemplo, funcionan en su proceso de perfeccionamiento como antecedentes de sistemas de información y comunicación posteriores a gran escala y que facilitan el proceso y desarrollo de la traducción en fases posteriores, con lo cual se evita una refutación causal integrándose los eventos de información como ejes generadores de eventos de globalización en modelos de recepción de macroestructuras sociales y económicas.

Otra variante es considerar que estos eventos pertenecen al proceso de escala planetaria, en cuyo caso el comportamiento antes descrito, puede expresarse matemáticamente a través del modelo modificado caracterizado mediante las dos ecuaciones siguientes, donde los símbolos utilizados tienen el mismo significado que en el modelo original:

$$\Delta t_j = \frac{500}{2^j}$$

$$I_n = 1398 + \sum_{j=0}^{n} 0.7\Delta t_j$$

Cálculo del límite de I_n para el modelo original

De acuerdo al modelo propuesto $\mathbf{I_n}$ viene dado por:

$$I_n = 1492 + \sum_{j=0}^{n} 0.74\Delta t_j$$

Sustituyendo en la expresión anterior el valor

$\Delta t_j = \frac{400}{2^j}$, se obtiene:

$$I_n = 1492 + \sum_{j=0}^{n} \frac{0.74*400}{2^j} = 1492 + 0.74*400 \sum_{j=0}^{n} \frac{1}{2^j} = 1492 + 296 \sum_{j=0}^{n} \frac{1}{2^j}$$

Si se toma límite cuando
$n \to \infty$

en la expresión se obtiene el valor de I_∞ como se indica a continuación:

$$I_\infty = \lim_{n\to\infty} I_n = \lim_{n\to\infty} \left(1492 + 296 \sum_{j=0}^{n} \frac{1}{2^j} \right) = 1492 + 296 \lim_{n\to\infty} \sum_{j=0}^{n} \frac{1}{2^j} = 1492 + 296 \sum_{j=0}^{\infty} \frac{1}{2^j} \qquad (1)$$

Ahora, $S_x = \sum_{j=0}^{\infty} \frac{1}{2^j}$, es una serie geométrica convergente, cuya suma (S) puede determinarse a través de la fórmula $S = \sum_{j=0}^{\infty} a_o r^j = \frac{a_0}{1-r}$, que en el caso de S_x los parámetros r y a_0 se corresponden con los valores 0.5 y 1. Sustituyendo estas magnitudes en la expresión para S_x se obtiene que el valor S_x viene dado por:

$$S_x = \sum_{j=0}^{\infty} \frac{1}{2^j} = \frac{1}{1-0.5} = 2$$

Sustituyendo el valor de S_x en la expresión (1) se obtiene que el valor de $I_\infty = \lim_{n\to\infty} I_n$ se corresponde con el año 2084.

Cálculo del límite de I_n para el modelo modificado

De acuerdo al modelo propuesto **I_n** viene dado por: $I_n = 1398 + \sum_{j=0}^{n} 0.7\Delta t_j$.

Sustituyendo en la expresión anterior el valor $\Delta t_j = \frac{500}{2^j}$, se obtiene:

$$I_n = 1398 + \sum_{j=0}^{n} \frac{0.7 * 500}{2^j} = 1398 + 0.7 * 500 \sum_{j=0}^{n} \frac{1}{2^j} = 1398 + 350 \sum_{j=0}^{n} \frac{1}{2^j}$$

Si se toma límite cuando $n \to \infty$ en la expresión se obtiene el valor de I_∞ como se indica a continuación:

$$I_\infty = \lim_{n \to \infty} I_n = \lim_{n \to \infty} \left(1398 + 350 \sum_{j=0}^{n} \frac{1}{2^j} \right) = 1398 + 350 \lim_{n \to \infty} \sum_{j=0}^{n} \frac{1}{2^j} = 1398 + 350 \sum_{j=0}^{\infty} \frac{1}{2^j} \quad (1)$$

Ahora, del apartado anterior se conoce que $S_x = \sum_{j=0}^{\infty} \frac{1}{2^j} = \frac{1}{1 - 0.5} = 2$.
Por tanto, sustituyendo el valor de S_x en la expresión (1) se obtiene que el valor de $I_\infty = \lim_{n \to \infty} I_n$ se corresponde con el año 2098.

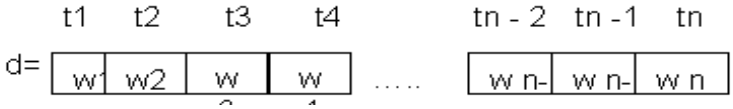
Teniendo presente la duración del proceso de globalización y dado que es demostrable el proceso de ralentización en el desarrollo de la globalización de la información y por tanto de la Traducción hay que considerar los sistemas de recuperación de información existentes para iniciar una mejora en los mismos que se adecúe a esta realidad informacional.

2. LOS MODELOS DE RECUPERACIÓN.

Los modelos de recuperación tienen como objetivo el facilitar el proceso de comparación entre una consulta determinada y un conjunto de textos sobre los que se realiza la consulta[1], para esto definen distintas formas de representar los documentos. Estos modelos de recuperación están pensados únicamente para documentos de contenido textual. Su

[1] Baeza-Yates, R. A., and Ribeiro-Neto, B. (2010). *Modern Information Retrieval* (2nd ed.). Reading, Massachusetts: Addison-Wesley; Croft, W. B., Metzler, D., and Strohman, T. (2010). *Search Engines: Information Retrieval in Practice*. London, England: Pearson; Grossman, D. A., and Frieder, O. (2004). *Information Retrieval: Algorithms and Heuristics* (2nd ed.). Berlin, Germany: Springer; Hearst, M. A. (2009). *Search User Interfaces*. Cambridge, England: Cambridge University Press; Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press; Manning, C. D., and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: MIT Press; Özsu, M. T., and Liu, L., editors (2009). *Encyclopedia of Database Systems*. Berlin, Germany: Springer; Salton, G. (1968). *Automatic Information Organziation and Retrieval*. New York: McGraw-Hill; van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). London, England: Butterworths; Witten, I. H., Moffat, A., and Bell, T. C. (1999). *Managing Gigabytes: Compressing and Indexing Documents and Images* (2nd ed.). San Francisco, California: Morgan Kaufmann; Zipf, G. K. (1949). *Human Behavior and the Principle of Least-Effort*. Cambridge, Massachusetts: Addison-Wesley.

funcionamiento es sencillo, para cada documento se construye un índice determinado en función del texto contenido en el documento. Derivado de esto tenemos el concepto de índice invertido que equivale a decir que la relación de los documentos en los que aparece una determinada palabra. Los índices de los documentos tienen en cuenta la frecuencia de aparición de las palabras. Cada documento se representa a través de un vector como los que se muestran a continuación:



Donde w1 indica la importancia del índice ti en el documento d. Suele tomar valor en el intervalo [0,1]. A las distintas formas de obtener el valor w1 se les denomina esquemas de asignación de pesos. Entre los distintos esquemas de asignación de pesos cabe distinguir:

Esquema Binario: Se asigna peso 1 (w1 = 1) si la palabra aparece en el documento y peso 0 (wi = 0) en caso contrario.

Frecuencia Inversa de Documento

(IDF): IDF = log2 N / Ni + 1 Wn =IDF i * Fn N = Nº total de documentos Ni = Nº de documentos en los que aparece el término i.

Fn = Frecuencia interna del término i en el documento j

Modelo espacio de vectores

En este modelo de recuperación de información cada documento se representa a través de un vector de n dimensiones cuyas componentes son los términos que aparecen en el texto. El valor de cada componente se calcula a partir del IDF (Inverse Document frequency) y se obtiene una representación vectorial para las consultas, que se comparan con los vectores de los documentos empleando una función de similitud. Para obtener la similitud de ente documento y consulta se puede utilizar algunas de las funciones siguientes:

Producto escalar:

$$SIM\ (t,q) = \sum_{i=1}^{n} t_i q_i$$

Distancia Euclídea:

$$SIM\ (t,q) = \sqrt{\sum_i |t_i - q_i|^2}$$

Fórmula del coseno:

$$SIM(t,q) = \frac{\sum_{i}^{n} t_i q_i}{\sqrt{\sum_{i=1}^{n} \left(t_i^2 q_i^2\right)}}$$

Entre las ventajas de este modelo de recuperación se encuentran:

Es posible obtener una lista ordenada de documentos que satisfacen la consulta.

Es posible controlar la respuesta ante una consulta, ya sea limitando el número de resultados o estableciendo un umbral de similitud. Como principal desventaja a destacar es que se supone que los términos de indexación son independientes.

Modelo probabilístico

El objetivo de este modelo de recuperación es determinar la probabilidad de que un documento determinado satisfaga la consulta del usuario, esta probabilidad depende Únicamente de la caracterización del documento y de la consulta. Para ello se trata de determinar aquellas propiedades que definen el conjunto de documentos relevantes. P(dj relevant to q)/ P(dj non-relevant to q)

· La consulta q es un subconjunto de términos de indexación.
· R = Conjunto de documentos relevantes
· R = Conjunto de documentos no relevantes
· P(R/d1) = Probabilidad de que el documento d1 sea relevante para la consulta q.

P(R/d) = Probabilidad de que el documento d1 NO sea relevante para la consulta q.

Frente a estos modelos generales nos encontramos con desarrollos más avanzados en los que es imprescindible seleccionar aquella información que es esencial. La Recuperación de Información (*Information Retrieval*, IR) trata con grandes colecciones de material textual, casi toda ella en formato http siendo su objetivo resolver sintaxis de búsqueda y entropía informativa de los usuarios[2].

La tradicional IR consiste en la búsqueda automática de todos los documentos relevantes en una colección de documentos independientemente de que estos estén estructurados o no, intentando

[2] K.L. Kwok, *A neural network for probabilistic information retrieval*. Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval. Cambridge, Massachusetts, United States. 1989.

además que el número de documentos no relevantes producidos en la solución de la sintaxis de búsqueda sea el mínimo posible. Los objetivos principales de IR y, con frecuencia su método, son el indexado de textos y la búsqueda de los documentos útiles de una colección, siempre que esta colección sea localizable. Hoy en día, dado el grave problema de multiplicación informativa, la recuperación de información abarca el modelado, la clasificación y la categorización de documentos, interfaces de usuario, visualización de datos, filtrado, análisis documental de contenido, etc.[3]. En definitiva, su objetivo es mejorar los resultados de recuperación en relación con mejores valores obtenidos en relación con la memoria y la precisión, siempre que la memoria (*recall)* sea el número de documentos relevantes recuperados dividido por el número de todos los documentos relevantes[4]:

$$\text{Memoria} = \frac{\textit{número de documentos relevantes recuperados}}{\textit{número total de documentos relevantes}}$$

Frente a la precisión[5] (*precision*) que es el número de documentos relevantes recuperados dividido por el número de todos los documentos recuperados[6]:

$$\text{Precisión} = \frac{\textit{número de documentos relevantes recuperados}}{\textit{número total de documentos recuperados}}$$

Es tradicional, por tanto, que desde hace tiempo los investigadores del campo de la recuperación de información se han dedicado al proceso de gestión de esta gran cantidad de información. Por lo que la recuperación se ha empleado, principalmente, para la clasificación y categorización de textos[7] construyendo modelos y propuestas como el Modelo de Espacio

[3] R. Baeza-Yates y B. Ribeiro-Neto, *Modern Information Retrieval*. ACM Press / Addison-Wesley. 1999.

[4] Ropero Rodríguez, J. *Método general de Extracción de información basado en el uso de Lógica Borrosa. Aplicación en portales web.* Universidad de Sevilla. Noviembre de 2009, T. D. p. 42.

[5] Ropero Rodríguez, J. *Método general de Extracción de información basado en el uso de Lógica Borrosa. Aplicación en portales web.* Universidad de Sevilla. Noviembre de 2009, T. D. p. 42.

[6] M.E. Ruiz y P. Srinivasan, *Automatic Text Categorization Using Neural Networks.Advances in Classification Research* vol. 8: Proceedings of the 8th ASIS SIG/CR Classification Research Workshop. Ed. Efthimis Efthimiadis. Information Today, Medford:New Jersey. 1998. pp 59-72.

[7] A.R. Aronson, T.C. Rindflesch y A.C. Browne, *Exploiting a large thesaurus for information retrieval*. Proceedings of RIAO, 197-216. 1994; S. Liu, M. Dong, H. Zhang, R. Li, Z. Shi, *An approach of multi-hierarchy text classification* Proceedings of the International Conferences on Info-tech and Info-net, 2001. Beijing. Vol. 3, pp. 95–100. 2001.

Vectorial (*Vector Space Model*, VSM), el método del vecino K más próximo (*K nearest neighbour method*, KNN), también denominado *k-means*, el modelo de clasificación Bayesiano, Redes Neuronales y las Máquinas de Soporte Vectorial (*Support Vector Machine*, SVM)[8]. De todos ellos el Vector Space Model es el modelo usado con más frecuencia por su simplicidad y su alta velocidad de procesado[9].

En este Modelo de Espacio Vectorial (*Vector Space Model*), el contenido (textual) de un documento está representado por un espacio multidimensional representado por un vector, es decir todos los conceptos son limitados por vectores que podemos valorar numéricamente. A posteriori, podemos diferenciar las clases correspondientes del vector designado analizando comparativamente estas distancias entre vectores. La metodología de procedimiento de este modelo puede ser dividido en[10]:

- 1º Indexación del documento, donde los términos/conceptos más relevantes son extraídos a través de un análisis de contenido del texto del documento.
- 2º Introducción y cálculo de pesos para los términos indexados, mejorando la velocidad y precisión en la búsqueda para la respuesta a la sintaxis de búsqueda comparándola con la relevancia prevista para el usuario.
- 3º Clasificación del documento en relación a la cuestión presentada según una medida de semejanza que reduzca la entropía informativa.

3. EL PROBLEMA LINGÜÍSTICO EN RECUPERACIÓN INFORMATIVA

Lo anteriormente expuesto demuestra que los modelos existentes necesitan, y obedecen a unos circuitos de reconocimiento previo de la información que los alejan de la velocidad de recuperación y de la eficacia en la recuperación que necesita el usuario especializado[11].

---

[8] M. Lu, K. Hu, Y. Wu, Y. Lu y L. Zhou, *SECTCS: towards improving VSM and Naïve Bayesian classifier*. IEEE International Conference on Systems, Man and Cybernetics, Vol. 5, p. 5. 2002.

[9] S. Liu, M. Dong, H. Zhang, R. Li, Z. Shi, *An approach of multi-hierarchy text classification* Proceedings <of the International Conferences on Info-tech and Info-net, 2001. Beijing. Vol. 3, pp. 95 – 100. 2001; M. Lu, K. Hu, Y. Wu, Y. Lu y L. Zhou, *SECTCS: towards improving VSM and Naïve Bayesian classifier*. IEEE International Conference on Systems, Man and Cybernetics, Vol. 5, p. 5. 2002; Y. Zhao y G. Karypis, *Improving precategorized collection retrieval by using supervised term weighting schemes*. Proceedings of the International Conference on Information Technology: Coding and Computing, 2002. pp 16 – 21.

[10] V.V. Raghavan y S. K. Wong, "A critical analysis of vector space model for information retrieval". *Journal of the American Society for Information Science*, Vol.37 (5), p. 279-87, 1986.

[11] Artandi, S. (1973). Information Concepts and Their Utility. *Journal of the American Society for Information Science, 24*, 242-245.; Baird, J. C. (1984). Information Theory and Information Processing. *Information Processing & Management, 20* (3), 373-381; Belkin, N. J. (1975). Some Soviet Concepts of Information for Information Science. *Journal of the American Society for Information Science, 26* , 56-64; Belkin, N. J. (1978). Information Concepts for Information Science. *Journal of Documentation, 34*, 55-85.; Beniger, J. R. (1988). Information and Communication: The New Convergence. *Communication Research, 15* (2), 198-218; Boulding,

K. E. (1961). Image: Knowledge of Life and Society, *The Image: Knowledge of Life and Society* (pp. 3-18). Ann Arbor, Mich.: University of Michigan Press; Brillouin, L. (1965). Observation, Information, and Imagination. In S. Dockx & P. Bernays (Eds.), *Information and Prediction in Science.* New York: Academic Press; Buckland, M. K. (1991). Information as Thing. *American Society for Information Science, Journal, 42* (5), 351-360; Derr, R. L. (1985). Concept of Information in Ordinary Discourse. *Information Processing & Management, 21* (6), 489-499; Diener, R. A. V. (1981, October 25-30, 1981). *Relational Analysis: On the Representation and Analysis of Knowledge.* Paper presented at the Information Community: An Alliance for Progress. Proceedings of the 44th ASIS Annual Meeting, Washington, DC; Dolby, J. L. (1977). On the Notions of Ambiguity and Information Loss. *Information Processing and Management, 13* (1), 69-77; Dolby, J. L. (1984). Data as Information. *Information Processing & Management, 20* (3), 407-415; Fairthorne, R. A. (1967). Morphology of 'Information Flow'. *Journal of the Association for Computing Machinery, 14*, 710-719; Farradane, J. (1979). Nature of Information. *Journal of Information Science, 1*, 13-17; Gordon, D. B., & Sager, N. (1985). Method of Measuring Information in Language, Applied to Medical Texts. *Information Processing & Management, 21* (4), 269-289; Green, R. (1991). Profession's Models of Information: A Cognitive Linguistic Analysis. *Journal of Documentation, 47* (2), 130-148; Green, R. T., & Courtis, M. C. (1966). Information Theory and Figure Perception: The Metaphor that Failed. *Acta Psychologica, 25*, 12-36; Hammarberg, R. (1981). Cooked and the Raw. *Journal of Information Science, 3*, 261-267; Hayes, R. M. (1991). Measurement of Information and Communication [MS] (pp. C1-C22); Hoffman, E. (1980). Defining Information: An Analysis of the Information Content of Documents. *Information Processing and Management, 16* , 291-304; Kasanof, R. (1968). Right to Lie? *Center Magazine, I*, 42-43; Kochen, M. (1984). Coding for Recording and Recall of Information. *Information Processing & Management, 20* (3), 343-354; Kolata, G. b. (1978). Information Theory: a Surprising Proof. *Science, 199 (Jan. 6*, 42; Kreither, H., & Kreither, S. (1976), *Cognitive Orientation and Behavior*. New York: Springer; Leide, J. E. (1981). Emmorphosis: Information as Process. *Canadian Journal of Information Science, 6*, 35-37; Leupolt, M. (1978). Some Considerations on the Nature of Information. *International Forum on Information and Documentation, 3* (3), 29-34; Levi, I. (1984). Information and Ignorance. *Information Processing & Management, 20* (3), 355-362; Levine, M. M. (1977). Informative Act and Its Aftermath: Toward a Predictive Science of Information. *Journal of the American Society for Information Science, 28*, 101-106; Lynch, M. F. (1977). Variety Generation -A Reinterpretation of Shannon's Mathematical Theory of Communication, and its Implications for Information Science. *Journal of the American Society for Information Sicence, 24*, 19-25; Mackay, D. M. (1965). Information and Prediction in Human Sciences. In S. Dockyx & P. Bernays (Eds.), *Information and Prediction in Science.* New York: Academic Press; McCarthy, J. (1966). Information. *Scientific American, 215*, 64-73; Meadow, C. T., & Yuan, W. (1997). Measuring the Impact of Information: Defining the Concepts. *Information Processing and Management, 33* (6), 697-714; Molholt, P. (1984). Nature of Information and Its Influence on Libraries. *Special Libraries, 75* (3), 347-251; Pearson, C., & Slamecka. (1979). Semiotics Foundations of Information Science, I. Theory of Sign Structure. *Foundations, FIS-2*, 7-19; Pérez, A., & Tondl, L. (1965). On the Role of Information Theory in Certain Scientific Procedures. In S. Dockx & P. Bernays (Eds.), *Information and Prediction in Science*. New York: Academic Press; Pratt, A. D. (1977). Information of the Image: A Model of the Communications Process. *Libri, 27*, 204-220; Pratt, A. D. Information of the Image [MS] (pp. 204-220); Roberts, N. (1976). Social Considerations Towards a Definition of Information Science. *Journal of Documentation, 32*, 249-257; Rudd, D. (1983). Do We Really Need World III? Information Science With or Without Popper. *Journal of Information Science, 8*, 99-105; Schreider, Y. A. (1965). On the Semantic Charateristics of Information. *Information Storage and Retrieval, 2*, 221-233; Shannon, C. E., & Weaver, W. (1975). Mathematical Theory of Communication, *Mathematical Theory of Communication* (pp. 3-19). Urbana, Ill.: University of Illinois Press;

Por ello una de las soluciones utilizadas es el empleo de técnicas de Tratamiento del Lenguaje Natural (*Natural Language Processing*, NLP) se pueden usar para la Recuperación de Información (IR) de varias formas. el objetivo principal de aplicar el tratamiento del Lenguaje Natural a IR es alcanzar una mejora en los resultados de recuperación en relación con mejores valores obtenidos para la memoria (*recall*) y la precisión (*precision*). Desde el punto de vista histórico, las técnicas NLP centradas en el Indexado Motivado Lingüísticamente (*Linguistically Motivated Indexing*, LMI) han sido el foco principal de investigación; LMI ha sido diseñado y evaluado en contraste con el Indexado No Lingüístico
(*Non-Linguistic Indexing*, NLI) [SPARCK-JONES99]. Los sistemas LMI utilizan técnicas lingüísticas, usando la semántica y la sintaxis para identificar términos, encontrar unidades compuestas de varias palabras o caracterizar la estructura interna de una frase o documento. Por otra parte, los sistemas NLI no utilizan ninguna de estas técnicas y se limitan a aplicar criterios estadísticos, utilizando las denominadas *stop words*.

Hay dos aproximaciones diferentes para integrar técnicas NLP y recursos en la IR:

El Indexado Motivado Lingüísticamente (LMI) se usa para crear términos índice para un modelo de espacio vectorial o para sistemas de búsqueda booleana. En el primer caso, los documentos o búsquedas son convertidas en vectores con un cierto peso, devolviéndose aquellos vectores similares a los de la consulta; el segundo caso se basa en el álgebra de Boole.

Sistemas basados en Inteligencia Artificial (*Artificial Intelligence*, AI), que tratan de emparejar una consulta con las representaciones semánticas de los textos de entrada. Además de para NLP, otros usos de LMI se encuentran por ejemplo en la traducción automática (*Machine Translation*, MT), complementando a la traducción basada en reglas, a la traducción basada en estadística y a la traducción basada en ejemplos. Para ello es necesario identificar los problemas de variación (de la variación que supone la existencia de formas diferentes de decir lo mismo) y que se pueden

Shannon, C. H., & Weaver, W. (1962). Introductory Note on the General Setting of the Analytical Communication Studies, *Mathematical Theory of Communication* (pp. 3-18). Urbana, IL: Univ. of Illinois Press; Shaw, D., & Davis, C. H. (1983). Entropy and Information: A Multidisciplinary Overview. *Journal of the American Society for Information Science (JASIS), 34* (1), 67-74; Tenner, E. (1991). Impending Information Implosion. *Harvard Magazine*, 31-34; Thompson, F. B. (1968). Organization is the Information. *American Documentation, 19*, 305-308; Weaver, W. Recent Contributions to the Mathematical Theory of Communication. In C. E. Shannon (Ed.), *Mathematical Theory of Communication*; Wersig, G., & Neveling, U. (1975). Phenomena of Interest to Information Science. *The Information Scientist, 9* (4), 127-140; Wyatt, H. V. (1972). When does information become knowledge? *Nature, 235*, 86-89; Ziman, J. M. (1969). Information, Communication, Knowledge. *Nature, 224*, 318-324.

suponer fruto de factores extralingüísticos, factores como la geografía, la historia, la situación comunicativa, en su sentido más amplio, o determinados elementos sociales. Pero no siempre es así. Por eso se ha dicho[12] que los factores que determinan la aparición de unas variantes lingüísticas en ciertas circunstancias y de otras variantes en circunstancias diferentes, dentro de una comunidad de habla, pueden responder a estas cuatro posibilidades:

a. que las variantes vengan determinadas exclusivamente por factores lingüísticos;
b. que las variantes vengan determinadas exclusivamente por factores extralingüísticos;
c. que las variantes vengan determinadas conjuntamente por factores lingüísticos y extralingüísticos;
d. que las variantes no vengan determinadas por factores lingüísticos ni por factores extralingüísticos.

En el último caso, lo más probable no es la ausencia de determinación por parte de esos tipos de factores, sino tal vez la falta de respuestas o de explicaciones por parte de los especialistas. Si algo no viene determinado aparentemente ni por unos factores ni por otros, es más razonable pensar en la incapacidad de los estudiosos que en lo inexplicable de la lengua.

Considerando todo lo anterior, es preciso insistir en el hecho de que la variación, definida como el uso alterno de formas diferentes de decir lo mismo, se puede encontrar prácticamente en todos los niveles de la lengua, desde el más concreto (fonético-fonológico) al más amplio (discurso, por ejemplo), pasando por la gramática y el léxico. Para explicar el funcionamiento de estos usos, se puede prestar atención, separada o conjuntamente, a la forma en que ejercen su influencia los factores lingüísticos y a la forma en que lo hacen los factores extralingüísticos, esto es, los factores históricos, geográficos, contextuales y sociales. Entre esos factores sociales está naturalmente la profesión u ocupación de los hablantes. Este factor es el que principalmente nos interesa para la elaboración de funciones y algoritmos de recuperación especializados para traductores, pues la traducción especializada se alimenta de textos con jerga profesional, altamente especializada y jerarquizada, cuya comunicación puramente profesional no ha de tener una intención o un carácter críptico, por más que su dominio corresponde normalmente a

[12] H. Cedergren, (1993): *Sociolingüística*, en H. López Morales (coord.), *Introducción a la lingüística actual*, Madrid, Playor, 1983, p. 150. H. López Morales, *Sociolingüística*, 2ª ed., Madrid, Gredos, pp. 84-85.

individuos iniciados[13]. Nos encontramos ante variedades sectoriales, especializadas[14],de grupo o tecnolectos, con diferente grado de hermetismo, que pueden ser de muchos tipos: aquí se incluiría la jerga médica y de la enfermería, la economista y empresarial, la jurídica, la militar, la periodística, la informática y multitud de jergas de oficios, que en ocasiones han gozado de una larga tradición. En conexión con estas variedades sectoriales, estarían también los lenguajes científico - técnicos, formados principalmente por nomenclaturas específicas[15].

[13] Bates, M. J. (1976). Rigorous Systematic Bibliography. *RQ, 16*, 7-26; Bates, M. J. (1986). What Is a Reference Book? A Theoretical and Empirical Analysis. *RQ, 26*, 37-57; Bates, M. J. (1997). Exploratory Profile of Personal Home Pages: Content, Design, Metaphors. *Online & CDROM Review, 21*, 331-340; Buckland, M. K. (1997). What Is a 'Document'? *Journal of the American Society for Information Science, 48* (9), 804-809; Burnett, K. (1993). Toward a Theory of Hypertextual Design. *Postmodern Culture, 3* (2); Cheney, G., & Tompkins, P. K. (1988). On the Facts of the Text as the Basis of Human Communications Research. *Communications Yearbook, 11*, 455-481; Clapp, V. W. (1971). Greatest Invention since the Title-Page? Auto-bibliography from Incipit to Cataloging-in-Publication. *Wilson Library Bulletin, 46*, 348-359; Dillon, A. (1994). *Designing usable electronic text: ergonomic aspects of human information usage*. London: Taylor & Francis; Ellis, D., Ford, N., & Wood, F. (1993). Hypertext and Learning Styles. *Electronic Library, 11* (1), 13-18; Garfield, E., ed. (1976). Is there a future for the scientific journal? In E. Garfield (Ed.), *Essays of an information scientist* (Vol. 2, pp. 318-322); Golovchinsky, G., & Chignell, M. H. (1997). The Newspaper as an Information Exploration Metaphor. *Information Processing and Management, 33* (5), 663-683; Gordon, D. R. (1975). Print as a Visual Medium. *Library Quarterly, 45* (1), 34-45; Keyhani, A. (1993). Online Journal of Current Clinical Trials: An Innovation in Electronic Journal Publishing. *Database, 176* (1), 14-23; Lacy, D. (1982). Culture and the Media of Communication. *Scholarly Publishing, 13* (3), 195-210; Rada, R. (1991). Focus on Links: A Holistic View of Hypertext. *International Classification, 18* (1), 13-18; Rada, R. (1991). Small, Medium, and Large Hypertext. *Information Processing & Management, 27* (6), 659-677; Renwick, H. L., & Cutter, S. (1983). Map Postcards and Images of Place. *Landscape, 27*, 30-38; Rowley, J. E. (1983). Future for printed indexes? *Aslib proceedings, 35* (4), 234-238; Ryan, B. (1991). Dynabook Revisited with Alan Kay. *Byte*, 203-208; Shackel, B., Pullinger, D. J., & Maud, T. I. (1983); Spencer, H., Reynolds, L., & Coe, B. (1975). Spatial and Typographic Coding in Printed Bilbiographic Materials. *Journal of Documentation, 31*, 59-70; Taylor, R. S. (1984). Value-Added Processes in Document-Based Systems: Abstracting and Indexing Services. *Information Services & Use, 4* (3), 127-146; Veltman, K. (1991). Computers and a New Philosophy of Knowledge. *International Classification, 18* (1), 2-12; White, H. D., Bates, M. J., & Wilson, P. (1992). *For Information Specialists: Interpretations of Reference and Bibliographic Work*. Norwood, New Jersey: Ablex

[14] Denominación dada por B. Rodríguez en *Las lenguas especiales. El léxico del ciclismo*, León, Colegio Universitario de León, 1981, pp. 9-153. Véase también B. Rodríguez, "Argot y lenguaje coloquial", en A. Briz, J.R. Gómez Molina, M.J. Martínez y Grupo Val.Es.Co. (eds.), *Pragmática y gramática del español hablado*, Valencia, Universidad de Valencia-Pórtico, 1997, pp. 225-239.

[15] Véase B. Rodríguez, "Lo específico de los lenguajes científico-técnico", *Archivum*, XXVII-XXVIII (1977-1978), pp. 485-521. También M.T. Cabré, *La terminología. Teoría, metodología, aplicaciones*, Barcelona, Antártida, 1993.

Y aquí habría que decir que no todos los lenguajes especializados comparten plenamente unos rasgos lingüísticos, aunque sí es posible fijar como caracteres comunes los siguientes[16].

1) En primer lugar, son variedades especializadas aquellas que sirven como instrumento de comunicación formal y funcional entre especialistas en una determinada materia.

2) Desde un punto de vista lingüístico, los lenguajes de especialidad se caracterizan por utilizar, en términos generales, la gramática de la lengua común, matizada por ciertos usos que pueden destacar cualitativa o cuantitativamente: así, es frecuente que aparezcan formas de tratamiento específicas o habituales en ciertos ámbitos profesionales (pensemos, para el ámbito jurídico, en formas como señoría o letrado o, para el ámbito académico, en formas como profesor o doctor), como frecuente es el uso específico de ciertas formas verbales (pensemos en el futuro de subjuntivo o en el gerundio, en la lengua jurídico-administrativo: si no compareciere, alegación solicitando revisión de pruebas) y como frecuente puede ser el uso de procedimientos específicos de formación de palabras.

3) Desde un punto de vista estilístico, los lenguajes de especialidad se caracterizan por ser utilizados en contextos formales, por lo que se ven favorecidos los rasgos que expresan una mayor impersonalidad y una menor implicación afectiva.

4) Desde un punto de vista comunicativo, las variedades de especialidad se caracterizan por subordinar lo estético y lo expresivo a lo objetivo y a la eficacia comunicativa. De esta forma, se ven favorecidos los usos lingüísticos capaces de expresar orden, claridad, concisión.

5) Teniendo en cuenta el modo del discurso –entendiendo "modo" según lo hace M.A.K. Halliday al hablar del registro–, los tecnolectos –digamos que en una buena parte de ellos– se caracterizan por conceder un lugar preeminente al discurso escrito. En la práctica del lenguaje jurídico-administrativo es importante conocer la forma de los decretos, las instancias, las demandas, las actas, los oficios, los certificados o los acuerdos; en la práctica del lenguaje de los negocios se necesita conocer la forma de las cartas comerciales o de los pedidos; en la práctica de la sanidad se debe conocer la forma de los historiales clínicos.

A esta dificultad general, la de fijar los límites entre lo específico y lo común, se puede añadir la derivada de los tipos tan diferentes que existen entre los lenguajes de especialidad. Es ya conocida una clasificación según el grado de abstracción del lenguaje, la artificiosidad, la sintaxis y los participantes en la comunicación especializada, que lleva a distinguir hasta

[16] Moreno Hernández, F. (1999): *Lenguas de especialidad y variación lingüística,* en S. Barrueco, E. Hernández y L. Sierra (eds.*), Lenguas para fines específicos (VI). Investigación y enseñanza*, Alcalá de Henares, Universidad de Alcalá, 1999, pp. 3-14.

cuatro clases de variedades de especialidad: lenguajes profesionales, lenguajes técnicos, lenguajes científicos y lenguajes simbólicos. Considerados en ese orden, los lenguajes profesionales serían los que disfrutarían de un menor grado de abstracción, de una menor artificiosidad y de una sintaxis más libre; los lenguajes más abstractos y pre-determinados serían los simbólicos[17].

Sin embargo, lo que más nos interesa para lograr una efectividad en la recuperación de conocimiento especializado para la traducción son las características del léxico de las lenguas de especialidad, muy especialmente con el léxico del lenguaje científico-técnico. En las unidades léxicas de este lenguaje se produce una circunstancia poco frecuente en el léxico común o general: el significante y el significado de estos signos establecen una relación unívoca que impide la polisemia o la connotación. En el libro de Martín, Ruiz, Santaella y Escánez, titulado *Los lenguajes especiales*, se afirma a este respecto, con toda razón:

> La denotación expresa el significado de las palabras, sin mezcla de nota cualitativa alguna. La significación de los vocablos científicos es denotativa. Estos lenguajes, que por definición son unívocos y objetivos, evitan las equivalencias laterales de valor estilístico y expresivo: protozoo tiene un significado único e invariable en cualquier contexto en que se integre. (...) El vocabulario científico no se puede ver modificado por el contexto, ni intra ni extratextualmente, pues supondría, además, atentar contra la coherencia que debe mantener todo texto científico a lo largo de su trayectoria. De acuerdo con lo que estamos exponiendo, este léxico, a diferencia del léxico común, es un vocabulario inmóvil, sin posibilidad de traslado de su significado por

[17] Abel, M. J. (1990). Experiences in an Exploratory Distributed Organization. In e. a. Jolene Galegher (Ed.), *Intellectual Teamwork: Social and Technological Foundations of Cooperative Work* (pp. 489-510). Hillsdale, New Jersey: Lawrence Erlbaum; Agre, P. E. (1997). Designing Genres for New Media: Social, Economic, and Political Contexts (pp. 131-160); Buckland, M. (1987). Combining Electronic Mail with Online Retrieval in a Library Context. *Information Technology & Libraries, 6*, 266-271; Budd, J. M., & Connaway, L. S. (1997). University Faculty and Networked Information: Results of a Survey. *Journal of the American Society for Information Science (JASIS), 48*, 1-10; Doty, P., Bishop, A. P., & McClure, C. R. (1991). *Scientific Norms and the Use of Electronic Research Networks.* Paper presented at the ASIS '91: Proceedings of the 54th Annual Meeting of the American Society for Information Science, Washington, DC; Hill, W., et al. (1995, May 7-11, 1995). *Recommending and Evaluating Choices in a Virtual Community of Use.* Paper presented at the Human Factors in Computer Systems CHI '95 Conference Proceedings, Denver, Colorado; Rosenbaum, H., & Newby, G. B. (1990, November 4-8, 1990). *Emerging Form of Human Communication: Computer Networking.* Paper presented at the ASIS '90: Proceedings of the 53rd Annual ASIS Meeting, Toronto, Ontario; Schramm, W. (1957). Twenty Years of Journalism Research. *Public Opinion Quarterly, 21*, 91-107; Shardanand, U., & Maes, P. (1995, May 7-11, 1995). *Social Information Filtering: Algorithms for Automating 'Word of Mouth'.* Paper presented at the Human Factors in Computer Systems CHI '95 Conference Proceedings, Denver, Colorado.

motivos afectivos: diuresis no tiene la misma capacidad de transformación que posee la voz azul.

Con todo esto podemos afirmar que la elección de lenguajes altamente especializados como el científico-técnico es el más adecuado para las pruebas de algoritmos de recuperación fundamentados en lógica difusa. Logrando así un proceso automático de agrupamientos semánticos.

4. DESARROLLO DEL PROYECTO DE RECUPERACIÓN.

En 1999 Sierra y McNaught[18] propusieron un algoritmo para la generación automática de agrupamientos semánticos basado en analogías[19]. El algoritmo, se aplicó originalmente sobre un diccionario terminológico en el área de metrología en el idioma inglés. Del análisis de los resultados obtenidos por el algoritmo básico de alineamiento semántico y de un riguroso estudio sobre el algoritmo se han identificado una serie de opciones que derivan en un conjunto de alternativas para mejorar el número de pares-semánticos reconocidos en el algoritmo[20]. Las observaciones dan

[18] Sierra G. (1999). Design of a concept-oriented tool for terminology. PhD Thesis University of Manchester, Institute of Science and Technology; Sierra G. & McNaught J., (2000) "Design of an onomasiological search system: A concept-oriented tool for terminology". *Terminology*. Vol. 6 (1).

[19] Amir A., Auman Y., Landau G., Lewenstein M., & Lewenstein N., (1997). "Pattern matching with swaps". In Proc. FOCS'97. pp. 144-153; Bellman R., Dreyfus S. (1962) Applied Dynamic Programming. Princeton University Press, Princeton NJ; Baeza-Yates R., Ribeiro-Neto B. (1999) Modern Information Retrieval. ACM press, Addison Wesley. Frakes W. B., (1992). "Stemming algorithms". In Information retrieval: Data Structures & Algorithms. W.B. Frakes and R.Baeza-Yates (eds.). New Jersey. Prentice Hall; Sierra G. & McNaught J., (2000), "Extracting semantic clusters from MRD for an onomasiological search dictionary". International Journal of Lexicography. Vol. 13 (4) Ukkonen, E. (1985). Algorithms for approximate string matching. Information and control, Vol. 64. Pp. 100-118; Velichko V. M., Zagoruyko N. G. (1970). Automatic Recognition of 200 words . International Journal of Man-Machine Studies, Vol. 2. pp. 223-234; Vintsyuk T.K. (1968) Speech discrimination by dynamic programming. Cybernetics, Vol 4 (1). pp. 52-57; Wagner R. A., Fisher M. J. (1974). The string-to-string correction problem. Journal of the ACM, Vol. 21(1), pp. 168-173.

[20] Arnovick, G. N., & Gee, L. G. (1978). Design and Evaluation of Information Systems. *Information Processing & Management, 14*, 369-280; Belkin, N. J., Oddy, R. N., & Brooks, H. M. (1982). ASK for Information Retrieval: Part II. Results of a Design Study. *Journal of Documentation, 38*, 145-164; Blair, D. C., & Maron, M. E. (1985). Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System. *Communications of the ACM, 28* (3), 289-299; Blair, D. C., & Maron, M. E. (1990). Full-Text Information Retrieval: Further Analysis and Clarification. *Information Processing & Management, 26* (3), 437-447; Bonzi, S., & Liddy, E. (1989). Use of Anaphoric Resolution For Document Description in Information Retrieval. *Information Processing & Management, 25* (4), 429-441; Brajnik, G., Guida, G., & Tasso, C. (1987). User Modeling in Intelligent Information Retrieval. *Information Processing & Management, 23* (4), 305-320; Buckland, M. K., & Larson, R. R. (1992). *Entry Vocabulary and Cross-Database Searching: Proposal to the Database Activities Program, Division of Instrumentation and Resources, NSF* (grant proposal): Univ. of California, Berkley; Burket, T.

pie a líneas de trabajo, divididas en dos grandes grupos: Heurísticas alternativas y modificaciones a la interfaz hombre-computadora de las cuales se ha trabajado en las siguientes líneas: pares semi-iguales y semi-nulos, intercambio de palabras, rutas múltiples de alineamiento. Como resultado, se desarrollo el algoritmo de alineamiento semántico múltiple La expectativa respecto al desempeño de las diferentes alternativas desarrolladas, indicaba que el algoritmo de rutas múltiples junto con modificación de costos y pares semi-iguales y semi-nulos ofrecerían los mejores resultados posibles. Con el fin de establecer la certeza de las observaciones cualitativas, se evaluó el desempeño de las diferentes alternativas a través del método *recall* y *precision*. Además, como una

G., Emrath, P., & Kuck, D. J. (1971). Use of Vocabulary Files for On-line Information Retrieval. *Information Processing and Retrieval, 15* (6), 281-289; Cleverdon, C. (1967). Cranfield Tests on Index Language Devices. *Aslib Proceedings, 19*, 173-194; Cleverdon, C. (1970). Evaluation Tests of Information Retrieval Systems. *Journal of Documentation, 26*, 55-67; Cleverdon, C. Conclusions.*, N78-18979. no ref. available. Excerpt, p.53*; Cleverdon, C. W., & Mills, J. (1963). Testing of Index Language Devices, *Aslib Proceedings* (Vol. 15, pp. 106-130); Cooper, W. S. (1971). Definition of Relevance for Information Retrieval. *Information Storage and Retrieval, 7*, 19-37; Cuadra, C. A., & Katter, R. V. (1967). Opening the Black Box of 'Relevance'. *Journal of Documentation, 23* (4), 291-303; Curtice, R., & Jones, P. E. (1967). *Distributional Constraints and the Automatic Selection of an Indexing Vocabulary.* Paper presented at the Proceedings of the 30th Annual Meeting of the American documentation Institute, New York, October 1967; Dillon, M. (1974). Experiment in Superficial Indexing. *Information Storage and Retrieval, 10*, 63-71; Dumais, S. T. (1988). Textual Information Retrieval. In M. Helander (Ed.), *Handbook of Human-Computer Interaction* (pp. 673-699). North-Holland: Elsevier Science Publishers; Efthimiadis, E. N. (1995). User Choices: A New Yardstick for the Evaluation of Ranking Algorithms for Interactive Query Expansion. *Information Processing & Managment, 31* (4), 605-620; Eisenberg, M., & Schamber, L. (1988). *Relevance: The Search for a Definition.* Paper presented at the AISIS '88: Information Technology: Planning for the Next Fifty Years. Proceedings of the 51st Annual Meeting of the American Society for Information Science; Ellis, D. (1984). Effectiveness of Information Retrieval Systems: The Need for Improved Explanatory Frameworks. *Social Science Information Studies, 4* , 261-272; Ellis, D. (1984). Theory and Explanation in Information Retrieval Research. *Journal of Information Science, 8* (1), 25-38; Froehlich, T. J. (1988, August 28-September 1, 1988). *Relevance and the Relevance of Social Epistemology.* Paper presented at the Information*Knowledge*Evolution: Proceedings of the 44th FID Congress, Helsinki; Gluck, M. (1995). Understanding Performance in Information Systems: Blending Relevance and Competence. *Journal of the American Society for Information Science, 46* (6), 446-460; Harman, D. (1995). Overview of the Second Text Retrieval Conference (TREC-2). *Information Processing & Management, 31* (3), 271-289; Harmon, G. (1970). *Information Need Transformation During Inquiry: A Reinterpretation of User Relevance.* Paper presented at the The Information Concious Society. Proceedings of the ASIS, 33rd Annual Mtg, Philadelphia, PA, Oct 11-15; Hendry, I. G., Willet, P., & Wood, F. E. (1986). INSTRUCT: A Teaching Package for Experimental Methods in Information Retrieval. Part I. The Users' View. *Program, 20* (3), 245-263; Hendry, I. G., Willet, P., & Wood, F. E. (1986). INSTRUCT: A Teaching Package for Experimental Methods in Information Retrieval. Part II. Computational Aspects. *Program, 20* (4), 382-393; Hillman, D. J. (1964). Notion of Relevance (I). *American Documentation, 15*, 26-34; Janes, J. W. (1991). Relevance Judgements and the Incremental Presentation of Document Representations. *Information Processing & Management, 27* (6), 629-646

medida de comparación entre las diferentes alternativas se propone el índice de identificación de pares-vinculados y el índice de recuperación de pares-semánticos.

Hemos empleado la Teoría de Conjuntos Difusos introducida por L. Zadeh (1965), con el fin de suministrar un esquema para dar tratamiento matemático adecuado a un sinnúmero de problemas, en los que juega un papel fundamental cierta imprecisión o vaguedad que procede de una especie de ambigüedad intrínseca en los textos y así evitar las variaciones semánticas con los que los pares sean lo más exhaustivos posible en su uso correcto en equivalencias lingüísticas LP-LM.

El empleo de modelos ideales, no realizables en la práctica, ha dado buenos resultados en algunos campos[21]; sin embargo, una aproximación que en muchos problemas resulta más realista, es la de considerar conjuntos difusos, de forma que se asocien propiedades a cierto universo de objetos (pares, conceptos lingüísticos, conceptos científicos de especialidad, etc.) sin imponer acotaciones severas sobre los sistemas que estamos estudiando[22].

Nuestro fin en el algoritmo múltiple desarrollado es la introducción de una familia parametrizada de medidas de borrosidad, construida a partir de una familia de índices de "afinidad" entre dos distribuciones probabilísticas que se conviertan en espejo de conceptos definibles a textos y marcados como índices o pesos semánticos. Sea X un conjunto de elementos (referencial). Un subconjunto difuso de X se define intuitivamente a través de una propiedad, que los elementos de X no necesariamente tienen que satisfacer o no, sino que pueden cumplir con "cierto grado"[23]. De

[21] Geckeler, H (1976). *Semántica estructural*, Madrid, Gredos; Haton J. P. (1973). Contribution à l'Analyse, Paramétrisation et la Reconnaissance Automatique de la Parole, Thèse de doctorat d'état, Université de Nancy, France; Hirschberg, D. S. (1975). A linear Space algorithm for computing maximal common subsequences. Communication of the ACM. Vol 18(6). pp. 341-343; Lee J., Kim D., Park K., Cho Y. (1997). Efficient algorithms for approximate string matching with swaps. In Proc. CPM'97. LNCS 1264, Springer-Verlag. Pp. 28-39; Levenshtein, V.I. (1965). Binary codes capable of correcting spurious insertions and deletions of ones. Problems of information Transmission, (1). pp. 8-17; Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics - Doklady, 10 (8), pp. 707-710; Lowrence R., Wagner R. A., (1975). An extension of the string-to-string correction problem. Journal of ACM, Vol. 22 (2), pp 177-183.

[22] Needleman S. B., Wunsch C. D. (1970). A general method applicable to search for similarities in amino-acid sequence of two proteins. *Journal of Molecular Biology*, Vol. 48. pp 443-453; Porter, M.F. (1980). An algorithm for suffix stripping. Program, 14(3). pp 130-137; Reitchert T. A., Cohen D. N., Wong A. K. C. (1973). An application of information theory to genetic mutations and matching of polypeptide sequences. *Journal of Theoretical Biology*, Vol 42. pp. 245-261.

[23] De Kerf, J. (1975). A bibliography on fuzzy sets. J. Comput. Appl. Math, 1:205--212; Díaz Nafría, J.M. (2008). "Indeterminación de la observación", en Díaz y Salto (eds.) *¿Qué es información?* León: Universidad de León, pp. 489-502; Kandel, A. and Byatt, W.J. (1978). Fuzzy

este modo: el concepto de pertenencia o no de un concepto especializado a un conjunto A puede expresarse numéricamente mediante la función de pertenencia, también llamada a veces función característica. Esta función asigna a cada elemento x del universo de discurso un dígito binario (1 ó 0) según x pertenezca o no al conjunto A (independientemente de la lengua meta del traductor)

$$\mu_A : \mathbf{X} \rightarrow \{0,1\} \mid \quad \mu_A(x) = \begin{cases} 1 & \text{cuando } x \in A \\ 0 & \text{cuando } x \notin A \end{cases}$$

cualquier conjunto A ⊂ X se puede definir por los pares semánticos que forman cada elemento x del universo y su función de pertenencia, expresándose de a la siguiente forma:

$$A = \{(x, \mu_A(x)) \forall x \in \mathbf{X}\}$$

En la teoría de conjuntos borrosos, los conjuntos clásicos se denominan conjuntos *crisp*, con el fin de distinguirlos de los conjuntos borrosos[24]. Sea A un conjunto clásico definido en el universo X, entonces para cualquier elemento x dentro de X, x ∈ A o x ∉ A. En la teoría de conjuntos borrosos esta propiedad está generalizada, por lo tanto, en un conjunto borroso A (texto referencial, conjuntos de textos referenciales, fuentes de consulta, texto en lengua de partida, etc.), no es necesario que x ∈ A o x ∉ A.

---

sets, fuzzy algebra and fuzzy statistics. *Proc. IEEE, pp.* 1619-1639; Kandel, A. and Davis, H.A. (1976). The first fuzzy decade. (A bibliography on fuzzy sets and their applications). Comput. Sci. Dep. New Mexico Inst. Min. Technol, Socorro, CSR-140, 1976; Menger, K. (1951). Ensembles flous et foncions aleatoires. *C. R. Acad. Sci., (232)*, 2001-2003; Negoita, C.V. and Ralescu, D.A. (1975). *Applications of Fuzzy Sets to System Analysis*, Chaps. 1 and 2. Basel: Birkhaeuser; Pawlak, Z. (1985). Rough sets and fuzzy sets. *Fuzzy Sets and Systems, 17*, 99-102; Pérez-Amat, (2008). "Hacia una teoría semántica de la información", en Díaz y Salto (eds.) *¿Qué es información?* León: Universidad de León, pp. 51-72.

[24] Bellman, R.E., Kalaba, R., and Zadeh, L.A. (1964). Abstraction and pattern classification. RAND Memo, RM-4307-PR. [Online] RAND access [retrived: 27/02/2010]; Chang, S.S.L. (1972). Fuzzy mathematics, man and his environment. *IEEE Trans. Sys. Man Cybern., 2*, 92-93; Driankov, D., Hellendoorn, H., and Reinfrank, M.(1993). *An Introduction to fuzzy control*. Berlin: Springer Verlag; Dubois, D. and Prade, H. (1987). Twofold fuzzy sets and rough sets-some issues in knowledge representation. *Fuzzy Sets and Systems, 23*, 3-18; Gaines, B.R. and Kohout, L.J. (1977). The fuzzy decade: A bibliography on fuzzy systems and closely related topics. *Int J. Man-Math. Stud., 9*, 1-69 (también en Gupta et al. (1977), pp. 403-490); Gale, S. (1975). Boundaries, tolerance spaces and criteria for conflict resolution. *J. Peace Sci., 1(2)*, 95-115; Gupta, M.M., Saridis, G.N. and Gaines, B.R. (1977). *Fuzzy automata and Decision Processes*. Amsterdam: North-Holland Publ; Gusev, L.A. and Smirnova, I.M. (1973). Fuzzy sets: Theory and applications (a survey). *Autom. Remote Control(USSR), 6(5)*, 66-85; Kaufmann, A. (1980). Bibliography on fuzzy sets and their applications. BUSEFAL (LSI Lab, Univ. Paul Sabatier, Toulouse, France), (1-3).

En los últimos años se han propuesto varias definiciones que presentan la generalización de la propiedad de pertenencia (Dubios 1987), (Pawlak 1985), (Shafer 1976), pero parece que la teoría de conjuntos borrosos es la más intuitiva entre el resto de teorías y teoremas existentes. La generalización se realiza como sigue por lo que el algoritmo identifica los pares semánticos dudosos como correctos agrupándolos semánticamente en sus conjuntos tipo más necesarios para el traductor.

Para cualquier conjunto clásico A es posible definir la función característica μ_A:X→{0,1} como en la ecuación. En la teoría de conjuntos borrosos, la función característica está generalizada de manera que la función de pertenencia asigna <mentos {0,1}. El conjunto que se basa en esta pertenencia extendida se denomina Conjunto Borroso. Con esto hemos definido *Universo de Discurso* como el conjunto **X** de posibles valores que puede tomar la variable x en un texto científico. Se representa:

$$\mathbf{X} = \{x\}$$

La *función de pertenencia* $\mu_A(x)$ de un conjunto borroso A es una función

$$\mu_A : X \to [0,1]$$

Así, cualquier elemento x (concepto de semipar, para adjudicar con el par semántico completo en las dos lenguas) en X tiene grado de pertenencia $\mu_A(x) \in [0,1]$. A queda completamente determinado por:

$$A = \{(x, \mu_A(x) \mid x \in X\}$$

es decir, se puede describir el conjunto borroso en la ecuación como sigue:

$$\begin{aligned} A &= \mu_A(x_1)/x_1 + \mu_A(x_2)/x_2 + \ldots + \mu_A(x_n)/x_n \\ &= \sum_{i=1}^{n} \mu_A(x_i)/x_i \end{aligned}$$

donde el símbolo de división no es más que un separador de los conjuntos de cada par, y el sumatorio es la operación de unión entre todos los elementos del conjunto. El + satisface a/x+b/x = max(a,b)/x, es decir, si el mismo elemento tiene dos grados de pertenencia diferentes 0.8 y 0.6, entonces el grado de pertenencia será 0.8. Se puede escribir cualquier universo discreto en la siguiente forma:

$$A = \sum_{x \in X} \mu_A(x)/x$$

pero cuando X es incontable o es continuo, se describe la ecuación anterior como:

$$A = \int_X \mu_A(x)/x$$

Se puede escribir la ecuación con la notación clásica como sigue:

$$\{\mu_A(x)/x \mid x \in X\}$$

Sean A y B dos conjuntos borrosos definidos respectivamente sobre el universo X e Y, y sea la relación borrosa R definida sobre X×Y. El *soporte* de un conjunto borroso A es el conjunto clásico que contiene todos los elementos de A con los grados de pertenencia que no son cero. Esto se define por S(A).
Se define el soporte de un conjunto borroso A como sigue:

$$S(A) = \{x \in X \mid \mu_A(x) > 0\}$$

Como la borrosidad representa un tipo especial de incertidumbre, en ocasiones se han considerado algunas medidas de la Teoría de la Información Estadística como base para definir medidas de borrosidad. En este sentido, De Luca & Termini (1972) definieron una *entropía no probabilística* (según su propia denominación) que tomaba como base la entropía de Shannon, y que para un subconjunto difuso Ã de X venía dada por:

$$\sum_{x \in X} \left\{\mu\tilde{A}(x) \log \mu\tilde{A}(x) + [1 - \mu\tilde{A}(x)] \log[1 - \mu\tilde{A}(x)]\right\}$$

(bajo el supuesto de que X sea un conjunto discreto) El objetivo perseguido en este trabajo ha sido el de emplear otras medidas de la Teoría de la Información con el mismo fin. Estas medidas son ciertos índices de *afinidad* entre dos distribuciones de probabilidad, que evalúan el grado de consonancia o similitud entre esas distribuciones. Las razones para esta elección se encuentran en algunas ideas directrices que orientan las últimas investigaciones sobre la medición de la borrosidad. Estas ideas señalan que la forma más natural de expresar el grado de borrosidad de un subconjunto difuso en recuperación de traducción especializada es en términos de la falta de distinción entre el subconjunto difuso y su complementario (considerando complementario el conjunto de conceptos contenidos en un texto de partida o de llegada).

Por lo tanto, un conjunto borroso A es *convexo* si y sólo si X es convexo y

$\forall x, y \in X\, \forall \lambda \in [0,1] \mid \mu_A(\lambda x + (1-\lambda)y) \geq min(\mu_A(x), \mu_A(y))$

se define la altura de un conjunto borroso A sobre X, que se denota por Alt(A) como:

$$Alt(A) = \sup_{x \in X} \mu_A(x)$$

y lo consideraremos normal si un conjunto borroso A, si Alt(A)=1, y es subnormal si Alt(A) < 1.

En la teoría de control borroso, es usual tratar sólo con conjuntos borrosos convexos, que nos resultan altamente útiles para agrupar términos de en los que la reducción polisémica no sea factible como los términos de notas de alcance y sólo entonces, dado un número $\alpha \in [0,1]$ y un conjunto borroso A, definimos el *α-corte* de A como el conjunto clásico A_α que tiene la siguiente función de pertenencia:

$$\mu_{A\alpha}(x) = \begin{cases} 1 & \text{cuando } \mu_A(x) \geq \alpha \\ 0 & en\ cualquier\ otro\ caso \end{cases}$$

En definitiva, el *α-corte* se compone de aquellos elementos cuyo grado de pertenencia supera o iguala el umbral α. Las operaciones como la igualdad y la inclusión de dos conjuntos borrosos derivan de la teoría de conjuntos clásicos. Dos conjuntos borrosos son iguales si cada elemento del universo tiene el mismo grado de pertenencia en cada uno de ellos. El conjunto borroso A es un subconjunto del conjunto borroso B si cada elemento del universo tiene grado de pertenencia menor en A que en B. Dos conjuntos borrosos son *iguales* (A=B) sí y sólo sí

$$\forall x \in X : \mu_A(x) = \mu_B(x)$$

Con esta función definimos las equivalencias lingüísticas de par semánticos iguales en lenguas de llegada y partida, para incluirlas en el algoritmo y realizaremos un subconjunto de términos reutilizables siempre que

$$\forall x \in X : \mu_A(x) \leq \mu_B(x)$$

Así los conjuntos borrosos se pueden operar entre sí del mismo modo que los conjuntos clásicos, puesto que los primeros son una generalización del los segundos. La interpretación semántica para la generalización de pares semánticos equivalentes con conjuntos borrosos no es tan simple como con conjuntos clásicos porque se usan las características de funciones de pertenencia. Es posible definir las operaciones de intersección, unión y complemento haciendo uso de las mismas funciones de pertenencia. Zadeh propuso lo siguiente (Zadeh

1965): La *intersección* entre dos conjuntos borrosos se representa como sigue[25]:

$$\forall x \in X : \mu_{A \cap B}(x) = min(\mu_A(x), \mu_B(x))$$

La *unión* entre dos conjuntos borrosos se representa como sigue:

$$\forall x \in X : \mu_{A \cup B}(x) = max(\mu_A(x), \mu_B(x))$$

El *complemento* de un conjunto borroso se representa como sigue:

$$\forall x \in X : \mu_{\overline{A}}(x) = 1 - \mu_A(x)$$

CONCLUSIONES

Los resultados obtenidos indican que el algoritmo propuesto tiene un alto grado de precisión con base en que del número total de pares identificados (32 pares) casi todos (30 pares) son efectivamente pares-semánticos. Siendo básicamente un algoritmo heurístico, el algoritmo de alineamiento semántico puede mejorar la recuperación de los pares-semánticos si se relajan las restricciones sobre las cuales descansa; sin embargo, esa misma relajación puede llevar a una disminución importante en la precisión de los pares identificados. En términos de la aplicabilidad del algoritmo a los problemas de recuperación de información, es necesario buscar un balance entre la recuperación y la precisión. Cualquier variante de las heurísticas subyacentes en el algoritmo debe evaluarse cuantitativamente a fin de tener un criterio para su elección.

[25] Kosko, Bart (1995). *Pensamiento borroso*. Barcelona. Crítica (original: *Fuzzy thinking. The new science of fuzzy logic*. New York: Hyperion, 1991); Shafer, G. (1976). *A Mathematical Theory of Evidence. Princeton*, NJ: Princeton University Press, 1976; Watanabe, S. (1969). Modified concepts of logic, Probability and information based on generalized continuous characteristics function. *Inf. control, 15*, 1-21, 1969; Watanabe, S. (1975). Creative learning and propensity automata. *Trans. Syst., Man Cybern.*, *5*, 603-609; Weber, S. (1983). A general concept of fuzzy connectives, negations and implications based on t-norms and t-co-norms. *Fuzzy Sets and Systems*, *11*, 115-134. Zadeh, L.A. (1964). *Fuzzy sets. Memorandum ERL*. Berkley: Univ. of California. (publicado en *Information and Control*, 1965); Zadeh, L.A. (1965). Fuzzy sets. *Information and Control, 8(3),* pp. 338-353.