

# La consistencia como indicador de la fiabilidad (corrección) de la indización: los casos de LISA, PASCAL y VOCED

**Luis Miguel Moreno Fernández**

Universidad de Murcia (España)

**Mónica Izquierdo Alonso**

Universidad de Alcalá de Henares (España)

**Antonio Jesús Vázquez Montávez**

Licenciado en Documentación

## Resumen

Se analizan las posibilidades de la consistencia como indicador de calidad de la indización mediante un doble procedimiento. En primer lugar, se calcula la coherencia habida en las bases de datos de LISA y PASCAL aplicando solo las fórmulas de Hooper y Rolling. Esto permite demostrar que la consistencia no necesariamente muestra que estemos ante una indización de calidad, si no tenemos en cuenta el contexto en el que se lleva a cabo esa descripción de los documentos. Después se procede a investigar en la base de datos VOCED la vinculación entre consistencia y calidad de la indización, estudiando la interrelación entre aquella y los restantes factores que modulan la indización: corrección, exhaustividad y especificidad. Para ello se compara la indización hecha en VOCED con una indización efectuada como “modelo” o “patrón oro”, lo que permite asimismo añadir a las dos fórmulas citadas la del índice de Kappa. Este procedimiento se ha concretado en un modelo que contempla toda la casuística que incide en una correcta indización (semántica, pragmática y ambiental), a partir de un patrón de referencia. El método revela que la consistencia por sí sola no constituye un indicador de calidad, pero que sí es condición necesaria para alcanzarla, correlacionando la consistencia con las demás características de la indización. Asimismo determina las causas de variabilidad en las ratios de consistencia y detecta los errores u omisiones que se produzcan en la indización.

**Palabras clave:** Consistencia en la indización. Bases de datos. LISA. PASCAL. VOCED. Calidad de la indización.

## Abstract

The possibilities of the consistency as an indexing quality indicator are analyzed through a double procedure. First of all, the consistency found within the LISA and

PASCAL databases after applying both the Hooper's and Rolling equations is stated. This theory demonstrates that, if the context in which the document is described is not taken into account, consistency may not necessarily show the indexing quality of the text. Next, the link between consistency and indexing quality in the VOCED database is also stated, by studying the relationship between consistency itself and the remaining factors that build indexing: correction, exhaustivity and specificity. In order to reach to that conclusion, the VOCED indexing is compared to a so called "model" or "golden pattern" indexing, which enables us to add the Kappa index to the two already mentioned formulas (Hooper and Rolling). This procedure therefore constitutes a model that includes all the factors (semantic, pragmatic and environmental factors) that may affect the correct indexing process from a quality indicator. The method states that consistency itself is not a quality indicator, but a required element to complete it. It also defines the reasons of consistency ratio variability and finds the mistakes or possible omissions in the indexing process.

**Keyword:** Indexing consistency. Databases. LISA. PASCAL. VOCED. Indexing quality.

## 1. Introducción

La calidad de la indización incide fundamentalmente —por lo menos para la mayoría de los autores— en la eficiencia que puede tener la búsqueda documental por materias en el contexto de un sistema de recuperación de la información (SRI).

Haciendo abstracción de los matices que ofrecen los estudios de consistencia, hay tres posturas en torno a esta cuestión (Moreno, 2003b). Son las siguientes:

- a. Teorías que cuestionan o relativizan la coherencia entre indizadores como indicador de la calidad de la indización y orientan las investigaciones hacia el plano de la coherencia establecida entre los descriptores escogidos por el indizador y los que escoge el usuario para la recuperación de información: Cooper (1969), Diodato (1981), Fugmann (1984), Lancaster (1991, 1998), Soergel (1985, 1994), Braam y Bruil (1992), Fidel (1994).
- b. Teorías que dan por sentada la relación directa entre coherencia de los indizadores y calidad de la indización, y, a su vez, entre estas y el rendimiento del sistema de recuperación de la información: Leonard (1975), Foskett (1982), Funk y Reid (1983), Chan (1989), Sievert y Andrews (1991), Aluri, Kemp y Boll (1991), Hodge (1992), Furner, Ellis y Willet (1994, 1999), Plaunt y Norgard (1998), Extremeño (1999), Gil-Leiva (2001), Leininger (2000), Uren (2000), Berrios (2000), Berrios, Cucina y Fagan (2002).
- c. Teorías que disocian coherencia y calidad de la indización de la eficiencia o rendimiento de la recuperación: Zunde y Dexter (1969ab), Rolling (1981), White y Griffith (1987).

El sector mayoritario (*b*) acepta la vinculación entre coherencia y calidad de la indización y entre esta y la eficiencia del SRI. Pero estos autores no constituyen un grupo monolítico, sino muy variado, que ha desarrollado trabajos de desigual calado, limitándose en general a dar por buena la tesis de Leonard (1975) acerca de la correlación positiva entre coherencia y rendimiento de las búsquedas. Es decir, que centran sus trabajos en aplicar unas fórmulas (por ejemplo, la más difundida, de Hooper, 1965) u otras (por ejemplo, la de Rolling, 1981) para calcular la consistencia, al margen de cualesquiera consideraciones sobre las circunstancias que pueden mediatizar aquella, y, por ende, influir en la calidad de la indización.

Dejando a un lado las ventajas e inconvenientes que tenga la aplicación de una determinada fórmula, lo cierto es que buena parte de esos trabajos arrojan unos porcentajes sobre consistencia o inconsistencia *escasamente fiables*, porque no tienen en consideración las condiciones en las que se ha llevado a cabo la indización. Así, por ejemplo, aplican las fórmulas al margen de que los documentos obtenidos de diferentes bases de datos hubieran sido descritos por medio de *lenguajes de indización diferentes* y siguiendo *procedimientos o criterios distintos*. Esto es, sencillamente, incongruente con el criterio de coherencia o consistencia, que por principio refiere al acuerdo/desacuerdo respecto de términos de indización escogidos por indizadores que desempeñan su actividad sirviéndose de las mismas herramientas y siguiendo unas pautas de trabajo homogéneas. Otro inconveniente añadido que suelen tener es que ignoramos en qué grado los registros con los que se ha trabajado constituyen una muestra representativa del conjunto de documentos que integra la base de datos. En estas condiciones creemos que los datos no son verdaderamente representativos de una indización de calidad.

Hay autores (Sievert y Andrews, 1991; Leininger, 2000) que han soslayado este problema aprovechando los registros duplicados existentes en la misma base de datos; un enfoque más acertado, si lo que se pretende analizar es una hipotética “calidad” de la indización, pues presuponemos que dentro un mismo contexto y manejando iguales herramientas, los indizadores habrían de coincidir en la descripción característica de un documento. Este procedimiento presenta la ventaja añadida de revelar cuál ha sido la indización de los documentos en condiciones “reales” dentro del sistema de información, más allá de las limitaciones impuestas por una prueba “experimental”. Pero no siempre resulta factible disponer de registros duplicados.

Por lo tanto, optamos por barajar otra opción; se fundamenta en la hipótesis en virtud de la cual para obtener unos datos de consistencia mínimamente fiables hay que partir de las circunstancias en las que aconteció la indización, con el fin de confeccionar un modelo de referencia que permita comparar las indizaciones. Este patrón también incluye los demás factores que inciden en la caracterización del documento, puesto que la consistencia por sí misma no revela demasiado sobre

la calidad de la indización. De ahí que hayamos desarrollado el trabajo en dos fases, la segunda consecuencia de la primera. Así, en principio procedimos a calcular la consistencia de la indización entre las bases de datos de LISA y PASCAL siguiendo las pautas anteriormente *criticadas y habituales* en bastantes trabajos; y después, para ilustrar que la coherencia o consistencia de la indización *no necesariamente* es sinónimo de calidad, investigamos la relación entre consistencia y calidad tratando de analizar empíricamente la vinculación entre la coherencia y los restantes factores que caracterizan la indización en la base de datos VOCED.

## 2. Objetivos I

Calculamos la consistencia de la indización entre las bases de datos LISA y PASCAL, aplicando simplemente las conocidas fórmulas de Hooper y Rolling a los registros comunes capturados en ambas bases de datos.

### 2.1. Metodología I

LISA y PASCAL son dos bases de datos especializadas en temas de biblioteconomía y documentación, pero americana la primera y francesa la segunda. Ambas indizan sus documentos con lenguajes documentales diferentes. El tema elegido para el estudio ha sido la “indización automática” (*automatic indexing*), aunque también encontraremos los vocablos *automatic support to indexing*, *automated support to indexing* y *automated assisted indexing*. La recuperación de la información sobre el tema la hemos restringido a los documentos publicados entre 1990 y 2004.

Realizadas las búsquedas, obtuvimos 240 documentos de LISA y 516 de PASCAL. Sin embargo, el número que barajamos para la investigación se reduce a 60 registros, que son los comunes a ambas bases de datos. Dado que los lenguajes de indización son diferentes, su morfología también varía, por lo que seguimos las siguientes pautas para comparar los términos de indización:

- a. Si un término aparece en singular en una base de datos y en plural en la otra, se considera que son dos términos distintos.
- b. Las expresiones constituidas por una frase en una base de datos tendrán que ser exactamente igual en la segunda para que las estimemos coincidentes.
- c. Los sinónimos también son tratados como dos términos distintos.
- d. Los términos genéricos en una base de datos y los específicos derivados o dependientes de ese general en la otra base de datos son tratados como dos términos diferentes sin acuerdo entre los indizadores.

Para calcular la consistencia de la indización entre los documentos coincidentes de las dos bases de datos, decidimos elegir las fórmulas de Hooper y Rolling. De este modo compararemos los resultados que arrojan estos dos métodos a partir de una misma muestra, y posteriormente veremos las particularidades de estos datos

respecto de los que obtenemos por procedimientos tan distintos como el anteriormente citado.

La ecuación de Hooper, utilizada por Sievert y Andrews en su estudio sobre *Information Science Abstracts* (1991), suele enunciarse así:

$$CP (\%) = 100 A / A + M + N = A \cap B / A \cup B$$

Donde *A* es el número de términos de acuerdo entre ambas bases de datos, *M* es el número de términos únicos usados solamente por un indizador, *N* es el número de términos únicos usados por el otro indizador y 100 es el factor que sitúa a la consistencia entre un rango de 0 a 100%.

La medida de la consistencia de la indización de Rolling, usada por Leininger (2000) en sus estudios, se puede resumir en:

$$C (\%) = 2c / A + B$$

Siendo *2c* el número de términos en los que están de acuerdo los dos indizadores, y *A + B* el número total de términos usados por ambos indizadores. Con esto Rolling intenta evitar que los términos coincidentes o concordantes tengan un peso específico tan elevado a favor de la coherencia como los discordantes en contra de ella.

## 2.2. Resultados I

En la siguiente tabla se muestran los resultados obtenidos de calcular la consistencia de la indización de los documentos coincidentes por ambas ecuaciones:

Documentos	L*	P**	L + P***	C. Hooper	C. Rolling
Documento 1 (Moya Anegón. <i>Genetic algorithms...</i> )	4	8	2	9,09%	16,66%
Documento 2 (Hauer. <i>Intelligent capture...</i> )	4	4	1	14,28%	25,00%
Documento 3 (Toht. <i>Az Automaticus...</i> )	6	7	3	30,00%	46,15%
Documento 4 (Godby. <i>The WordSmith...</i> )	5	7	1	9,09%	16,66%
Documento 5 (Godby. <i>Experimental...</i> )	10	8	2	12,50%	22,22%
Documento 6 (Bloomfield. <i>Indexing...</i> )	5	10	2	15,38%	26,66%
Documento 7 (Toht. <i>Innovate...</i> )	7	6	3	30,00%	46,15%
Documento 8 (Drott. <i>Indexing...</i> )	3	7	1	11,11%	20,00%
Documento 9 (Rapke. <i>Automatische...</i> )	9	7	1	6,66%	12,50%
Documento 10 (Ladewig. <i>Verfahren...</i> )	2	7	1	12,50%	22,22%

<b>Documentos</b>	<b>L*</b>	<b>P**</b>	<b>L + P***</b>	<b>C. Hooper</b>	<b>C. Rolling</b>
Documento 11 (Kim. <i>Corpus...</i> )	3	2	0	0	0
Documento 12 (Grumann. <i>Sind...</i> )	6	5	1	10,00%	18,18%
Documento 13 (Roberts. <i>The automation...</i> )	7	14	3	16,60%	28,57%
Documento 14 (Lepsky. <i>Katalogerweiterung...</i> )	8	6	2	16,66%	28,57%
Documento 15 (Lepsky. <i>Automatische...</i> )	4	8	1	9,09%	16,66%
Documento 16 (Humphrey. <i>Automatic...</i> )	2	7	1	12,50%	22,22%
Documento 17 (Fuketa. <i>A fast...</i> )	3	8	1	10,00%	18,18%
Documento 18 (Plaunt. <i>An association-based...</i> )	2	9	2	22,22%	36,36%
Documento 19 (Chen. <i>Semantic...</i> )	7	12	1	5,55%	10,52%
Documento 20 (Wan. <i>Experiments...</i> )	6	9	1	7,14%	12,50%
Documento 21 (Liu. <i>Decomposing...</i> )	6	12	2	12,50%	22,22%
Documento 22 (Hmeidi. <i>Design...</i> )	4	10	1	7,69%	14,28%
Documento 23 (Kim. <i>An automatic...</i> )	4	5	1	12,50%	22,22%
Documento 24 (Gil-Leiva. <i>Análisis...</i> )	6	7	1	8,33%	15,38%
Documento 25 (Duffield. <i>Tory...</i> )	4	6	0	0	0
Documento 26 (Bordón. <i>Documents...</i> )	6	8	1	7,69%	14,28%
Documento 27 (Leung. <i>A statistical...</i> )	4	13	2	13,33%	23,52%
Documento 28 (Ward. <i>The future...</i> )	2	8	1	11,11%	20,00%
Documento 29 (Gil Leiva. <i>Tendencias...</i> )	3	6	1	12,50%	22,22%
Documento 30 (Guedes. <i>Estudo...</i> )	2	6	1	14,28%	25%
Documento 31 (Lepsky. <i>Automatische...</i> )	4	9	2	16,66%	30,76%
Documento 32 (Alexander. <i>Automatic...</i> )	6	7	1	8,33%	15,38%

Documentos	L*	P**	L + P***	C. Hooper	C. Rolling
Documento 33 (Aptagiri. <i>A frame...</i> )	4	14	4	28,57%	44,00%
Documento 34 (Hebrail. <i>The SPHERE...</i> )	6	13	1	5,55%	10,52%
Documento 35 (Godert. <i>Inhaltliche...</i> )	2	13	1	7,14%	13,33%
Documento 36 (Cristal. <i>Some indexing...</i> )	3	13	0	0	0
Documento 37 (Lepsky. <i>Maschinelles...</i> )	7	13	2	11,11%	20,00%
Documento 38 (Chen. <i>Automatic...</i> )	5	17	0	0	0
Documento 39 (Cohen. <i>Highlights...</i> )	1	12	1	8,33%	15,38%
Documento 40 (Wu. <i>ACTS...</i> )	4	17	2	10,00%	19,04%
Documento 41 (Sridhar. <i>Generation...</i> )	3	13	1	6,66%	12,50%
Documento 42 (Larroche. <i>Das Nominalsyntaxma...</i> )	3	16	0	0	0
Documento 43 (Chang. <i>The...</i> )	1	9	1	11,11%	20,00%
Documento 44 (Broadhurst. <i>Caere...</i> )	3	9	0	0	0
Documento 45 (Smart. <i>Using...</i> )	3	16	2	11,76%	21,05%
Documento 46 (Gretz. <i>Indexierungen...</i> )	12	7	0	0	0
Documento 47 (Hersh. <i>A comparison...</i> )	9	18	1	3,84%	7,40%
Documento 48 (Rabitti. <i>Automatic...</i> )	2	16	1	5,88%	11,11%
Documento 49 (Harada. <i>Automatic...</i> )	7	7	1	7,69%	14,28%
Documento 50 (Sierverts. <i>Software...</i> )	9	11	1	5,26%	10,00%
Documento 51 (Pereira. <i>Representação...</i> )	5	5	0	0	0
Documento 52 (Teixeira. <i>Indexação...</i> )	5	8	0	0	0
Documento 53 (Milstead. <i>Methodologies...</i> )	5	12	0	0	0
Documento 54 (Schulz. <i>Einführung...</i> )	4	8	0	0	0

Documentos	L*	P**	L + P***	C. Hooper	C. Rolling
Documento 55 (Gibb. <i>Structured...</i> )	6	12	0	0	0
Documento 56 (Krause. <i>Padok-II...</i> )	17	11	2	7,40%	14,28%
Documento 57 (Anderson. <i>The nature...Part II</i> )	3	13	1	6,66%	12,50%
Documento 58 (Anderson. <i>The nature... Part I</i> )	3	14	1	6,25%	11,76%
Documento 59 (Godert. <i>Maschinelle...</i> )	7	8	2	15,38%	26,66%
Documento 60 (Wellisch. <i>Book...</i> )	2	11	0	0	0
				media = 9,06%	media = 15,91%
* Número de términos asignados por la base de datos LISA a un documento. ** Número de términos asignados por la base de datos PASCAL a un documento. *** Numero de términos coincidentes por ambas bases de datos para un documento.					

*Tabla I. Consistencia de la indización entre LISA y PASCAL*

Posteriormente, y simulando el estudio de Stubbs, Mangiaterra y Martínez (1999), hemos hallado la consistencia media para el grupo entero de documentos recuperados mediante la siguiente fórmula:

$$\text{media} = \sum x / n = \text{MCI} = \sum \text{PCI} / n \text{ PCI}$$

Los resultados, considerados en sí mismos, arrojan porcentajes de coherencia muy reducidos, en contra de lo que en principio supusimos. Dado que el tema escogido para nuestro análisis es muy concreto, pensábamos que la consistencia debería ser, por lo menos, relativamente elevada. Sin embargo, no ha sido así. Los porcentajes más pobres se obtienen mediante la aplicación de la fórmula de Hooper. Los valores extremos nos muestran una consistencia del 3,84% y del 30%. En cambio, mejoran apreciablemente con la fórmula de Rolling, aunque tampoco son elevados: 7,40% y 46,15% respectivamente también en los casos extremos. Este ligero aumento se debe a que se multiplican por 2 los términos comunes, a los que se otorga un peso específico mayor que a los discordantes. Pero no perdamos de vista que de los 60 documentos analizados, nada menos que 12, es decir, el 20%, ofrecen una inconsistencia absoluta, que se traduce en 0. Los promedios continúan siendo bajos: el 9,06% según el procedimiento de Hooper, y el 15,91% siguiendo el método de Rolling.

En ambos casos son muy bajos; no figuran dentro del rango que se considera más o menos válido: entre el 25% y el 50% sería lo normal (Leininger, Markey y Sievert).



### **2.3. Conclusiones I**

Hemos comprobado que a pesar de la concreción del tema sometido a investigación, la consistencia ha sido baja. Esto nos permite conjeturar que la coherencia no aumenta por el simple hecho de que los documentos analizados revelen contenidos específicos y homogéneos, aunque investigaremos más este asunto, puesto que el número de documentos analizados nos parece insuficiente para teorizar sin dejar resquicio a la duda en este sentido.

A la vista de estos resultados, ¿cabría concluir sin más que la indización efectuada en estas bases de datos es de baja calidad? Las dos son bases de datos que tratan de temas relativos a biblioteconomía y documentación y que parece que cuentan con indizadores expertos. En estas circunstancias, tal juicio resulta temerario, porque la metodología empleada, como comentamos anteriormente en nuestra hipótesis de trabajo, no es la más acertada, dado que no tuvimos en cuenta la política de indización y el protocolo que se ha seguido en cada una de ellas y las diferencias habidas entre los lenguajes documentales empleados para describir los documentos. Por ejemplo: observamos que en PASCAL los términos asignados a los documentos son muchos; incluso, en determinadas ocasiones, duplican en número los utilizados por LISA. Además, PASCAL se sirve de términos muy genéricos, mientras que LISA escoge descriptores más específicos.

¿Cómo buscar entonces la relación entre consistencia y calidad de la indización, suponiendo que la hubiera?

### **3. Objetivos II**

Posteriormente, en esta segunda fase, investigamos la relación entre consistencia y calidad tratando de analizar empíricamente la vinculación entre la coherencia y los restantes factores que caracterizan la indización: consistencia o coherencia, corrección o exactitud, exhaustividad y especificidad.

Partiendo de la tesis de Zunde y Dexter y de Rolling, elaboramos un modelo o “patrón oro” de indización de un conjunto de  $n$  documentos que permite contrastar la indización de una tercera base de datos (VOCED) respecto de un modelo de referencia, que ha sido confeccionado compartiendo la misma herramienta y técnica de indización que se siguió en la base de datos. Para la realización del patrón procuramos tener en cuenta las condiciones necesarias para una adecuada elaboración de estos instrumentos de trabajo: a) definición clara de los datos que pretendemos valorar; b) considerar las normas y criterios que se siguen en la selección de los términos; c) ejercitación en el uso de esas normas; d) acceso a toda la información disponible que resulte precisa para la designación de los términos de indización (Abad, 1997).

Los factores que concurren durante el proceso de la indización no actúan de manera independiente, sino de forma interrelacionada, dado que hay “determinantes” que les son comunes. Circunstancia que no se suele tener en consideración cuando se escribe sobre la coherencia o consistencia (Moreno, 2003a). En nuestra opinión, el inconveniente que tienen algunos estudios acerca de la consistencia que niegan o relativizan su impacto en la calidad de la indización reside en que la presentan como una variable “aislada” de las restantes, por lo que resulta “fácil” cuestionar su incidencia en la calidad de aquella y explicar que una indización puede ser “coherentemente mala” (Cooper, 1969), como efectivamente ocurre si es imprecisa o errónea.

Para que eso no suceda, por ejemplo, la *corrección* de la indización requiere la adecuada identificación de los conceptos, y, además, la asignación de términos con el nivel preciso de *exhaustividad* y *especificidad*; la coherencia requiere que el indizador escoja unos descriptores que se ajusten a las normas de *corrección*, *exhaustividad* y *especificidad*, si no queremos contar con una indización coherentemente mala. Y, en fin, una elevada exhaustividad comporta, si bien de manera indirecta, cierto nivel de especificidad (Moreno, 2003a).

Por lo tanto, si convenimos que la indización adecuada de un documento debe tener cierto grado de corrección y especificidad, podremos aceptar también que la consistencia, cuando aparece imbricada con la corrección, la exhaustividad y la especificidad, puede constituir un indicador de la “calidad” de la indización.

### **3.1. Metodología II**

Para asegurar en cierta medida que estos tres últimos factores se hallan detrás de la consistencia verdaderamente, el procedimiento que hemos seguido para estimarla en este trabajo es la elaboración de una indización “estándar” o modelo que responda a esos requisitos, con el fin de compararla con la que presentan los registros recuperados de la base de datos.

El procedimiento consiste en obtener de una base de datos cierto número de documentos descritos en los registros recuperados al efectuar búsquedas documentales sobre temas concretos. Seguidamente se han vuelto a indizar mediante el mismo lenguaje documental y procurando aplicar las pautas que han seguido los indizadores de aquella. Con esto pretendemos elaborar un “patrón” o “modelo” de indización, de manera que podamos comparar la indización hecha en la base de datos con la efectuada por nosotros. Elaborar un “patrón oro” con el que contrastar la calidad de la indización hecha por otros agentes gana adeptos en el campo de la indización automática (por ejemplo, Uren, 2000; Berrios, 2002; véase también Moreno, 2003b). El objetivo fundamental es comparar estos resultados con los que se obtienen al calcular la coherencia por los procedimientos que marginan el “contexto” en el que se ha efectuado la indización, tal y como ocurre analizando sin más

la consistencia de un grupo de documentos escogidos de dos o más catálogos de bibliotecas. La finalidad de este análisis es demostrar que estos últimos datos son escasamente significativos para estimar la calidad de una indización, dado que —insistimos en ello— evalúan la consistencia al margen del contexto en el que acontece la indización.

Por otro lado, adoptar un “modelo” o “estándar” de referencia para comparar la consistencia permite manejar una tercera fórmula para estimar la consistencia. Nos referimos al *índice de concordancia de Kappa*, enunciado en 1960. Gracias a él podemos establecer hasta qué extremo el grado de concordancia entre dos personas, métodos de análisis o evaluaciones de un fenómeno, según categorías mutuamente excluyentes, es superior al que se espera obtener por puro azar (Molinero, 2001). El índice de Kappa se obtiene dividiendo el número de veces que el valor del dato registrado en el documento primario se confirma en el patrón por el número total de ocasiones en las que se verifica la comparación (Abad, 1997). Y se formula de la manera siguiente:

$$K = (Po - Pe / 1 - Pe)$$

Donde  $Po$  es la proporción de concordancia apreciada (en tanto por 1) y  $Pe$  es la proporción de concordancia esperada por puro azar (Molinero). Si hay acuerdo absoluto, la concordancia será de 1; por eso  $1 - Pe$  constituye el margen de acuerdo posible que no puede atribuirse al azar.  $Pe$  es la suma o producto de ambas probabilidades.

Esta fórmula permite contrastar la coincidencia-divergencia de trabajos con relación a un modelo o patrón. Esto enlaza con las tesis de Zunde y Dexter y de Rolling sobre la conveniencia de comparar la indización de los documentos para establecer la consistencia respecto de un “modelo” o patrón establecido. Por medio del índice de concordancia de Kappa, además, se suprime el azar entre los individuos que realizan la acción.

### 3.2. Resultados II

De modo que el planteamiento genérico del trabajo que estamos llevando a cabo y que avanzamos aquí queda establecido en la tabla II.

Para que el conjunto de  $n$  documentos resulte representativo de una base de datos que cuenta con cerca de 26.500 registros, no nos hemos limitado a escoger solo un determinado número de documentos, sino que estos responden a temas diversos que tienen cabida en la base de datos. Por ejemplo: economía, democracia, infancia, internet, sida y ONG (organizaciones no gubernamentales). De los documentos obtenidos, se analiza una muestra comprendida entre el 5% y el 10% para cada uno de esos temas. Así pretendemos detectar también si existieran diferencias de consistencia en la indización de documentos pertenecientes al abanico de temas escogidos.

Documento	BD	Modelo	% t. cons.	% t. no cons.	Totales	Hooper	Rolling	Kappa
Descriptores principales								
Descriptores secundarios								
Totales								
Términos incorrectos: morfosintácticos asignación* relevancia** uso***								
Totales								
* Términos que no figuran en el Tesauro de la base de datos. ** Elección de conceptos (términos) errónea por no adecuarse al tema principal o a los temas secundarios del documento. *** Confundir en el Tesauro el término preferente con el o los no preferentes.								

*Tabla II. Consistencia y fiabilidad (corrección) de la indización en VOCED*

### 3.3. Conclusiones II

La consistencia, por sí sola, no revela a ciencia cierta que estemos ante una indización de calidad o no, pero creemos que sí es condición necesaria para alcanzarla. Por eso, si podemos mostrarla en el contexto de las demás características que acompañan a la descripción del contenido informativo de los documentos, contaremos con un indicador de calidad de mayor fiabilidad, que nos servirá para detectar los errores u omisiones que se produzcan en esa actividad, e incluso las posibles inconsistencias del lenguaje documental. A partir de los datos de concordancia que obtengamos pretendemos correlacionar la consistencia de la indización con la “calidad” de ella, en el contexto de los “términos incorrectos” que aparecen en el registro, y que comprenden vocablos morfosintácticos discordantes, asignación de términos en lenguaje natural, que no tienen nada que ver con los que figuran en el tesauro, expresiones irrelevantes para el contenido informativo del documento y elección errónea del término.

### Referencias

- Abad García, M. F. (1997). Investigación evaluativa en documentación. Aplicación a la documentación médica. Valencia: Universitat de València, 1997.
- Aluri, R.; Kemp, D. A.; Boll, J. J. (1991). Subject analysis in online catalogs. Englewood (Colorado): Libraries Unlimited, 1991.
- Berrios, D. C. (2000). Automated indexing for full text information retrieval. // 2000 AMIA Fall Symposium, Los Ángeles, 2000. (Actas en CD-Rom).
- Berrios, D. C.; Cucina, R. J.; Fagan, L. M. (2002). Methods for semi-automated indexing for high precision information retrieval. // Stanford Medical Informatics. 9:6 (2002) 637-652.

- Braam, R. R.; Bruil, J. (1992). Quality of indexing information: authors's views on indexing of their articles in chemical abstracts online CA-file. // *Journal of Information Science*. 18 (1992) 399-408.
- Chan, L. M. (1989). Inter-indexer consistency in subject cataloging. // *Information Technology and Libraries*. 8:4 (1989) 349-358.
- Cooper, W. S. (1969). Is interindexer consistency a hobgobling? // *American Documentation*. 20 (1969) 268-278.
- Diodato, V. P. (1981). *Author indexing in Mathematics*. Urbana: University of Illinois, Graduate School of Library and Information Science, 1981.
- Ellis, D.; Furner-Hines, J.; Willet, P. (1994). On the measurement of inter-linker consistency and retrieval effectiveness in hypertext databases. // *Proceedings of the 17<sup>th</sup> Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland, July 3-6 1994). Londres: Springer, 1994.
- Extremeoño, A. (1999). Calidad de la indización e incidencia de errores en la base de datos ECOSOC. // *Revista Española de Documentación Científica*. 22:2 (1999) 157-173.
- Fidel, R. (1994). User-centered indexing. // *Journal of the American Society for Information Science*. 45:8 (1994) 572-576.
- Foskett, A. C. (1982). *The subject approach to information*. Londres: Clive Bingley, 1982.
- Fugmann, R. (1985). The five-axiom theory of indexing and information supply. // *Journal of the American Society for Information Science*. 36:2 (1985) 116-129.
- Funk, M. E.; Reid, C. A. (1983). Indexing consistency in MEDLINE. // *Bulletin of the Medical Library Association*. 71 (1983) 176-183.
- Furner, J.; Ellis, D.; Willet, P. (1994). On the measurement of inter-linker consistency and retrieval effectiveness in hypertext databases. *Annual ACM Conference on Research and Development in Information Retrieval*. //: *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, Dublin, Ireland, 1994, p. 51-60
- Furner, J.; Ellis, D.; Willet, P. (1999). Inter-linker consistency in the manual construction of hypertext documents. // *ACM Computing Surveys*. 31 (4es), art. n° 18, 1999.
- Gil-Leiva, I. (2001). Consistencia en la asignación de materias en bibliotecas públicas del Estado. // *Boletín de la Asociación Andaluza de Bibliotecarios*. 63 (2001) 69-86.
- Hodge, G. M. (1992). *Automated support to indexing*. Philadelphia: National Federation of Abstracting and Information Services, 1992.
- Hooper, R. S. (1965). *Indexer consistency tests: origin, measurements, results, and utilization*. // Bethesda, MD: IBM Corp., 1965.
- Lancaster, F. W. (1991). *Indexing and abstracting in theory and practice*. Londres: The Library Association, 1991.
- Lancaster, F. W. (1998). *Indexing and abstracting in theory and practice*. Londres: The Library Association, 1998.
- Leininger, K. (2000). Interindexer consistency in PsycINFO. // *Journal of Librarianship and Information Science*. 32:1 (2000) 4-8.

- Leonard, L. E. (1975). Interindexer consistency and retrieval effectiveness: measurement of relationships. University of Illinois, Graduate School of Library Science, 1975. (Inédito).
- Moliner, L. M. (2001). Medidas de concordancia para variables cualitativas (online). URL: <<http://www.seh-lilha.org/concor2.htm>>. Consultado: 2003-02-19.
- Moreno Fernández, L. M. (2003a). La consistencia de la indización: I. Precisiones conceptuales, medidas, usos en los procesos documentales y relación entre coherencia e indización. // *Revista AIBDA (Asociación Interamericana de Bibliotecarios, Documentalistas y Especialistas en Información Agrícola)*. XXIV: 1-2 (2003) 1-30.
- Moreno Fernández, L. M. (2003b). La consistencia de la indización: II. Estado de la cuestión y tendencias de la investigación. // *Revista AIBDA (Asociación Interamericana de Bibliotecarios, Documentalistas y Especialistas en Información Agrícola)*. XXIV: 1-2 (2003) 31-66.
- Plaunt, Ch.; Norgard, B. A. (1998). An association based method for automatic indexing with a controlled vocabulary. // *Journal of the American Society for Information Science*. 49:10 (1998) 888-902.
- Rolling, L. (1981). Indexing consistency, quality and efficiency. // *Information Processing and Management*. 17 (1981) 69-76.
- Sievert, M. C.; Andrews, M. J. (1991). Indexing consistency in information science abstracts. // *Journal of the American Society for Information Science*. 42:1 (1991) 1-6.
- Soergel, D. (1985). *Organizing information. Principles of data base and retrieval systems*. Orlando, FL: Academic Press, 1985.
- Soergel, D. (1994). Indexing and retrieval performance: The logical evidence. // *Journal of the American Society for Information Science*. 45:8 (1994) 589-599.
- Stubbs, E. A.; Mangiaterra, N. E.; Martínez, A. M. (1999). International quality audit of indexing: a new application of interindexer consistency. // *Cataloging & Classification Quarterly*. 28:4 (1999) 53-69.
- Uren, V. (2000). An evaluation of text categorization errors. // *Proceedings of the One-day Workshop on Evaluation of Information Management Systems (15 September 2000)*. Londres: Queen Mary and Westfield College, 2000. 79-87.
- White, H. D.; Griffith, B. C. (1987). Quality of indexing in online data bases. // *Information Processing and Management*. 23:3 (1987) 211-224.
- Zunde, P.; Dexter, M. E. (1969a). Factors affecting indexing performance. // *Proceedings of the 32<sup>nd</sup> Annual Meeting of the American Society for Information Science*. Westport, CT: Greenwood, 1969.
- Zunde, P.; Dexter, M. E. (1969b). Indexing consistency and quality. // *American Documentation*. 20:4 (1969) 259-267.