Análisis bibliométrico de la lírica textual. El caso de *Martina Portocarrero en vivo en el Teatro Municipal*

Rubén Urbizagástegui Alvarado
Universidad de California, Riverside
Flori Martha Urbizagástegui Alvarado
CBSR/CCILA
Universidad de California, Riverside
ruben@urc.edu

Resumen

Con respecto al análisis bibliométrico de la lírica textual del casete *"Martina Portocarrero en vivo en el Teatro Municipal"*, fueron encontradas 1999 palabras que representan un total de 387 palabras diferentes. De esas palabras diferentes y a través del punto de transición de Goffman, fueron aisladas hasta 11 palabras que pueden servir como palabras claves para la indización de la lírica textual. El aislamiento de las palabras claves fue realizado por los métodos de los rangos mínimos y rangos máximos, que no parecen producir diferencias significativas. El modelo Gauss-Poisson inverso generalizado, y la prueba del chi-cuadrado fueron usados para evaluar el ajuste de los datos observados a los datos esperados del vocabulario activo de Martina Portocarrero. Se verificó que esta distribución se ajusta adecuadamente al modelo Gauss Poisson inversa generalizada.

Palabras claves: Martina Portocarrero, Ley de Zipf, Bibliometría, Música ayacuchana, Frecuencias de palabras, Punto de transición de Goffman.

Abstract

Bibliometric analyses of the lyrical text of the cassette tape *"Martina Portocarrero alive in the Municipal Theater"*. In the lyrical text of this cassette, 1999 words were count representing a total of 387 different words. The words mostly repeated were conjunctions, prepositions, personal pro-

nouns and articles. From the 387 different words, up to 11 words that can be used as keywords for indexing purposes of the lyrics, were isolated. The isolation of the keywords was made by the minimum and maximum rank methods, which do not seem to produce significant differences. The Gauss-Poisson inverse generalized model and the chi-square test was used to evaluate the fitting of the observed to expected data. It was verified that the distribution of word frequencies of this lyrical text fits very well to the Gauss-Poisson inverse generalized model.

Keywords: Martina Portocarrero, Zipf's law, Bibliometrics, Ayacuchan music, Word frequency distribution, Goffman's transition point.

Introducción

En ciencia de la información y bibliotecología, la indización consiste en la identificación del contenido de un documento y su representación a través de "palabras-claves". Se busca así expresar la información más significativa del documento con la asignación de términos descriptores, creando de esa manera un lenguaje de mediación entre el lector y el documento. Estas palabras-claves son usadas posteriormente en la recuperación de la información en los catálogos de centros de información, documentación o en las bases de datos especializadas. La indización tiene dos formas de expresión: indización manual, cuando es realizada por una persona y la indización automática, cuando se realiza a través de programas especiales ejecutadas por la computadora. En la indización manual los descriptores son escogidos libremente de acuerdo a la capacidad del indizador o siguiendo un conjunto de reglas de selección de palabras-claves de un vocabulario controlado. En la indización automatizada es la computadora la que realiza la selección de las palabras presentando un listado alfabético o jerarquizado de su presencia en los documentos.

Uno de los problemas que enfrenta la indización manual es el tiempo disponible para su ejecución y el volumen de documentos disponibles esperando oportunidad para ser indizados. Otro de los problemas se relaciona al dominio del texto, pues es necesario que los descriptores expresen el vocabulario de uso corriente específicos a cada campo de conocimiento. La familiaridad y el conocimiento del indizador sobre la terminología usada en los campos científicos son factores que inciden en la calidad de la indización. Igualmente el dominio del idioma en que está escrito el documento. Irónicamente, es en esta indización manual que la calidad de la indización se viene mostrando como inadecuada pues, además de ser un proceso moroso y caro, no hay formas de minimizar la subjetividad del indizador (Bruzinga; Maculan & Lima 2007). Para superar estos problemas se impulsaron las investigaciones en la indización automática. Este proceso consiste en la mecanización de la indización con el pro-

pósito de reducir la interferencia de la subjetividad del indizador, tanto en el análisis del documento como en la selección de las palabras-claves (Mamfrim, 1991) minimizando, al mismo tiempo, los problemas impuestos por el idioma. Uno de los mecanismos que actualmente se explora en la indización automática es el punto de transición de Goffman, derivada de la ley de Zipf. Utilizando la técnica del punto de transición de Goffman se han realizado diversos estudios exploratorios en el campo de la medicina (Boyce, 1975), ciencia y tecnología (Guedes, 1994), periodismo (Ribeiro, 1974), lingüística (Mamfrim, 1991). Estas investigaciones lidian con textos académicos y científicos con un lenguaje depurado y aséptico. Hasta donde es del conocimiento de los autores de este texto, no se conocen trabajos que centren sus análisis en la lírica musical y menos en la música popular andina peruana.

La *indización* de la lírica o textos de la música popular andina peruana aún no ha sido estudiado ni explorado. El mayor volumen de ese material, en la forma de LPs, CDs, casetes y similares, están reservados y esperando su oportunidad en los depósitos de la Biblioteca Nacional del Perú o en las bibliotecas privadas de etnólogos, folcloristas o coleccionistas interesados. Esa música a pesar de ser reproducida y vendida ilegalmente por ambulantes en las calles de las grande ciudades, especialmente Lima "la horrible", tampoco es del interés de las bibliotecas extranjeras y están ausentes de sus programas de desarrollo de colecciones. A pesar de esa marginación y olvido, se afirma que la espina dorsal de la industria musical en el Perú es la música andina. Una de sus múltiples variedades es la música ayacuchana. La melodía más conocida de ésta variedad musical quizá sea *"Adiós pueblo de Ayacucho"*, un huayno que narra los motivos de la migración de los indígenas desposeídos a las grandes ciudades de la costa. Esta migración es la expresión de la búsqueda de oportunidades para escapar a la pobreza y a la indiferencia de las autoridades que gobiernan el país desde la capital donde se centralizan el poder económico y político.

Una de las muchas y periódicas sublevaciones que ocurren en el Perú contra ese poder económico, político, racista y clasista, que margina y olvida a las poblaciones indígenas andinas, se produjo en la década de los 80s. En medio de ese torbellino, la música ayacuchana pareció alcanzar su mayor expresión y esplendor. Una de sus representantes más destacadas fue la cantante, compositora y folclorista Martina Portocarrero que a través de su voz transformó en canto y esperanza la furia contenida de los oprimidos. En 1987 ofreció un concierto en el Teatro Municipal de Lima, que después, con el nombre de *"Martina Portocarrero en Vivo en el Teatro Municipal",* se convirtió en un casete que era copiado, intercambiado y vendido en todos los rincones del Perú. Su fama surgida con su participación musical en sindicatos, fábricas en huelga, paros estudiantiles, barrios populares, marchas de protesta y mítines políticos partidarios, llegó a

su punto apoteósico con la canción *"Flor de Retama"*, un huayno ayacuchano que no quería ser escuchado por el gobierno "democrático" de Alán García (1985-1990) que lo censuró y prohibió. También la dictadura de Alberto Fujimori (1990-2000) persiguió su canto, pero a pesar de todos esos impedimentos, las masas populares de la costa, sierra y selva del país, seguían escuchando sus músicas y apoyándola. Finalmente, la lucha popular y la izquierda radical de esos años adoptaron y reivindicaron ese huayno hasta transformarlo en himno de la resistencia y la rebelión popular.

Este trabajo tiene dos objetivos: el primero, es analizar la lírica de las músicas de ese casete *"Martina Portocarrero en Vivo en el Teatro Municipal"*. La intención es realizar un análisis bibliométrico de las letras de las canciones transmitidas en ese cuerpo textual. Recordemos que la realización de un concierto, la grabación de un casete o la impresión de un CD-ROM, no es un acto aventurero ni un amontonado amorfo de una serie de composiciones musicales sin conexión ni sentido. Contrariamente a lo que se pueda pensar y percibir, es un trabajo conscientemente planificado y ejecutado para producir el mayor impacto posible en la audiencia y la crítica musical. Es un paciente trabajo de selección, organización y difusión que tiene una estructura temática muy similar a la elaboración de un libro de poemas. De hecho, las letras de las músicas de ese casete son bellísimos poemas indígenas. Por lo tanto, es lícito preguntarse si un análisis bibliométrico basado en la frecuencia de ocurrencias de palabras en el cuerpo textual musical aislaría palabras significativas que podrían ser utilizadas como palabras claves en un proceso de indización automática orientadas a la recuperación de la información musical. Es decir, ¿es posible identificar palabras significativas que podrían ser seleccionadas por el indizador como ítems de pre-coordenación o post-coordenación para la recuperación de la información? ¿Cuáles serían esas palabras mas frecuentemente utilizadas? ¿Esas palabras representarán el sentir y vivir de los indígenas creadores de la música ayacuchana?

El segundo objetivo es medir el tamaño o volumen del vocabulario usado en ese caset por Martina Portocarrero. Sichel (1986) sostiene que si en un determinado texto, el número de las palabras diferentes son ordenadas según la frecuencia de su uso, se genera una distribución de las palabras que generalmente es de la forma de una J invertida. La variable randómica r, relacionada al número de veces que una palabra especifica es usada en el texto, es discreta con origen en $r = 1$. Ese tipo de distribución de palabras fue investigada por Zipf (1949) lo que dio lugar a lo que hoy se conoce como la "Ley de Zipf". A pesar de que este modelo ha sido ampliamente utilizado para estimar el tamaño de los textos lingüísticos en otros campos, no se conocen estudios de aplicaciones del modelo propuesto por Sichel (1986) a textos de

líricas musicales siguiendo la organización según el tamaño-de-las-frecuencias de las palabras usadas. Tal vez la única excepción sea el trabajo de Rousseau & Rousseau (1993) pero ese estudio está restringido al idioma Inglés y al uso del modelo de Lotka, a la formulación de Leimkuhler, a la función de Mandelbrot y a la distribución de Bradford. Sin embargo, Sichel (1986) insiste en que el modelo Gauss Poisson inverso generalizado es el más adecuado para describir la dispersión de la distribución de frecuencias de las palabras usadas en un texto. Un ejemplo de su aplicación a textos bíblicos es dado por Pollatschek & Radday (1980). Así que siguiendo sus propuestas se pretende aplicarlo a la lírica del casete de Martina Portocarrero, pero agrupando las frecuencias del número de palabras que ocurrieron en la lírica textual. Con este segundo objetivo se pretende responder a la siguiente pregunta: ¿la distribución de Sichel, mide adecuadamente el tamaño del vocabulario utilizado en el caset *"Martina Portocarrero en Vivo en el Teatro Municipal"*?

Revisión de la literatura

Debido al estilo especial y particular de cada hablante o escritor así como a la existencia de una multiplicidad de lenguajes, nunca se pensó que la frecuencia de ocurrencias de palabras en un texto tuviera un tipo especial de comportamiento. Sin embargo, tratando de diseñar un eficiente sistema de estenografía para la lengua francesa, Estoup (1908) ya había observado que la frecuencia de las palabras del lenguaje natural sigue leyes estadísticas. Cuando las frecuencias de las palabras son trazadas sobre un papel gráfico en orden descendiente de frecuencia se forma una hipérbola muy similar a lo que hoy día es conocida como la "ley de Zipf". Esta ley es definida como *"un término aplicado a cualquier sistema de clasificación de unidades de tal manera que la proporción de clases con exactamente s unidades es aproximadamente proporcional a s-(1+a) para cualquier constante a > 0. Esta ley es familiar en una variedad de áreas empíricas que incluyen la lingüística, la distribución de rentas personales y la distribución de géneros y especies biológicas"* (Everitt, 1998). El nombre de esta ley es un homenaje a su formulador George Kingsley Zipf, un profesor de filología de la Universidad de Harvard, en los Estados Unidos, que mientras estudiaba lingüística en la Universidad de Berlín, percibió que el lenguaje como fenómeno natural en realidad era una serie de comunicaciones gestuales. Después de una extensiva investigación encontró que *"la longitud de una palabra, lejos de ser un asunto randómico, estaba relacionada a la frecuencia de su uso, de tal manera que cuanto mayor es la frecuencia, menor es la longitud de la palabra"* (Zipf, 1935). Ese investigador observó también que si las palabras de un texto suficientemente extenso fuesen contadas y ordenadas de acuerdo a su rango, la distribución de

las frecuencias resultantes sería inversamente proporcional a la posición que las palabras ocupasen en el rango de frecuencias del texto. En otras palabras, el producto de ambas variables sería una cantidad que podría ser aproximada por una constante, de tal manera que

$$R \times F = C \qquad (1)$$

donde R es el rango, F es la frecuencia de ocurrencia de las palabras y C es una constante. En términos logarítmicos, ésta ecuación se convierte en una línea recta con una inclinación igual a – 1, pero tomando la forma de

$$\log R + \log F = \log C \qquad (2)$$

Zipf (1949) notó también que esta relación era verdadera para varios tipos de objetos, incluyendo ciudades en los Estados Unidos por número de habitantes, libros por número de páginas, número de palabras que ocurren en un texto y géneros biológicos por el número de especies (*Encyclopedia of statistical sciences,* 1982). Sin embargo, el propio Zipf proporcionó más de una fórmula para sus observaciones porque sus intereses académicos fluctuaban entre la lingüística y la demografía. Aunque propuso el principio del mínimo esfuerzo, nunca lo enunció claramente. Uno de los enunciados referidos a este principio afirma que

> "… una persona resolverá sus problemas inmediatos minimizando el trabajo total que debe realizar para resolver al mismo tiempo sus problemas inmediatos y sus posibles problemas futuros"

Este principio parece aplicarse a fenómenos donde existe un mecanismo de libre competencia. Por ejemplo palabras compitiendo con otras palabras por el uso social y almacenamiento en la memoria humana. Sin embargo, la ley de Zipf puede ser establecida también como

> "Si se ordena libremente grupos de competencia en orden de rangos de tal manera que el mayor tenga rango 1, el segundo mayor rango 2, y así sucesivamente. Entonces, el producto del rango y el tamaño de cada grupo es aproximadamente una constante"

En el caso de la lengua inglesa, la palabra "the" es la más comúnmente usada y es usada dos veces más frecuentemente que la palabra "of" colocada en el segundo rango, y así sucesivamente. Sin embargo, esta ley se mantiene verdadera para todos los textos cuyos vocabularios han sido contados en una gran variedad de idiomas incluyendo los idiomas de los indígenas americanos (Simon, 1978). Una evidencia empírica de la ley de Zipf puede ser ilustrada trazando la dispersión de las ocurrencias versus el rango de las palabras sobre

un papel logarítmico de doble escala. Si la ley de Zipf es obedecida, los puntos se alinearán siguiendo una recta igual a – 1.

Una interesante revisión de su significado para el caso de la bibliometría es hecha por Wyllys (1981) y Hertzel (1987). Tague (1990) revisa las ventajas y desventajas de la organización y análisis de los datos por rangos y tamaños de las frecuencias. White & McCain (1989) incluyen comentarios sobre los estudios de la distribución de Zipf en el campo de la bibliometría. Brookes (1984) argumenta que existen dos tipos de distribuciones estadísticas: Gausiana y Zipfiana. En las distribuciones Zipfianas los momentos de la muestra serían determinados por el tamaño de la muestra, por tanto, las técnicas gausianas no serían adecuadas para tratar con ese tipo de distribución, sin embargo, sostiene que la distribución de Zipf es básica para dar cuenta de los fenómenos sociales. Haitun (1986) considera a la ley de Zipf como "la ley cuantitativa fundamental para las actividades humanas". Scarrott (1974) revisa la ley de Zipf con relación al tamaño de las ciudades. Ridley (1982) aplicó esta ley a las palabras provenientes de una entrevista como especificación de una muestra oral. Bender & Gill (1986) aplicaron la ley de Zipf al campo de la genética y Urzúa (2000) ofrece un simple pero eficiente método para probar la ley de Zipf.

En el campo de la bibliometría, existen dos formas de organizar los datos para efectuar un estudio de la ley de Zipf. Una forma es la organización de las palabras por rangos-de-frecuencias. En este tipo de ordenamiento existen tres formas de establecer los rangos cuando se observan palabras con el mismo número de frecuencias (empate de frecuencias): rango mínimo, rango ajustado y rango máximo. Por ejemplo, si se observan que 2 palabras ocurrieron 20 veces, el rango mínimo otorga a las dos palabras el mismo rango. El rango ajustado otorga la media del rango de ambas palabras. El rango máximo otorga el máximo valor del rango observado.

La otra forma de organización de los datos es por el tamaño-de-las frecuencias. Este tipo de organización es abogado por Sichel (1975, 1986) por su parentesco con la ley de Lotka y la ley de Bradford. Una didáctica explicación de la forma y significado de estas formas de organización de los datos son proporcionados por Tague (1990) y Rousseau & Rousseau (1993).

Material y métodos

Como unidades de análisis fueron tomadas cada una de las palabras que aparecieron en las canciones del casete *"Martina Portocarrero en Vivo en el Teatro Municipal"*. Las letras de las músicas fueron transcritas del casete a la forma impresa en papel. Después, usando ésta forma impresa, cada una de las palabras fueron nuevamente transcritas a fichas especiales de 3.5 x 7.5 cm., luego

organizadas alfabéticamente y después clasificadas y contadas las frecuencias con que cada palabra apareció en la lírica textual. Para asegurar homogeneidad en el conteo de las frecuencias resultantes, se adoptaron las convenciones siguientes:

- Una palabra es expresada como una serie de caracteres tipográficos precedida y seguida de espacios en blanco.
- Las palabras unidas por guión fueron tratadas como una sola palabra.
- Las palabras que expresaban formas singulares o plurales fueron contadas como una sola.
- Las palabras fonéticamente diferentes fueron contadas y tratadas como diferentes.
- Se omitieron los nombres de personas y lugares presentes en las letras del texto.
- Se omitieron las palabras que en el texto aparecen en lengua Quechua.
- Se omitieron las palabras que en el texto aparecieron en la forma de números.

La ejecución de cada pieza musical de este casete está intercalada con partes habladas que se refieren al contexto de la música, descripción de la región Andina de donde proceden los huaynos o simplemente sirven como pretexto para prestar homenaje a los que fueron asesinados en la lucha popular. Esas partes habladas de la lírica no fueron contabilizadas en las frecuencias de las palabras, es decir, fueron omitidos del conteo de las frecuencias. El casete estaba compuesto de 13 canciones cuyos títulos son: *Hierba silvestre, Mártires de Uchuraccay, Flor de retama, El hombre, Qué encanto tienen tus ojos, Mamacha de las Mercedes, Lluvia en el alma, De canto a canto, Javier Heraud, Cuna takiraki, Amigo, Agua rosada y Pajonal*. A pesar de que la autoría de esos huaynos son de diversos compositores, la unidad y coherencia del casete es dada por la selección, organización y ejecución intencionada de la lírica.

El análisis se realizó empleando tres formas de organización de las palabras identificadas en la investigación. Primero, usando el *método del rango mínimo* siguiendo los procedimientos establecidos por Pao (1977). Este método ya ha sido experimentado por Urbizagástegui (1999) en el análisis de textos en lengua inglesa así como por Basilio; Braga & Pierotti (1978?) en textos científicos en lengua portuguesa. Respecto a este método, autores como Goffman, citado por Pao (1977) como comunicado a través de una comunicación personal, han afirmado que la ley de Zipf considera solamente las palabras de alta frecuencia de ocurrencia y que esas palabras tienen la tendencia a ocupar posiciones de orden única en la distribución de palabras. En otras palabras, en un determi-

nado texto, dos palabras no pueden tener la misma frecuencia. Esta misma idea es compartida por Pao (1977) quien sugiere también que, ya que Zipf desarrolló dos leyes diferentes -uno para palabras de alta frecuencia y otro para palabras de baja frecuencia- esta ley predice y describe los dos extremos de la distribución de palabras en un determinado texto. Como la distribución de las palabras están posicionadas en dos extremos, talvez sería posible identificar una *región crítica* en la que ocurra la transición de las palabras de alta frecuencia para las palabras de baja frecuencia. Esta llamada "*región crítica*" nuclearia a las palabras más significativas del texto analizado. Para llegar a esa *región crítica* como punto de transición, Pao (1977) comenzó de la ecuación para las palabras de baja frecuencia propuesto por Booth (1967) que las derivó de Zipf (1949):

$$\frac{I_1}{I_n} = \frac{n(n+1)}{2} \tag{3}$$

Substituyendo I_n por 1, en la Ecuación (3) tenemos

$$\frac{I_1}{1} = \frac{n(n+1)}{2} \tag{4}$$

Reordenando la Ecuación (4) se obtiene

$$n^2 + n - 2\,I_1 = 0 \tag{5}$$

La Ecuación (5) es una ecuación cuadrática general, y resolviendo por la raíz resulta en

$$n = \frac{\left(-1 \pm \sqrt{1+8\,I_1}\right)}{2} \tag{6}$$

Pero como solamente estamos interesados en los valores positivos de *n*, podemos considerar solamente,

$$n = \frac{\left(-1 + \sqrt{1+8\,I_1}\right)}{2} \tag{7}$$

donde,

I_1 representa al número de palabras que ocurrieron solamente una vez.

Para alcanzar el primer objetivo se usó esta Ecuación (7) con la cual es posible calcular e identificar las palabras alrededor de la región crítica, i.e. calcular

el punto de transición y la región donde se encontrarían las palabras más significativas del texto lírico del casete de Martina Portocarrero. La distribución de las palabras se organizó siguiendo los rangos mínimos y los rangos máximos. Para la aplicación del primer *método del rango mínimo* se siguió le procedimiento de Pao (1977). Los mecanismos de organización de los datos para la aplicación de este método pueden ser observados en las *Tablas 1 y 2*.

Para la aplicación del segundo *método del rango máximo* se siguió el procedimiento sugerido por Sun; Shaw & Davis (1999). Los mecanismos de organización de los datos para la aplicación de este método pueden ser observados en la *Tabla 4*. Esos autores proponen la siguiente ecuación para estimar la región crítica:

$$n \geq (- 1 + (1 + 4 D)^{1/2}) / 2 \qquad (8)$$

Esta expresión puede aún ser simplificada hasta convertirse en

$$n^{*} = (D)^{1/2} \qquad (9)$$

donde
D es el rango máximo de clasificación de las palabras

Para alcanzar el segundo objetivo, se analizó los datos empleando el *método del tamaño de las frecuencias* usando la propuesta de Sichel (1986). Esta propuesta se refiere a la relación entre el número total de palabras y el número de palabras distintas que aparecen en un texto literario. Esta preocupación se investiga principalmente en el campo de la lingüística computacional con la intención de observar y estimar el número de clases existentes en un texto así como el número de palabras presentes en el vocabulario de un autor y el número de palabras distintas que aparecen en un texto. En un determinado texto con un número total de palabras N (llamado tokens) habrá cierto número de palabras distintas (llamadas tipos). Esta información esta implícita en la distribución de probabilidades de las palabras que suministran el número de veces que palabras únicas son usadas en un texto que contiene N tokens lo que facilita la estimación de lo que se ha convenido en llamar "el vocabulario activo" de un autor. Para este último caso fue usada la distribución Gauss Poisson inversa generalizada propuesta por Sichel (1986). La prueba de ajuste del chi-cuadrado al 0.01 nivel de significancia fue utilizada para evaluar el ajuste de las frecuencias observadas a las frecuencias esperadas. Los datos fueron manipulados con el paquete estadístico SPSS versión 10.0 en ambiente Windows.

En esta distribución, la probabilidad de la primera frecuencia es dada por la siguiente ecuación:

$$\phi(1) = \frac{y\theta}{1-(1-\theta)^{y}} \tag{10}$$

y después, todas las otras probabilidades fueron calculadas usando la siguiente fórmula de recurrencia general,

$$\phi(r+1) = \left(\frac{r-y}{r+1}\right)\theta\ \phi(r) \tag{11}$$

Resultados y discusión

Se contabilizaron un total de 1999 palabras que representaban a 387 palabras diferentes. Esas palabras, ordenadas según la frecuencia descendiente están listadas en la *Tabla 1*. Las palabras están ordenadas desde la primera palabra con 100 ocurrencias hasta las últimas 54 palabras diferentes con una sola ocurrencia cada una al final de la lista. La conjunción "que" fue la palabra que ocurrió mas veces (100), seguida de la preposición denotativa "de" que ocurrió 86 veces, los pronombres personales "me" en su forma dativa o acusativa y "mi" como adjetivo posesivo ocurrieron 54 veces cada uno, y así sucesivamente, los artículos "el", "la", la conjunción copulativa "y", el adverbio de negación "no", la preposición "en", etc. Es obvio que conjunciones, preposiciones, pronombres personales, artículos, adverbios y adjetivos, fueron las palabras de mayor ocurrencia. Diferentemente del idioma inglés en el cual el artículo "the" es el más frecuente, en la lírica andina ayacuchana de Martina Portocarrero, la palabra que se repite con mayor frecuencia es la conjunción "que". En el orden de ocurrencias es seguida por una preposición, pronombres personales y artículos. Es esta una característica del comportamiento del lenguaje español o es un caso particular de manifestación de la lírica musical andina? Es acaso consecuencia del casete ser una compilación de canciones de diferentes autores y por eso muestran ese tipo comportamiento? Para obtener respuestas específicas sobre este punto serian necesarias mayores investigaciones que las establecidas como objetivo de este trabajo. La *Figura 1*, muestra la distribución del número de palabras diferentes observadas según la frecuencia de ocurrencias.

Las mayores ocurrencias se concentraron en 183 palabras diferentes que ocurrieron 2 veces, 54 palabras diferentes que ocurrieron 1 vez, 45 palabras diferentes que ocurrieron 4 veces, 19 palabras que ocurrieron 8 veces, 17 palabras que ocurrieron 3 veces, y 16 palabras que ocurrieron 6 veces cada una y así su-

cesivamente. Se puede observar cómo la distribución de ocurrencias de palabras está fuertemente sesgada a la izquierda. Este tipo de distribución sesgada es una característica típica del volumen de palabras presentes en un texto con una vertiginosa caída en las primeras frecuencias para después mostrar una larga cola de palabras de baja frecuencia de ocurrencias. En esta forma de visualización gráfica las conjunciones, preposiciones, pronombres, artículos y adverbios caen dentro de una larga cola de pequeñas frecuencias de utilización.

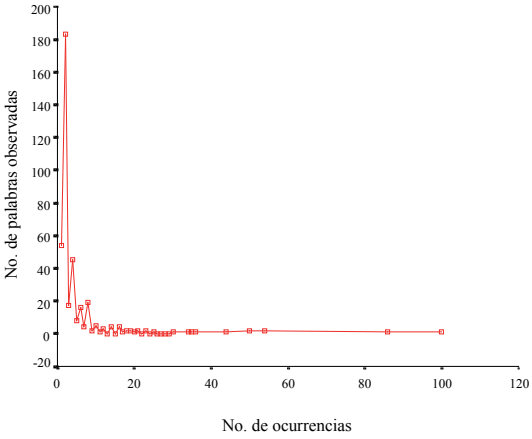

Fig. 1: Dispersión de la ocurrencia de las palabras

La *Tabla 2*, muestra la organización de las palabras por el método de los rangos mínimos de frecuencias de ocurrencias. Es evidente que las conjunciones, preposiciones, pronombres, artículos, adverbios, adjetivos, y los verbos auxiliares, son las palabras que ocurren con mayor frecuencia en el texto y se situaron en el tope de la lista. Pero también es obvio que algunas palabras de gran significado para el texto están situadas siguiendo a esos artículos, pronombres, preposiciones, adjetivos, adverbios, conjunciones, disyunciones y verbos auxiliares. Términos como *Alma, Quiero, Canto, Flor, Arde, Retama, Amarillito, Amigo, llorarás, Negro*, *Viento*, etc. son tópicos en el texto y pueden ser usados apropiadamente como términos de caracterización temática de la lírica del casete de Martina Portocarrero. No en vano, Zipf (1935) también sostenía que las palabras más frecuentes eran más polisémicas que las palabras menos frecuentes, con tendencia a mostrar que esos fonemas poseían también más *homófonas* o palabras similares con diferentes significados.

Sin embargo, una cuestión parece ser difícil de resolver y se relaciona al punto de corte. ¿De las *n* palabras situadas en el tope de la lista, cuáles de ellas podrían ser escogidas como palabras claves? solamente la primera palabra? las dos, tres, cuatro o cinco palabras situadas en los primeros rangos? Cómo determinar el punto de corte? Para responder a esas interrogantes se usó la ecuación (7) y se encontró que el punto de transición fue igual a 9.9043. Esto significa que la región crítica, es decir, el punto donde las palabras de alta frecuencia comienzan a transformarse en baja frecuencia deben ocurrir alrededor de las palabras que ocurren más o menos 10 veces en el texto. Si observamos las palabras mostradas en la *Tabla 2*, hubieron 4 palabras significativas (resaltadas en negritas) que ocurrieron 10 veces *(Amarillito, Amigo, Llorarás, Negro)*. Pero alrededor de ellas están agrupadas otras palabras como *Arde, Retama*, *Viento, Canta, Aroma, Sangre*, que son significativas para el texto y pueden ser usadas como términos de tipificación de la temática de la lírica del casete de Martina Portocarrero. Esto llega a ser más evidente cuando se eliminan las palabras sin contenido semántico, a pesar de que esas palabras son importantes para la estructura gramatical del texto. De hecho, las palabras demasiadas frecuentes no aportan información y por eso se considera que una palabra que aparezca en al menos 80% de los documentos de una determinada colección carece de utilidad para las tareas de recuperación de la información. Estas palabras se consideran como vacías y se eliminan durante el proceso de análisis automático del texto para evitar que puedan ser consideradas como potenciales términos de indización" (Moreira González, 2002).

Como la intención de este estudio es aislar las palabras mas significativas para elegir términos adecuados a un proceso de indización, tal vez esas palabras aparecerán más claramente si usamos los artículos, pronombres, preposiciones, adjetivos, adverbios, conjunciones, disyunciones, los verbos auxiliares, y otras palabras sin significado semántico, como palabras de contención, es decir, si las ignoramos. Con la intención de explorar esta idea, fue construida la *Tabla 3* que muestra el efecto de esa omisión. Ahora la tabla muestra más claramente en el tope de la lista las palabras que pueden ser usadas como términos de caracterización temática o palabras claves en un proceso de indización automática del texto, especialmente aquellas colocadas hasta el quinto rango. Estas palabras claves son: *Alma, Quiero, Canto, Flor, Arde, Haberlo, Retama, Amarillito, Amigo, Llorarás, Negro, Viento.* Nuevamente, a los datos de esta *Tabla 3*, se aplicó la fórmula de transición de Goffman (ecuación 7) pero no se encontró variaciones. Es decir, las palabras más significativas están dispersas alrededor de aquellas que ocurrieron 9.9043 ± 10 veces. Estas palabras fueron *Amarillito, Amigo, Llorarás,* y *Negro*. Pero si se sube o baja una o dos palabras sobre y debajo del punto de transición se pueden incorporar también los términos *Retama* y *Viento*.

Es claro que la lírica textual hace referencia a la retama que amarilla en los caminos, una de las canciones se llama precisamente "Flor de Retama". A los amigos que se perdieron en la lucha y que se continuarán perdiendo en las batallas, por eso se llora o se vaticina que se llorarán esas pérdidas. La música andina peruana esta repleta de referencias a la clareza o la negritud de la noche y es un símbolo que se repite en muchas canciones. Varias de las canciones analizadas (Flor de retama, Mamacha de las Mercedes, El hombre, Javier Heraud, etc.) tratan centralmente temas tales como las penas y sufrimientos por un hijo asesinado, los conflictos sociales, la lucha armada, la resistencia a la opresión, la memoria de los que han muerto, la solidaridad, etc. La selección de palabras extraídas evidentemente no representa esos temas de manera equilibrada con respecto a la fuente, incluso muchos de ellos quedan excluidos. Entre los términos presentes en las letras pero no seleccionadas como palabras claves están: pueblo, dolor, sufrir, diferentes formas del verbo matar, tierra, triste, vida, hijo, hermano, guerrillero, tirano, justicia, libertad, deseo, mundo, recuerdo, etc.

Recientemente, Sun; Shaw & Davis (1999) propusieron un nuevo modelo para estimar la frecuencia de ocurrencias de palabras en un texto así como para encontrar la región de transición de las palabras de alta frecuencia hacia las palabras de baja frecuencia. Las autoras usan el método del rango máximo para asignar los rangos a las palabras con las mismas frecuencias de ocurrencias (ver Tabla 4). Aparentemente éste método es mas simple de calcular que el método de Pao (1977). Para hacer un estudio comparativo entre ambos métodos, el método propuesto por Sun; Shaw & Davis (1999) fue aplicado a la lírica del casete de Martina Portocarrero. Con este método, la región de transición de las palabras de altas frecuencias para las palabras de bajas frecuencias fueron estimadas como aconteciendo en aquellas palabras que ocurrieron 19.67 ± 20 veces. Como se puede observar en el listado mostrado en la Tabla 1, esta fue la palabra *Si*. Alrededor de ésta palabra se localizan otras con valioso significado para la indización de la lírica. Palabras como *Alma, Quiero, Canto Flor*, pueden ser candidatas a esa opción.

Nuevamente, para explorar los efectos de la retirada de las palabras sin contenido semántico significativo para un proceso de indización, tales como artículos, pronombres, preposiciones, adjetivos, adverbios, conjunciones, disyunciones, y verbos auxiliares, fueron retirados de la tabla de frecuencia de ocurrencias de las palabras. La *Tabla 5* muestra los efectos de esa omisión. Ahora la tabla muestra más claramente en el tope de la lista las palabras que pueden ser usadas como términos o palabras claves en un proceso de indización automática del texto, especialmente aquellas colocadas hasta el quinto rango. Estas palabras claves son: *Alma, Quiero, Canto, Flor, Arde, Retama, Amarillito, Amigo, Llorarás, Negro, Viento*. Para esta situación la región de transición

fue calculada como ocurriendo a partir de las palabras que ocurrieron 18.17 ± 18 veces. Esta región es señalada por la palabra *Voy* (resaltada en negritas). Sin embargo, alrededor de esta palabra se agrupan otras de gran significación para la temática del casete. Estas palabras son: *Alma, Quiero, Canto, Flor, Arde, Retama, Amarillito, Amigo, Llorarás, Negro, Viento*. La *Tabla 6*, resume las palabras candidatas para ser seleccionadas como términos de indización por ambos métodos (rangos mínimos y rangos máximos). Como se puede observar, no parecen existir diferencias significativas entre ambos métodos de identificación de palabras claves para la caracterización automática de un texto.

Tabla 6: Palabras de indización seleccionadas por ambos métodos

Método de Pao (1977)	Método de Sun; Shaw & Davis (1999)
Rango mínimo	Rango máximo
Amarillito	Alma
Amigo	Quiero
Llorarás	Canto
Negro	Flor
Arde	Arde
Retama	Retama
Viento	Amarillito
Canta	Amigo
Aroma	Llorarás
Sangre	Negro
	Viento

Lamentablemente, este casete no forma parte de la colección musical de ninguna biblioteca nacional ni extranjera, por lo tanto, no es posible hacer una comparación de los encabezamientos de materias ni de las palabras claves usadas por algunas bibliotecas para describirlo. Es evidente que la Ley de Zipf, a través del punto de transición, identifica claramente el contenido textual de la lírica, pero las palabras identificadas no parecen ser adecuadas para operar como palabras claves orientadas a la recuperación de la información. Un estudioso de la música andina, difícilmente buscaría recuperar este tipo de lírica usando palabras claves como *Amarillito, Alma, Amigo, Quiero,* o cualquiera de las otras palabras listadas en la *Tabla 6*. Lo que sí parece claro es que estas palabras

tipifican adecuadamente la temática de la lírica del casete. Para buscar un padrón de comparación, se realizó una búsqueda en los registros de la Biblioteca del Congreso Americano. Como se sabe, en problemas relacionados a los encabezamientos de materias e indización, esta biblioteca es usada como modelo a seguir por bibliotecas del mundo entero. Se encontró que, para casos similares al casete de Martina Portocarrero, la Biblioteca del Congreso Americano ha usado los siguientes encabezamientos de materias:

Canciones folclóricas, Quechuas – Perú – Ayacucho
Música folclórica – Perú – Ayacucho
Música popular – Perú – Ayacucho
Indígenas de América del Sur – Perú – Ayacucho – Música
Huaynos – Perú – Ayacucho
Ayacucho (Perú) – Canciones y músicas

Estas palabras de tipificación de la música popular andina, proceden de un contexto cultural externo a la lírica de las músicas ejecutadas y parecen orientadas a describir la procedencia regional de la lírica y expresan una visión folclorista de la música andina pero no tienen ninguna semejanza con las palabras identificadas a través del punto de transición. En una relación de hegemonía y dependencia, esas palabras claves representan la visión oficial hegemónica de los que dominan, nominan y denominan las prácticas culturales de los hegemonizados. El punto de transición de Goffman, derivada de la ley de Zipf, parece identificar los motivos del canto y encanto de los indígenas andinos aunque las palabras que tipifican esos temas y motivos no son usados como términos de indización en la cultura oficializada de la academia representada por las palabras-claves y listas de encabezamientos de materias usados en la práctica bibliotecológica contemporánea. En este punto es necesario recordar que las propiedades formales de los términos de recuperación e indización solo entregan su sentido ideológico si se les relaciona con las condiciones sociales de su producción, es decir, con la posición que ocupan los elaboradores de esos términos en el campo de la producción intelectual y con el mercado en el que son producidos. Por más legítimo que sea tratar los encabezamientos de materias, tesauros y similares como meros instrumentos de recuperación de información hay que mostrar que la nominación implícita en sus términos implica también relaciones de poder simbólicos donde se actualizan las relaciones de clase entre los grupos dominantes y dominados. Como todo discurso, ese lenguaje que los especialistas en indización y tesauros producen y reproducen mediante una alteración sistemática del lenguaje común, son el

> *"producto de un compromiso entre un interés expresivo y una censura constituida por la misma estructura del campo en el que ese discurso se construye y circula. Mas o menos conseguido según la competencia especifica de cada productor, esa formación de compromiso, … es el producto de estrategias de eufemización … esas estrategias tienden a asegurar la satisfacción del interés expresivo … en los límites de la estructura de las posibilidades de beneficio material o simbólico que las diferentes formas de discurso pueden producir a los diferentes productores en función de la posición que ocupan … en la estructura de la distribución del capital especifico que esta en juego" (*Bourdieu, 1985:109) en cada campo académico.

De modo que esa censura estructural se ejerce a través de las sanciones que, dicho campo, funcionando como un mercado donde se constituyen los precios de las diferentes modalidades de expresión, se impone a cualquier productor de bienes simbólicos y se condena a los ocupantes de posiciones dominadas a la alternativa del silencio. Este es el caso de la música popular indígena peruana. Es más, esos discursos esotéricos producidos por los especialistas del lenguaje controlado,

> *"experimentan una especie de universalización automática y dejan de ser exclusivamente palabras de dominantes y de dominados en el interior de un campo especifico para convertirse en palabras válidas para todos los dominantes y todos los dominados"* (Bourdieu, 1985:15)

No es posible olvidar que cualquier sociedad esta constituida de clases sociales en lucha por la apropiación de diferentes capitales, contribuyendo con esa lucha, a las relaciones de fuerza que dan sentido a la perpetuación del orden social o a su cuestionamiento. Esta lucha por los sentidos y denominaciones lingüísticas se incluye en lo que Bourdieu (2003) denomina "teoría de la dominación simbólica". Según esa teoría, la realidad social expresa un conjunto de relaciones de fuerza entre clases históricamente en lucha permanente unas con las otras. En ese contexto de luchas, las expresiones lingüísticas de nominación y denominación como los expresados en los listados de encabezamientos de materias son apenas mecanismos de dominación legitimizados. La legitimidad lingüística es apenas la expresión de la cualidad de lo que es aceptado y reconocido como legítimo por los miembros de una comunidad sea esta académica o científica o de otro tipo. En este caso, la comunidad de los bibliotecarios y científicos de la información que como actores sociales producen la legitimidad de los vocabularios controlados, encabezamientos de materias, términos de indización, palabras-claves y tesauros para que su "competencia profesional" sea conocida y reconocida. De esa manera producen también "los arbitrarios culturales legitimizados" que, como en este caso, la lista de encabezamientos de materias producida por la biblioteca del congreso americano, lo expresan perfectamente. Esa forma de imponer una manera legítima de ver

el mundo andino y sus prácticas culturales musicales es motivo de lucha por la apropiación de la expresión simbólica expresada en sus canciones. Eso explica el divorcio entre la temática textual de la lírica del casete de Martina Portocarrero y las formas de nominación expresadas en las palabras-claves y encabezamientos de materias utilizadas por los bibliotecarios y las bibliotecas del mundo entero. Los intelectuales orgánicos de los grupos hegemónicos, tienen el poder de la codificación y producen esquemas de percepción y términos para designar una realidad que no es el de ellos. Esos términos producidos imperceptiblemente entran en el lenguaje cotidiano y parecen disponer de la fuerza de la evidencia certera de la nominación. De modo que esos vocabularios controlados o listados de encabezamientos no son neutros sino que encierran una concepción y una visión del mundo social y sirven para reforzar y legitimar la dominación hegemónica. Un tesauro que represente la visión y preocupación de los oprimidos, en este caso los indígenas andinos peruanos, aun esta por pensarse y hacerse.

Por otro lado, si en un determinado texto, el número de las palabras diferentes son ordenadas según la frecuencia de su uso, se genera una distribución de la frecuencia de usos de las palabras en ese texto, que generalmente es de la forma de una J invertida. La variable randómica r, relacionada al número de veces que una palabra especifica es usada en el texto, es discreta con origen en $r = 1$. Ese tipo de comportamiento de las palabras fue investigado por Zipf (1949) lo que dio lugar a lo que hoy se conoce como la "Ley de Zipf". A pesar de que este modelo ha sido ampliamente utilizado para estimar el tamaño de los textos lingüísticos en otros campos, no se conocen estudios de aplicación de la ley de Zipf a textos de líricas musicales siguiendo el método del tamaño-de las frecuencias de las palabras usadas. Tal vez la única excepción sea el trabajo de Rousseau & Rousseau (1993), pero ese estudio está restringido al idioma Inglés y al uso del modelo de Lotka, a la formulación de Leimkuhler, a la función de Mandelbrot y a la distribución de Bradford. Sin embargo, Sichel (1986) insiste en que el modelo Gauss Poisson inverso generalizado es el más adecuado para describir la dispersión de la distribución de frecuencias de las palabras usadas en un texto. Un ejemplo de su aplicación a textos bíblicos es dado por Pollatschek & Radday (1980). Así que siguiendo sus propuestas este modelo fue aplicado a la lírica del casete de Martina Portocarrero pero agrupando las frecuencias del número de palabras que ocurrieron en la lírica textual. La *Figura 2* muestra la dispersión de las palabras observadas versus el número de ocurrencias de las frecuencias agrupadas.

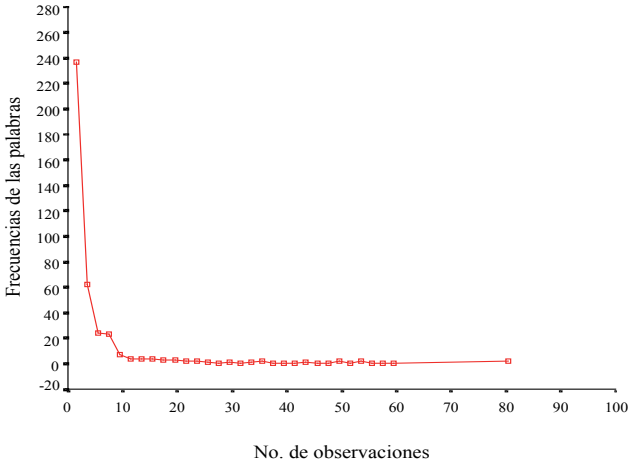

Fig. 2: Dispersión de la ocurrencia de las palabras

Se puede observar que las ocurrencias de las palabras están fuertemente sesgadas a la izquierda formando una especie de J inversa con una larga cola de palabras con pequeñas ocurrencias. La prueba del chi-cuadrado fue usada para evaluar el ajuste de los datos observados y esperados. Con una media estimada de 4.8488 palabras y la proporción de la primera frecuencia agrupada igual a 0.6124 se estimó $\hat{y} = 0.58228$ y $\hat{\theta} = 0.99294$. Con esos valores se estimaron las frecuencias esperadas mostradas en la *Tabla 7*.

Tabla 7: Valores observados y esperados

Frecuencias observadas	Frecuencias esperadas
237	237.00
62	49.15
24	23.06
23	13.84
7	9.39
8	12.15
7	7.64
7	7.43
6	11.76
6+	4.71
387	376.13

Al nivel de significancia de 0.01 y con 7 grados de libertad, el chi-cuadrado calculado fue igual a 14.74 menor que el valor crítico de 18.4753. Por lo tanto, se concluye que la distribución de frecuencias de las palabras de la lírica del casete de Martina Portocarrero se ajusta adecuadamente a la distribución Gauss Poisson inversa generalizada. En otras palabras, el "el vocabulario activo" de la folclorista Martina Portocarrero sigue una distribución Gauss Poisson inversa generalizada. La *Figura 3* muestra el trazado de la distribución de las frecuencias observadas y esperadas en la que se puede observar casi una perfecta aproximación de los puntos de dispersión.

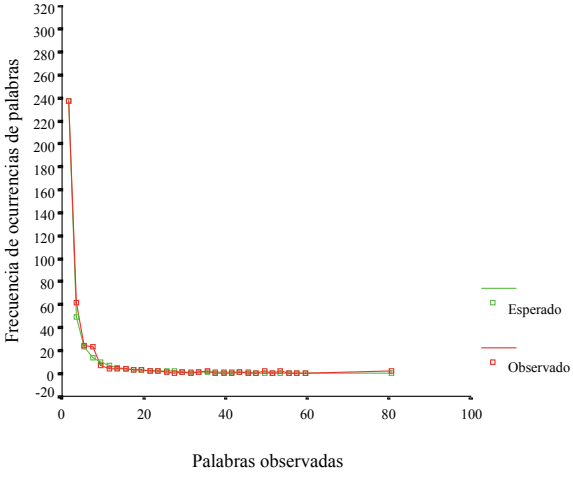

Fig. 3: Dispersión de las frecuencias observadas y esperadas

Conclusiones

La relevancia del análisis de la distribución de frecuencias de palabras en un texto para el campo de la Bibliotecología y Ciencia de la Información ha resultado de su aplicación al contenido de documentos de carácter científico o de sus resúmenes procedentes de bases de datos bibliográficas y de texto completo, en los que aparentemente el lenguaje formal utilizado es más preciso por su procedencia de la ciencia oficial hegemónica. Esa característica ha permitido que los términos identificados sean representativos del contenido de los documentos y puedan ser utilizados como descriptores en tesauros u otros tipos de lenguajes de búsqueda o vocabularios controlados. La aplicación del punto de transición de Goffman, derivada de la ley de Zipf, a obras literarias

que no son académicas (en este caso, música andina peruana), en las que sus contenidos semánticos son metafóricos parecen aportar mejores resultados para la antropología, la lingüística, la sociología, la etnología, la literatura, etc. que para la Bibliotecología y Ciencia de la Información. Esas áreas deberían explorar las ventajas ofrecidas por la ley de Zipf y, en especial, el punto de transición de Goffman.

En la lírica textual del casete *"Martina Portocarrero en Vivo en el Teatro Municipal",* fueron encontradas 387 palabras diferentes que ocurrieron un total de 1999 veces. De este total de palabras identificadas y mediante el método de organización de las palabras por los rangos mínimos y máximos, fueron aisladas hasta 11 palabras que aparentemente pueden servir como palabras claves para la caracterización temática de esa música ayacuchana. El aislamiento de las palabras claves por esos dos métodos, no parecen producir diferencias significativas. En esta investigación, ambos métodos identificaron casi las mismas palabras con pequeñas variaciones no significativas. Se observó coincidencias en el 80% de las palabras escogidas. Sin embargo, las palabras identificadas, aunque tipifican muy bien la temática y el contenido de la lírica textual, no son adecuadas para operar como palabras claves orientadas a la recuperación de la información. Las palabras claves empleadas para la recuperación, como ejemplificadas por el uso común en los encabezamientos de materias de la Biblioteca del Congreso Americano, parecen representar la visión oficial de la academia hegemonizante en relación a la cultura hegemonizada, en este caso, la lírica de los indígenas peruanos. Buscándose una explicación para esta visión hegemonizante, se encontró que la construcción de los encabezamientos de materias, tesauros, palabras claves y similares, están ligadas a un mercado lingüístico, que al igual que el mercado económico (donde hay monopolios, relaciones de fuerza objetivas que hacen que los productores y sus productos no sean todos iguales), expresan relaciones de fuerza y poder. De esa forma, en el mercado lingüístico de la bibliotecología y la ciencia de la información expresadas como la construcción de encabezamientos de materias, tesauros y similares vocabularios orientados a la recuperación de la información y que tipifican ciertos dominios, también existe una forma de determinación de precios que hacen que todos los productos lingüísticos no sean iguales y tengan precios diferentes. Esos precios como formas de expresión simbólicas, se muestran por su presencia o ausencia en esos vocabularios controlados siendo lo más común que las expresiones de las culturas dominadas sean condenadas al silencio. Este es el caso de la música andina ayacuchana.

En el caso de la distribución de las palabras según el tamaño-de-las-frecuencias de ocurrencias, el modelo Gauss-Poisson inverso generalizado, y la prueba del chi-cuadrado fueron usados para evaluar el ajuste de los datos observados

a los datos esperados. Al 0.01 nivel de significancia y con 7 grados de libertad, se verificó que el chi-cuadrado calculado fue igual a 14.74 menor que el valor crítico de 18.4753. Por lo tanto, se concluyó que "el vocabulario activo" de la folclorista Martina Portocarrero es adecuadamente modelada y prevista por la distribución Gauss Poisson inversa generalizada.

Referencias bibliográficas

Basilio, Margarida; Braga, Liliam Maria & Pierotti, Maria de Lourdes Carvalho (1978). *Estrutura de textos científicos em língua portuguêsa: Estudo bibliométrico-linguístico*. 22 folhas datilografadas.

Bender, M. L. & Gill, Pritmohinder (1986). "The genetic code and Zipf's law". En *Current anthropology,* 27(3) ; pp.280-283.

Booth, A. D. A (1967) "Law of ocurrences for words of low frequency". En *Information and Control*, 10(4); pp.388-393.

Bourdieu, Pierre (1985). *¿Qué significa hablar?: Economía de los intercambios lingüísticos*. Madrid: Akal.

Bourdieu, Pierre (1984). *Sociología y cultura.* México, D. F.: Grijalbo.

Bourdieu, Pierre (2003). *Primeiras lições sobre a sociología de P. Bourdieu*. Petrópolis : Editora Vozes.

Boyce, Bert (1975). "Automatic and manual indexing performance in a small file of medical literature". En *Bulletin of the Medical Library Association*, 63(4); pp.378-385.

Brookes, B. C. (1984) "Towards informetrics: Haitun, Laplace, Zipf, Bradford and the Alvey Program". En *Journal of Documentation*, 40(2); pp.120-143.

Bruzinga, Graciane Silva; Maculan, Benildes Coura Moreira dos Santos & Lima, Gercina Ângela Borém de Oliveira (2007). "Indexação automática e semântica: estudo da análise do conteúdo de teses e dissertações". En VIII ENANCIB – Encontro Nacional de Pesquisa em Ciência da Informação 28 a 31 de outubro de 2007, Bahia, Brasil.

Kotz, Samuel Norman Johnson (associate editor) and Campbell B. Read (1982) *Encyclopedia of statistical sciences* New York : Wiley, v. 9; pp. 674-676.

Estoup, J. B (1908). *Gammes sténographiques : recuil de textes choisis pour l'acquisition méthodique de la vitesse*. Paris : Institut stenographique.

Everitt, Brian (1998). *The Cambridge dictionary of statistics*. New York : Cambridge University Press.

Guedes, Vânia Lisbõa da Silveira (1994). "Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos". En *Ciencia da Informação*, 23, 3; pp. 318-326.

Haitun, S. D. (1986). "Stationary scientometrics distributions: Part III: The role of the Zipf distribution". En *Scientometrics*, 9 (3-4); pp.145-164.

Hertzel, Dorothy H. (1987) "History of the development of ideas in bibliometrics". En *Enciclopedia of Library and Information Science*. New York : M. Dekker, 42; pp. 146-219.

Mamfrim, Flavia Pereira Braga (1991). "Representação de conteúdo via indexação automática em textos integrais de língua portuguesa". Em *Ciência da Informação,* 20 (2); pp. 191-203.

Moreira González, José Antonio (2002). "Aplicaciones al análisis automático del contenido provenientes de la teoría matemática de la información". En *Anales de Documentación*, 5 ; pp.273-286.

Pao, Miranda Lee (1977). "Automatic indexing based on Goffman's transition of word ocurrences". En *American Society for Information Science.* Meeting (40th : 1977 : Chicago, Ill.) Information Management in the 80's : proceedings of the ASIS annual meeting 1977. Volume 14 : 40th annual meeting, Chicago, Illinois, September 26-October 1, 1977 / Bernard M. Fry, compiler, Clayton A. With Plains, N. Y. : Knowledge Industry Publications for American Society for Information Science, c1977.

Pollatschek, M. & Radday, Y. T. (1980). "Vocabulary richness and concentration in Hebrew biblical literatura". En *Bulletin Association for Literary and Linguistic Computing*, 8; pp. 217-231.

Ridley, Dennis R. (1982). "Zipf's law in transcribed speech". En *Psychological research*, 44(1); pp.97-103.

Rousseau, R. & Rousseau, S. (1993). "Informetric distributions: a tutorial review". En *Canadian Journal of Information and Library Science*, 18 ; pp.51-63.

Scarrott, Gordon (1974)). "Will Zipf join Gauss?". En *New Scientist*, 62(898); pp.402-404.

Sichel, H. S. (1975). "On a distribution law for word frequencies". En *Journal of the American Statistical Association*, 70 (351); pp.542-547.

Sichel, H. S. (1986). "Word frequency distributions and type-token characteristics". En *Mathematical Scientist*, 11; pp.45-72.

Simon, Herbert A. (1978). *The sizes of things. In: Statistics : a guide to the unknown.* San Francisco: Holden Day.

Sun, Qinglan; Shaw, Debora & Davis, Charles H. (1999). "A model for estimating the ocurrence of same-frequency words and the boundry between high and low frequency words in texts". En *Journal of the American Society for Information Science*, 50(3) pp. 280-286.

Tague, Jean (1990). "Ranks and sizes: some complementarities and contrasts". En *Journal of Information Science,* 16 ; pp.29-35.

Urbizagástegui Alvarado, Rubén (1999). "Las Posibilidades de la ley de Zipf en la indización automática". En *B3: Revista Electrónica de Bibliotecología* http://www.geocities.com/ResearchTriangle/2851).

Urzúa, Carlos M. (2000). "A simple and efficient test for Zipf's law". En *Economics letters*, 66 ; pp.257-260.

White, Howard D. & McCain, Katherine W. (1989). "Bibliometrics". En *Annual Review of Information Science and Technology*, 24; pp. 119-186.

Wyllys, R. E. (1981). "Empirical and theoretical bases of Zipf's law". En *Library Trends*, 30; pp.53-64.

Zipf, G. K. (1949). *Human behaviour and the principle of least effort*. Cambridge: Addison-Wesley.

Zipf, G. K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. Cambridge : MIT Press.

Tabla 1: Distribución de frecuencias de las palabras del caset "Martina Portocarrero en vivo en el Teatro Municipal"

No. de palabras	No. de ocurrencias	Palabras
1	100	Que
1	86	De
2	54	Me, Mi(s)
2	50	El, La(s)
1	44	Y
1	36	No
1	35	En
1	34	Lo(s)
1	30	A
1	25	Con
2	23	Yo, Te
2	21	Alma, Quiero
1	20	Si
2	19	Ay, Se
2	18	Tu(s), Voy
1	17	Como
4	16	Canto, Flor, Para, Por
4	14	Al, Arde, Todo, Un
3	12	Haberlo, Su(s), Tiene(s)
1	11	Retama
5	10	Amarillito, Amigo, Del, Llorarás, Negro
2	9	Es, Viento
19	8	Acaso, Aquí, Canta, Cuando, Esta, Están, Negra, Ni, Pajonal, Pasa, Recuerdo, Serás, Querido, Sin, Solo, Tan, Ver, Viday, Ya
4	7	Aroma, Mas, Sangre, Voz
16	6	Adorado, Amarillando, Andate, Brindar, Corazón(es), Donde, Eres, Lado, Mañana, Olvidar, Otros, Pero, Quieres, Quererte, Ser, Tanto(s)
8	5	Capullo, Lloras, Mío, Ojos, Porque, Puño, Silvestre, Yerba
45	4	(45 palabras diferentes)
17	3	(17 palabras diferentes)
183	2	(183 palabras diferentes)
54	1	(54 palabras diferentes)

Total de palabras = 1999
Total de palabras diferentes = 387

Tabla 2: Distribución de las frecuencias de ocurrencias de las palabras de acuerdo al rango mínimo

Rango	No. de ocurrencias	Palabras
R	F	C
1	100	Que
2	86	De
3	54	Me
	54	Mi(s)
4	50	El
	50	La(s)
5	44	Y
6	36	No
7	35	En
8	34	Lo(s)
9	30	A
10	25	Con
11	23	Yo
	23	Te
12	21	Alma
	21	Quiero
13	20	Si
14	19	Ay
	19	Se
15	18	Tu(s)
	18	Voy
16	17	Como
17	16	Canto
	16	Flor
	16	Para
	16	Por
18	14	Al
	14	Arde
	14	Todo
	14	Un
19	12	Haberlo
	12	Su(s)
	12	Tiene(s)
20	11	Retama
21	10	***Amarillito***
	10	***Amigo***
	10	***Del***
	10	***Llorarás***
	10	***Negro***

22	9	Es
	9	Viento
23	8	19 palabras diferentes
24	7	4 palabras diferentes
25	6	16 palabras diferentes
26	5	8 palabras diferentes
27	4	45 palabras diferentes
28	3	17 palabras diferentes
29	2	183 palabras diferentes
30	1	54 palabra diferentes

Tabla 3: Distribución de las palabras con contenido semántico por rango de frecuencias

Rango r	Frecuencia f	Palabras c
1	21	Alma
	21	Quiero
2	18	Voy
3	16	Canto
	16	Flor
4	14	Arde
5	12	Haberlo
	12	Tiene(s)
6	11	***Retama***
7	10	***Amarillito***
	10	***Amigo***
	10	***Llorarás***
	10	***Negro***
8	9	Es
	9	***Viento***
9	8	Acaso, Canta, Esta, Están, Negra, Pajonal, Pasa, Recuerdo, Serás, Querido, Ver, Viday
10	7	Aroma, Sangre, Voz
11	6	Adorado, Amarillando, Ándate, Brindar, Corazón(es), Eres Lado, Mañana, Olvidar, Quieres, Quererte, Ser
12	5	Capullo, Lloras, Ojos, Puño, Silvestre, Yerba
13	4	Acompañarme, Amor, Alegría, Camino, Caprichos, Ciego, Cielo, Cositas, Deseo, Dinamita, Envano, Florecida, Graduó, Guerrillero, Hermano, Hombre, Lejos, Llorando, Matando, Matan, Miran, Mundo, Perfumar, Pólvora, Pueblo, Puro, Quieras, Quitan, Ruego, Sentido, Sufrir,

		Tienen, Tierra, Triste, Vaya, Vida
14	3	Agua, Desdicha, Encanto, Falso, Gloria, Gorrión(nes), Hay, Montaña, Puna, Riqueza, Rosada, Tormento, Solitita
15	2	Aborrezcas (y otras183 palabras diferentes)
16	1	(54 palabras diferentes)

Tabla 4: Frecuencia de las palabras por rango máximo de clasificación (Propuesto por Sun; Shaw & Davis, 1999)

Palabras P	Orden O	Rango R	Ocurrencias F	Multiplicación R x F
1	1	1	100	100
1	2	2	86	172
2	3-4	4	54	216
2	5-6	6	50	300
1	7	7	44	308
1	8	8	36	288
1	9	9	35	315
1	10	10	34	340
1	11	11	30	330
1	12	12	25	300
2	13-14	14	23	322
2	15-16	16	21	336
1	17	17	20	340
2	18-19	19	19	361
2	20-21	21	18	378
1	22	22	17	374
4	23-26	26	16	416
4	27-30	30	14	420
3	31-33	33	12	396
1	34	34	11	374
5	35-39	39	10	390
2	40-41	41	9	369
19	42-60	60	8	480
4	61-64	64	7	448
16	65-80	80	6	480
8	81-88	88	5	440
45	89-133	133	4	532
17	134-150	150	3	450
183	151-333	333	2	666
54	334-387	387	1	387

Tabla 5: Distribución de las palabras con contenido semántico por rango máximo de frecuencias

Orden	Rango	Frecuencia Ocurrencia	Palabras
1-2	2	21	Alma, Quiero
3	3	18	***Voy***
4-5	5	16	Canto, Flor
6	6	14	Arde
7-8	8	12	Haberlo, Tiene(s)
9	9	11	Retama
10-13	13	10	Amarillito, Amigo, Llorarás, Negro
14	14	9	Viento
15-23	23	8	Acaso, Canta, Negra, Pajonal, Pasa, Recuerdo, Querido, Ver, Viday
24-26	26	7	Aroma, Sangre, Voz
27-38	38	6	Adorado, Amarillando, Andate, Brindar, Corazón(es), Eres Lado, Mañana, Olvidar, Quieres, Quererte, Ser
39-44	44	5	Capullo, Lloras, Ojos, Puño, Silvestre, Yerba
45-80	80	4	Acompañarme, Amor, Alegría, Camino, Caprichos, Ciego, Cielo, Cositas, Deseo, Dinamita, Envano, Florecida, Graduó, Guerrillero, Hermano, Hombre, Lejos, Llorando, Matando, Matan, Miran, Mundo, Perfumar, Pólvora, Pueblo, Puro, Quieras, Quitan, Ruego, Sentido, Sufrir, Tienen, Tierra, Triste, Vaya, Vida
81-93	93	3	Agua, Desdicha, Encanto, Falso, Gloria, Gorrión(nes), Hay, Montaña, Puna, Riqueza, Rosada, Tormento, Solitita
94-276	276	2	Aborrescas, y otras 183 palabras diferentes
277-330	330	1	(54 palabras diferentes)