anales de psicología, 2014, vol. 30, nº 3 (octubre), 1146-1150
http://dx.doi.org/10.6018/analesps.30.3.154931
© Copyright 2014: Servicio de Publicaciones de la Universidad de Murcia. Murcia (España)
ISSN edición impresa: 0212-9728. ISSN edición web (http://revistas.um.es/analesps): 1695-2294

Análisis de contenido y lingüística computacional: su rapidez, confiabilidad y perspectivas

Brenda Lía Chávez[1]* y Jorge Martín Yamamoto[2]

[1] (1) Pontificia Universidad Católica del Perú. Investigadora del Grupo de Investigación en Bienestar, Cultura y Desarrollo; Departamento de Psicología. (2) UNOPS Perú.
[2] (1) Pontificia Universidad Católica del Perú. Profesor Asociado y Coordinador del Grupo de Investigación en Bienestar, Cultura y Desarrollo; Departamento de Psicología. (2) Universidad de Bath. Investigador visitante; Departamento de Ciencias Políticas y Sociales. (3) B y P Bienestar y Productividad, Perú.

Resumen: El análisis de contenido es una técnica que convierte las respuestas abiertas de entrevistas en categorías. Este proceso es de gran utilidad dado que define las categorías de un estudio sobre la base de la percepción de la muestra, evitando la imposición de categorías creadas por el investigador. Sin embargo, este tipo de análisis conlleva un alto costo de tiempo, recursos y personal especializado. Programas como el ATLAS.ti o el NVivo no constituyen una solución eficaz ni eficiente. Los nuevos programas basados en lingüística computacional ofrecen un escenario diferente, dado que el programa "entiende e interpreta" las categorías. Para comprobar su eficacia y eficiencia se compara un análisis de contenido hecho por expertos con el análisis utilizando el programa SPSS *Text Analytics for Surveys* (TA). Se concluye que bajo la supervisión de un investigador especializado, siguiendo ciertos pasos de afinamiento de la extracción, el TA permite un ahorro de tiempo importante, una mayor confiabilidad y abre las posibilidades para análisis cualitativos con muestras grandes.
Palabras clave: análisis de contenido; análisis cualitativo; categorización; investigación émica; lingüística computacional; text analytics.

Title: Content analysis and computational linguistics: its quickness, reliability and perspectives.
Abstract: Content analysis is a technique that converts open-ended responses into categories. This process is of great value since it defines the categories of a study based on the perception of the sample, avoiding imposed categories created by the researcher. However, this type of analysis involves extensive use of time, resources, and expertise. Programs such as ATLAS.ti or NVivo do not constitute an effective nor efficient solution. New software based on computational linguistics offers a different scenario, as it allows the "understanding and interpretation" of categories. In order to prove its effectiveness and efficiency, content analysis made by experts is compared with analysis using SPSS Text Analytics for Surveys (TA). We conclude that under the supervision of a specialized researcher, TA allows for an important time saving, increased reliability, and opens up possibilities for qualitative analysis of large samples.
Key words: content Analysis; qualitative analysis; categorization; emic research; computational linguistics; text analytics.

Introducción

Los beneficios del análisis cualitativo como punto de partida en una secuencia de investigación cualitativa y cuantitativa son conceptualmente reconocidos, sin embargo los métodos de análisis tienden a consumir tiempo y recursos especializados, además de presentar dificultades para obtener una adecuada confiabilidad. Durante años, nuestro equipo de investigación ha probado diversos programas que ayuden a automatizar el proceso, manteniendo la riqueza de la interpretación en su contexto. Recientemente, hemos probado una solución que finalmente está permitiendo automatizar el proceso, reduciendo drásticamente el tiempo requerido y mejorando la confiabilidad.

El análisis de contenido y la investigación émica

Nuestro equipo de investigación ha venido estudiando el comportamiento social en comunidades andinas y amazónicas relativamente aisladas de las influencias occidentales modernas (Yamamoto, 2004b, 2008b; Yamamoto, Feijoo, & Lazarte, 2008). Allí, las teorías psicológicas "universales" descubrían su sesgo occidental. No se podían confirmar sus postulados, no reflejaban estas realidades e incluso, los instrumentos eran simplemente inservibles. Este choque intercultural de teorías nos llevó a optar por una línea de investigación émica. En contraste con la investigación ética, la investigación émica (Jahoda, 1977) consiste en un proceso inductivo que hace uso de métodos cualitativos que permiten el uso de conceptos generados desde los propios datos, evitando así caer en visiones apriorísticas.

Típicamente la recogida de datos cualitativos para la investigación émica se hace mediante entrevistas en profundidad o cuestionarios con preguntas de respuesta abierta. La herramienta utilizada para el procesamiento de esta información es el análisis de contenido. En la literatura, el análisis de contenido más difundido se basa en el conteo de la presencia o ausencia de categorías establecidas y codificadas previamente por el investigador (Bazeley, 2006; Piñuel, 2002; Weber, 1990).

En estudios hechos por nuestro equipo de investigación (Yamamoto, 2004b, 2007; Yamamoto & Feijoo, 2007) hemos optado por el análisis de contenido heurístico, el cual nos permite contar con categorías generadas durante el proceso de categorización. El proceso consiste en reducir las respuestas obtenidas en categorías mínimas que expresen la idea del entrevistado, agrupando las respuestas similares bajo una misma categoría. Esto permite contar con una lista de categorías que representan el total de respuesta de la muestra para determinada variable. Esta lista se convierte en la lista de variables de una base de datos. A través de un proceso de dicotomización, se asigna cero al individuo que no ha mencionado la categoría y uno al que sí. De esta forma, se pueden realizar diversos procedimientos estadísticos que soporten datos dicotómicos como análisis inferenciales no paramétricos y análisis factorial. Más aún, la lista de categorías producto del análisis de contenido puede convertirse en los ítems de una escala piloto que aplicada a una muestra adecuada es sensible a los análisis psicométricos establecidos, con la ventaja del origen émico de sus ítems. Esta integra-

*** Dirección para correspondencia [Correspondence address]:**
Brenda Lía Chávez. Pontificia Universidad Católica del Perú, Av. Universitaria 1801, Lima 32, Peru. E-mail: blchavez@pucp.pe

ción de métodos cualitativos y cuantitativos tiende a brindar índices de confiabilidad y validez muy altos y constituye un paso crítico en la producción de información consistente, precisa y sensible al contexto (Yamamoto, 2008a).

Sin embargo el análisis de contenido también tiene algunas desventajas. En primer lugar, el alto costo en tiempo y personal especializado. La tarea de categorización debe ser resuelta por un investigador con experiencia, criterio y agudeza en la comprensión de las variaciones regionales del lenguaje. En segundo lugar, la dificultad de obtener confiabilidad en los resultados. La creación de categorías y la inclusión de casos en una u otra categoría puede resultar distinta entre un investigador y otro, requiriendo la evaluación y comparación entre los resultados obtenidos por diferentes "codificadores" (Kolbe & Burnett, 1991; Tinsley & Weiss, 2000).

Para resolver estas limitaciones, se han desarrollado métodos automatizados para el análisis de contenido. No obstante, en la revisión de 21 programas para análisis de texto computerizado, Lowe (2003) encuentra que la mayoría cumple funciones básicas y no recomendables para el reemplazo del análisis de contenido manual. Por un lado están los programas más "estadísticos" como *WordStat*, que realizan el conteo de frecuencias basado en diccionarios cuyos términos deben ser introducidos por el mismo investigador. Bazeley (2006) señala que estos programas tienen un alto potencial para la generación de categorías sin sentido. Por otro lado están los programas más conocidos en el análisis cualitativo, como ATLAS.ti o NVivo, que constituyen un soporte para anotaciones o una suerte de fichero virtual, sin mayores posibilidades de categorización automática.

Los avances en el campo de la Lingüística Computacional y específicamente en el Procesamiento del Lenguaje Natural (PLN) brindan una perspectiva diferente. Desde la década de los años noventa, se considera que la categorización mediante programas basados en PLN tiene un gran potencial en el procesamiento de grandes cantidades de texto (Jacobs, 1992; Sable, McKeown, & Church, 2002). Estos programas cuentan con algoritmos que analizan la estructura y significado del lenguaje, es decir, no se limitan a hacer un conteo de términos sino que "entienden" el significado de las palabras en su contexto y facilitan la identificación de relaciones entre términos, pudiendo así analizar las ambigüedades inherentes a la comunicación verbal. Sin embargo, diferentes problemas dificultaban su puesta en práctica, especialmente en lenguaje español dada la limitación de módulos y diccionarios en este lenguaje.

Luego de diferentes pruebas, el programa de lingüística computacional SPSS *Text Analytics for Surveys* es el primero que nos ha brindado resultados satisfactorios en el análisis de contenido automatizado. Adicionalmente, es una solución de bajo costo. Sin embargo, se requiere contrastar sus resultados en relación con los métodos manuales establecidos. Por tanto el objetivo del presente estudio es comparar los resultados de un análisis de contenido hecho manualmente por analistas especializados con uno asistido por el programa mencionado.

Método

Participantes

La muestra está conformada por 27 sujetos, pobladores de la ciudad de Iquitos, en la Amazonía peruana. Fueron, 17 hombres y 10 mujeres, con edades entre los 24 y 68 años (M = 39.8, DT = 10.36). Todos los participantes contaban con educación superior universitaria o técnica.

Instrumento

Se aplicó la *Entrevista de Componentes Émicos del Bienestar* (ECB: Yamamoto, 2004). Este es un protocolo de entrevista estructurado de respuesta abierta que indaga los componentes del bienestar: metas y necesidades, recursos, valores, momentos más felices y momentos más infelices. En el presente estudio se analiza únicamente el módulo de metas y necesidades.

Técnica Analítica

Se aplicó la técnica de análisis de contenido heurístico. Las respuestas fueron sintetizadas en categorías mínimas que expresaran la idea central de lo respondido por cada sujeto. Las respuestas semejantes de distintos sujetos fueron agrupadas bajo una misma categoría. Al finalizar la categorización se calculó la frecuencia mediante el conteo del número de sujetos que mencionaron las ideas pertenecientes a cada una de las categorías formadas. Se utilizó una hoja en el programa Excel para organizar la información. Se estimó el tiempo requerido para todo el proceso.

Paralelamente se hizo un análisis de contenido utilizando el programa SPSS *Text Analytics for Surveys* 4.0. Se realizó la digitalización del texto de cada cuestionario. Se importaron estos datos al programa, el cual utiliza una combinación de técnicas lingüísticas y estadísticas automáticas para la extracción de términos y generación de categorías. El programa también realizó el conteo de frecuencias, con lo cual se obtuvo una lista de categorías para ser comparadas con el análisis de contenido manual. Se estimó el tiempo requerido para todo el proceso.

Finalmente se realizó la comparación de ambas técnicas. Los resultados fueron confrontados señalando las categorías compartidas y las particulares a cada técnica.

Resultados

El análisis de contenido manual (ACM) demandó aproximadamente 13 horas de trabajo divididas en tres días, mientras que el análisis con el programa de lingüística computacional (PLC) se realizó en tres horas y un solo día. Se obtuvieron 45 categorías con el ACM y 47 con el PLC, de las cuales 27 comparten nombres similares. La Figura 1 presenta las categorías compartidas con sus respectivos nombres y frecuen-

cias. Las categorías compartidas están conformadas por 14 categorías iguales y 13 con variaciones de frecuencia. Once de estas últimas se diferencian en un solo caso y dos se diferencian en dos casos.

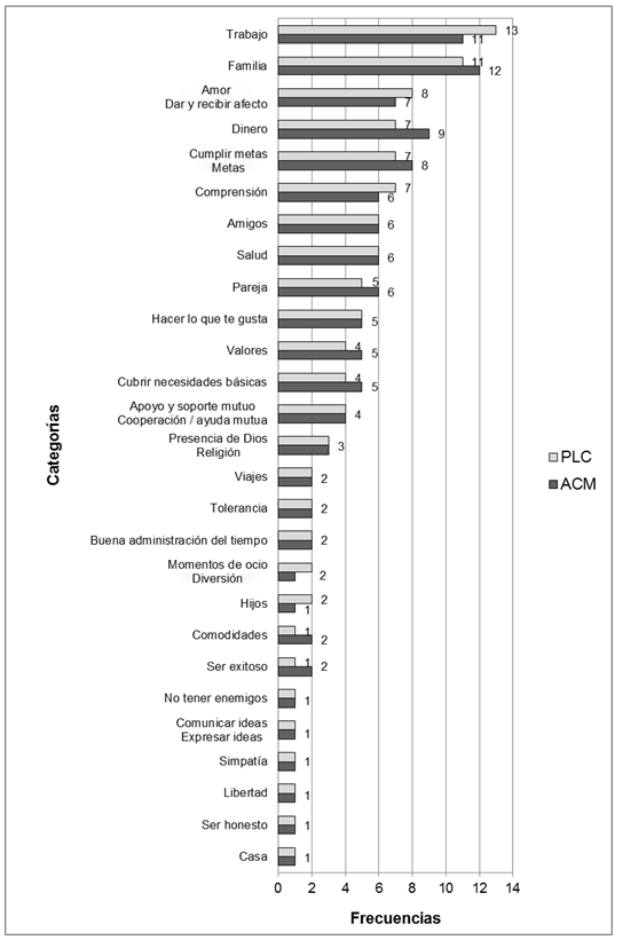

Figura 1. Categorías compartidas resultado del análisis de contenido manual (ACM) y el análisis de contenido del programa de lingüística computacional (PLC).

Las razones de las diferencias entre las categorías observadas son tres. En primer lugar, el ACM es más preciso al reconocer el significado de las palabras, ya que el analista maneja un nivel mayor de identificación de posibles sinónimos entre categorías. Un ejemplo de ello es que en el ACM se reconoce "labor" como un sinónimo de "trabajo" y en el PLC se categorizan de forma separada. Sin embargo se le puede "enseñar" al PLC que se deben considerar esas palabras como sinónimos añadiendo términos a la biblioteca de recursos. En segundo lugar, el PLC delimita las categorías de forma más concreta y ACM de forma más abstracta. Por ejemplo, el ACM reconoce el término "unión" como parte de la categoría "cooperación/ayuda mutua" mientras que el PLC lo deja fuera de ella. Mediante la revisión de las fórmulas de categorización generadas automáticamente por el PLC se puede agregar mayor abstracción cuando sea necesario. En tercer lugar, gracias a la rutina de lectura automática y una sección dedicada a las "extracciones no usadas", el PLC presenta términos que en algunos casos fueron pasados por alto en el ACM.

Las categorías particulares formadas sólo por el ACM o el PLC pueden observarse en la Figura 2. Se encuentra que en ambos métodos se reconoce el mismo contenido pero las categorías se organizan de forma distinta, un ejemplo se presenta en la Tabla 1. Finalmente, el PLC generó 14 categorías de frecuencia uno, mientras que el ACM generó sólo cuatro categorías con frecuencia uno. El PLC genera más categorías de baja frecuencia debido a que ubica los términos no utilizados en categorías unitarias. El ACM tiende a una mayor abstracción en la conformación de categorías de baja frecuencia, generando más categorías de frecuencia dos o tres. Nuevamente, el nivel de abstracción pertinente puede afinarse en la configuración de análisis con el PLC.

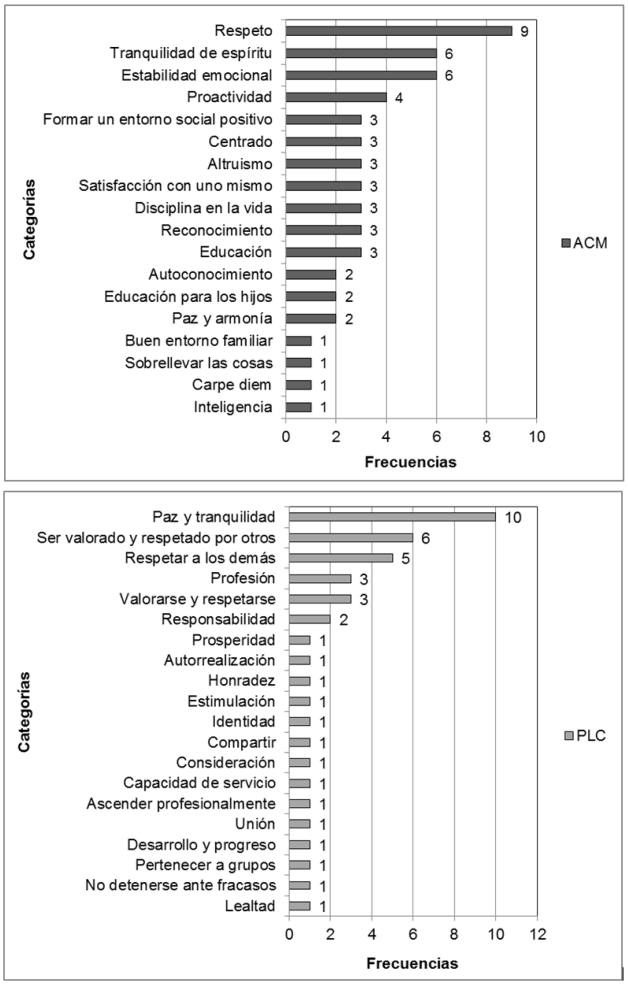

Figura 2. Categorías particulares resultado del ACM o el PLC.

Tabla 1. Ejemplo de organización del contenido en categorías según ACM y PLC. La categoría Respeto del ACM se divide en dos categorías con el PLC

ACM		**PLC**	
Categoría "Respeto"	Respetando el derecho de los demás Reconocer los derechos de los demás Respetar a los demás Respeto mutuo entre las personas Actuando conscientemente con nuestro entorno (con el prójimo), para tener la conciencia limpia Ser respetado Ser visto con respeto Ser respetado en la sociedad Consideración	**Categoría "Ser valorado y respetado por otros"**	Respetando el derecho de los demás Reconocer los derechos de los demás Respetar a los demás Respeto mutuo entre las personas Actuando conscientemente con nuestro entorno (con el prójimo), para tener la conciencia limpia
		Categoría "Respetar a los demás"	Ser respetado Ser visto con respeto Ser respetado en la sociedad Sentirse apreciado Ser reconocido Que la gente valoren tu trabajo

Discusión y conclusiones

En líneas generales, revisando las extracciones y generando recursos afinados en el PLC se pueden obtener resultados similares al ACM. Sin embargo, el uso del PLC tiene las ventajas del menor tiempo y recursos consumidos. Adicionalmente, la confiabilidad aumenta al poder utilizarse una misma plantilla de generación de categorías para todo el análisis, en vez de utilizar varios investigadores o asistentes que realizan el análisis de contenido en paralelo, requiriendo una revisión cruzada.

Otra ventaja del uso del PLC son las formas de visualización. En una ventana del programa, la respuesta original se muestra completa y únicamente se subraya la extracción (frase o término que luego será incluido en la categoría). En el ACM se suele leer la respuesta y reducirla a frases más cortas para insertarla en el archivo de trabajo, lo cual puede llevar a la pérdida de información o a caer en errores de lectura. Asimismo, el programa facilita la visualización de la categorización en el marco completo de la respuesta, lo cual permite revisar la pertinencia de la inclusión de términos en cada categoría y favorece la comprensión de la información en su contexto de respuesta.

El proceso repetido de extraer, revisar, refinar y volver a extraer potencializa la categorización con el PLC. A mayor uso del programa se van perfeccionando los recursos, mediante la adición de términos a la biblioteca y la creación de plantillas con fórmulas de categorización refinadas. Resulta una ventaja el hecho de que todos estos recursos puedan ser luego reutilizados y compartidos entre investigadores.

Finalmente, el PLC permite diseñar investigaciones que eran poco factibles en el pasado. Los estudios de base son fundamentales para el desarrollo de instrumentos psicométricos sólidos cuyos ítems representen con precisión la variedad y amplitud de las variaciones en una población. Esto ha estado limitado por la complejidad y costos del análisis de contenido de muestras grandes. Con los PLC, resulta factible y económico realizar estudios de base con 500, 1500 o más participantes, abriendo el camino para estudios basados en entrevistas con alternativas de respuesta abierta representativos a nivel país. Estimamos que un análisis de contenido para 500 participantes utilizando los PLC debe tomar unas nueve horas. Una vez afinadas las bibliotecas y las plantillas, la diferencia para analizar 1500 participantes sería marginal. Este tipo de estudios permitiría una nueva generación de instrumentos psicométricos émicos, con ítems realmente representativos de la complejidad de una población.

Existe un uso amplio de estudios basados en grupos focales o *focus groups* en la fase cualitativa, seguidos de una fase cuantitativa orientada en los resultados de los grupos focales. Sin embargo, es conocido que en los grupos focales prevalece el consenso grupal en contra del registro adecuado de la varianza individual, teniendo como ventaja un bajo costo. En este nuevo escenario, los costos de una serie de grupos focales seguidos de una fase cuantitativa serían menores a los estudios de entrevistas abiertas con muestras representativas. Adicionalmente, estos últimos tendrían la ventaja de que no se haría una inferencia desde los resultados cualitativos hacia la fase cualitativa, sino que se analizaría directamente la fase cualitativa aplicada a una muestra representativa. Esto tendría enormes implicaciones en los estudios aplicados en el ámbito de la investigación de mercados, los estudios organizacionales de base así como en los diagnósticos para programas de desarrollo social.

En conclusión, los PLC llegaron a su momento de madurez, existiendo programas como el *Text Analytics for Surveys* que ofrecen a un bajo costo, soluciones amigables y consistentes en una amplia variedad de lenguajes. Ofrece, con un debido manejo, un ahorro sustancial de tiempo, recursos así como un incremento importante en la confiabilidad y de la validez. Abre puertas a una nueva generación de estudios que no estén restringidos por los costos y confiablidad de análisis de contenido de respuestas abiertas en muestras grandes. Ofrece un promisorio escenario para la investigación pura y aplicada. Las ciencias sociales se han sustentado

con una sólida estructura cuantitativa que reposa en una deleznable estructura de estudios de base, sustentados en la cáscara de huevo de los grupos focales y los cimientos de entrevistas abiertas tan profundos como su tamaño muestral. En un futuro cercano, quizá seamos testigos de una importante reconstrucción de teorías y diagnósticos a través de estudios de base de una nueva generación.

Agradecimientos.- Los autores agradecen a Ángela Ríos Cardoso por su apoyo en la recolección de los datos en la Amazonía. A Roger Molina de SPSS Andino por las facilidades en la información y licencias para las pruebas con el programa *Text Analytics for Surveys*. A Cinthya Díaz por el apoyo y revisión en el desarrollo del presente estudio.

Referencias

Bazeley, P. (2006). The contribution of computer software to integrating qualitative and quantitative data and analyses. *Research in the Schools, 13*(1), 64-74. Doi: citeulike-article-id:4171634

Jacobs, P. S. (1992). *Joining statistics with NLP for text categorization*. Paper presented at the Proceedings of the third conference on Applied natural language processing, Trento, Italy.

Jahoda, G. (1977). In pursuit of the emic-etic distinction: Can we ever capture it? In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 55-63). Lisse: Swets and Zeitlinger.

Kolbe, R. H., & Burnett, M. S. (1991). Content-analysis research: An examination of applications with directives for improving research reliability and objectivity. *Journal of Consumer Research, 18*(2), 243-250. Doi: 10.1086/209256

Lowe, W. (2003). Software for Content Analysis – A Review. Retrieved from http://kb.ucla.edu/system/datas/5/original/content_analysis.pdf

Piñuel, J. L. (2002). Epistemología, metodología y técnicas del análisis de contenido. *Estudios de Sociolingüística, 3*(1), 1-42.

Sable, C., McKeown, K., & Church, K. W. (2002). *NLP found helpful (at least for one text categorization task)*. Paper presented at the Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10.

Tinsley, H. E. A., & Weiss, D. J. (2000). Interrater reliability and agreement. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling.* (pp. 95-124): San Diego, CA, US: Academic Press.

Weber, R. (1990). *Basic Content Analysis*. Londres: Sage Publications.

Yamamoto, J. (2004a). *El protocolo de entrevista a profundidad de componentes de bienestar*. Pontificia universidad católica del Perú. Documento inédito. Lima.

Yamamoto, J. (2004b). *Hacia una metodología de intervención con criterio intercultural*. SNV amazonía. Agencia holandesa de cooperación internacional.

Yamamoto, J. (2007). *Calidad de vida en comunidades rurales, peri urbanas y urbano marginales en un corredor andino. Hacia un modelo multinivel de bienestar y desarrollo*. Paper presented at the XIII congreso nacional y III congreso internacional de psicología, Cusco

Yamamoto, J. (2008a). Implications for Wellbeing Research and Theory. In J. Copestake (Ed.), *Wellbeing and Development in Peru. Local and Universal Views Confronted* (pp. 231-242). New York: Palgrave McMillan.

Yamamoto, J. (2008b). Un regard alternatif sur la subjectivité : le bien être des communautés andines. *Connexions, 89*, 147-170.

Yamamoto, J., & Feijoo, A. R. (2007). Componentes émicos del bienestar: Hacia un modelo alternativo de desarrollo. *Revista de Psicología (Lima), 25*, 197-231.

Yamamoto, J., Feijoo, A. R., & Lazarte, A. (2008). Subjective Wellbeing: An Alternative Approach. In J. Copestake (Ed.), *Wellbeing and Development in Peru. Local and Universal Views Confronted* (pp. 61-101). New York: Palgrave McMillian.

(Artículo recibido: 26-6-2012; revisión recibida: 20-3-2013; aceptado: 4-5-2013)