RECONOCIMIENTO AUTOMÁTICO DE RIMAS PARA EL *CANCIONERO FOLKLÓRICO DE MÉXICO*

Entre los trabajos de computación preparados para obtener los índices del *Cancionero folklórico de México* [1] hemos elaborado un programa rastreador para el reconocimiento automático de rimas, cuyo resultado será el índice de las mismas. Su importancia se halla en los interesantes problemas que plantea la construcción de un programa no solamente de ordenación, sino también de análisis. Queremos presentar aquí la forma en que lo realizamos y los resultados que hemos obtenido.

La finalidad de nuestro índice de rimas es la de reconocer tanto las rimas iguales que aparecen en distintas coplas como el juego de rimas que se da en el interior de una misma copla. El índice de rimas consta de todas las palabras rimantes, presentadas en todos los órdenes posibles (hay versiones que invierten los versos) de cada una de las coplas; por ejemplo, de la copla núm. 847:

> Mi corazón, pobrecito,
> preso en tus brazos está:
> castígale su delito
> y dale la libertad,
> que él te pagará solito
> y en buena conformidad.

necesitamos obtener las siguientes listas: pobrecito, delito, solito / delito, pobrecito, solito / solito, pobrecito, delito / está, libertad, conformidad / libertad, está, conformidad / conformidad, está, libertad.

La utilidad inmediata de este juego de combinaciones es la de

[1] El *Cancionero folklórico de México* es el resultado de una investigación colectiva realizada bajo la dirección de Margit Frenk Alatorre en el Centro de Estudios Lingüísticos y Literarios de El Colegio de México. A la fecha se han publicado dos tomos: *Coplas del amor feliz*, 1975 y *Coplas del amor desdichado*, 1977.

localizar coplas similares y relacionarlas dentro del *Cancionero,* pues debido al volumen del material con que se trabaja (12 000 coplas, muchas de ellas con más de una versión), es muy difícil poder encontrarlas únicamente con la ayuda de la memoria; también será un auxiliar muy útil para los lectores interesados en localizar una copla determinada de la obra, cuyos índices, incluyendo el de rimas, se publicarán en el quinto tomo.

Llamamos "verso" a cada una de las líneas que componen una estrofa o copla. Para la definición de "rima" nos hemos basado en un trabajo de A. M. Cirese [2] quien la define (siguiendo la tradición) como la identidad o similitud entre las terminaciones de la última palabra de los versos que componen una estrofa o copla. La "terminación" abarca desde la vocal tónica hasta el final de la palabra. Existe identidad cuando tanto las vocales como las consonantes de las terminaciones son iguales (rima consonante); existe similitud cuando únicamente las vocales son iguales (rima asonante). En esta etapa del programa no hemos tenido en cuenta otros casos de rima asonante como cuando hay dos vocales en la terminación y la vocal tónica es igual, pero la átona puede ser *e* o *i* u *o* o *u*. Tampoco hemos considerado casos de aliteración. Si quisiéramos comprobarlos en el *Cancionero* bastaría ampliar el programa.

El factor decisivo en la rima es la vocal tónica; es el punto a partir del cual medimos la terminación y es el único elemento necesario: puede o no haber otra vocal o consonante, pero la vocal tónica siempre está presente.

El material del *Cancionero* ha sido codificado y perforado en tarjetas. Las convenciones para la transcripción han sido las siguientes: cada tarjeta debe contener un verso de una copla copiado textualmente, acompañado por una clave que nos permite su localización dentro del corpus del *Cancionero;* esta clave consta de diez caracteres en total. El primero indica el número de tomo en que se encuentra la copla y es un dígito con valor 1, 2, 3 o 4. A continuación tenemos cinco caracteres que forman una cifra de cinco dígitos con valores posibles del 00001 al 99999 donde se marca el número de la copla: el séptimo carácter se utiliza para los casos de intercalación de coplas en la numeración original del *Cancionero* [3] y puede ser un dígito del 0 al 9, donde se prevé un caso extremo de nueve coplas intercaladas; el octavo carácter es alfabético y marca los casos

[2] "Inventaires et répertoires lexicaux, formulaires et métriques des chants populaires italiens" en *Linguistica matematica e calcolatori,* ed. A. Zampolli, Firenze, 1973, p. 227.

[3] Para evitar correr toda la numeración se les dio el número de la que las precede pero con las marcas *bis* o *ter* para diferenciarlas. Cf. *Cancionero,* t. 1, p. xxxii, § 4.

de familias de coplas [4]; los dos últimos caracteres forman una cifra de dos dígitos del 00 al 99 que nos indica el lugar que ocupa el verso dentro de la copla. Así la clave 1004920f01 se lee de la siguiente forma: es un verso que se encuentra en el tomo 1, que pertenece a la copla 00492, la cual está intercalada y pertenece a una familia de coplas ocupando el lugar *f* y es el primer verso de la copla.

Los pasos que debe seguir el programa rastreador de rimas son los siguientes:

1) Tomar como unidad de análisis la copla;

2) Aislar la última palabra de cada verso, pero con las siguientes restricciones:

a) No tomar en cuenta los elementos añadidos al final de la línea cuando van entre paréntesis. En esta situación podemos encontrar la palabra *sic*, un signo ?, o bien una palabra o grupo de palabras que llamamos apéndices y que quedan fuera de la configuración métrica del verso, por ejemplo en la copla núm. 351:

> Si porque te quiero, quieres (Llorona)
> quieres que te quiera más,
> te quiero más que a mi vida (Llorona):
> ¿qué más quieres?, ¿quieres más?

donde las últimas palabras de los versos 1 y 3 son *quieres, vida* y no *Llorona.*

b) Si el apéndice no va entre paréntesis, sino separado del cuerpo principal del verso con un guión largo [5], se toma la última palabra antes del guión y la última del apéndice; por ejemplo en la copla núm. 630:

> Dí si me quieres, Lola
> para comprarte
> una barca con remos — y con sus velas
> para pasearte.

tomaríamos las palabras *Lola, comprarte, remos, velas* y *pasearte.* Son muy pocos estos casos en el cuerpo general del *Cancionero,* pero es necesario tomarlos en cuenta.

[4] "En cuanto a las coplas que tienen versiones de distinto número de versos hay motivos para verlas como 'una sola'... Por razones de claridad, sin embargo, convenía imprimirlas por separado; lo que se ha hecho es ponerles un mismo número seguido de las letras a, b, c, etc." *(ibid.,* § 2).

[5] "En algunas seguidillas mexicanas el apéndice aparece sustituido por palabras que son indispensables para la comprensión del texto. Las hemos puesto a continuación del verso y separándolas de él por medio de un guión", *ibid.,* p. xxxvi.

c) En algunas ocasiones se utilizan corchetes cuando una fracción del verso (que se ha tomado de una grabación magnetofónica) no se escuchó bien, pero por algún medio se pudo completar; si esto sucede en la última palabra del verso, ésta debe leerse, para nuestros fines, como si los corchetes no estuvieran allí; por ejemplo en la copla núm. 1386*b*:

> Ariles y más ariles
> que ariles del que decía:
> "De noche te vengo a ver,
> porque no puedo de día,
> y si pudiera [viniera]
> a todas horas del [día]".

tanto *viniera* como *día* se consideran palabras finales.

3) Localizar la vocal tónica en cada última palabra. Como ya dijimos, la vocal tónica se identifica por medio del acento y por lo tanto para localizarla es necesario, en primer lugar, un algoritmo de reconocimiento de vocales (formado con un inventario de vocales y un cuadro de reconocimiento y reduccción de diptongos; cf. *infra*, p. 503) ; y, en segundo lugar, un proceso de pre-edición en donde se marquen los acentos prosódicos; o bien es necesario un algoritmo de identificación, que contenga las reglas de acentuación del español, por medio del cual se localice y acentúe la vocal tónica cuando ésta no lleva acento ortográfico. Puesto que las reglas de acentuación del español son bastante sencillas [6] hemos preferido la segunda posibilidad, que además elimina el trabajo de pre-edición. Las reglas para el reconocimiento de la vocal tónica (VT) son:

i) Si hay acento ortográfico, la VT es la que precede al acento: V/[7].
ii) Si no hay acento ortográfico:
  a) y la palabra sólo tiene una vocal, se considera que ésa es la VT;
  b) y la palabra tiene sólo dos vocales y éstas forman diptongo, se aplican las reglas de reducción de diptongo (cf. *infra*, p. 503) ; la que resulta será la VT;
  c) y la palabra termina en *n*, *s* o vocal, la VT es la penúltima, siempre y cuando ésta no forme diptongo con la

[6] Nos basamos en las reglas presentadas en un estudio sobre ortografía y pronunciación elaborado para el *Diccionario del español de México* por Lourdes Gavaldón, Lourdes Ros y Manuel Fernández.

[7] En la codificación que utilizamos, el acento ortográfico se marca "/" a continuación de la vocal acentuada, por ejemplo: CANCIO/N.

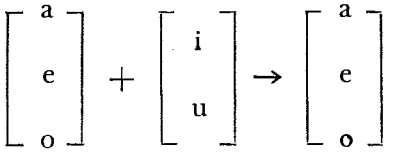
última, en cuyo caso se tomará la anterior; por ejemplo, en *gloria,* la VT no es la *i* sino la *o;*

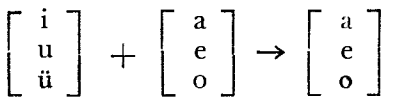
d) y la palabra no termina en *n, s,* o vocal, la VT es la última.

Las reglas para el reconocimiento y reducción de los diptongos son:

i) Si hay dos vocales unidas, pero una de ellas lleva acento ortográfico, se cuentan como dos vocales: V/V o VV/[8]. Igual tratamiento reciben si no llevan acento pero no son combinaciones con *i* o *u,* por ejemplo: *ea* dos vocales.

ii) Si hay dos vocales unidas, ninguna de las dos lleva acento ortográfico, y una de ellas es *i, u* o una combinación de ambas *iu, ui,* se cuentan como una sola vocal, de acuerdo con el siguiente cuadro:





iu	→ u
ui	→ i
üi	→ i

[9]

4) Separar la terminación a partir de la VT incluyendo a ésta.

5) Comparar las terminaciones entre sí y reunir lo que muestra una relación de identidad (rima consonante) o de similitud (rima asonante). También existe una rima "intermedia" en la que las vocales son idénticas y las consonantes muy similares, pues corresponden a los casos en que puede darse neutralización entre ellas porque casi todos sus rasgos distintivos son iguales; por ejemplo,

[8] No siguen este patrón las formas de la segunda persona plural terminadas en -áis, -éis; tampoco las palabras dieciséis, veintiséis, agnusdéi, aindamáis, buéis, marramáu (cf. G. SCAVNICKY and A. STAHL, *A reverse dictionary of the Spanish language,* Urbana, 1973), ni de los apellidos Monsiváis, Beristáin, Araquistáin. En todas ellas se mantiene el diptongo a pesar del acento. No hemos considerado estas excepciones porque es muy poco probable que aparezcan en un cancionero mexicano.

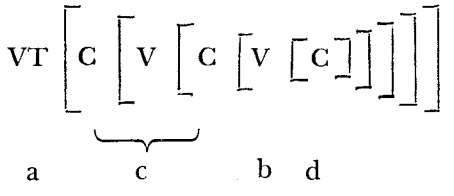
[9] La *ü* no es una vocal especial, sino una convención ortográfica para indicar que la *u* suena en las combinaciones *güi, güe.*

m y *n* son nasales y sonoras y sólo se oponen por ser una bilabial y la otra alveolar. Para las comparaciones se necesita:

a) un inventario de vocales; lo que no aparezca en éste se considerará consonante;
b) un inventario de las diferentes grafías para un mismo fonema (en el sistema fonológico del español mexicano): *b* = *v; g* = *j; z* = *s; c* = *s; y* = *ll;*
c) un inventario de las consonantes con las que se puede dar la rima intermedia: *m-n; l-r; r-rr* [10].

Como hemos encontrado en el *Cancionero* diferentes patrones de rima *(abcb, abab, abba, ababab,* etc.), es necesario comparar entre sí todas las palabras rimantes de una copla.

El patrón general de la terminación sería el que viene a continuación, donde todo lo que va entre corchetes es opcional:



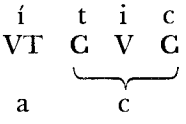
donde VT = vocal tónica, C = consonante o grupo de consonantes, V = vocal o diptongo. Por ejemplo, en las palabras *nombró, crió, fe* la terminación está constituida únicamente por la VT — *ó, e;* en *corazón* tenemos la terminación *-ón* con los elementos marcados con a y c en el esquema, o sea VT-*ó* y *C-n;* en la terminación de *flores* aparecen los cuatro elementos a, c, b y d: VT-*o,* C-*r,* V-*e,* y C-*s.*

En el esquema se incluye la posibilidad de analizar palabras esdrújulas rimantes, como en la copla núm. 377*a* (caso rarísimo en el corpus del *Cancionero*):

Por tu boquítica,
por tus piecíticos,
por tus ojíticos,
me muero yótico.

donde las cuatro palabras finales de verso son esdrújulas; así, la terminación *íticos* (de los versos 2 y 3) tiene el esquema

[10] Sólo hemos considerado estos tres casos pues parecen ser los más claros. Esto no elimina la posibilidad de que se den otros; el problema sería marcarnos un límite y explicarlos.



Los elementos que se comparan en las terminaciones son uno como mínimo (a) y cuatro como máximo: *a)* la VT; *b)* la última V de la terminación, si no es ésta la VT; *c)* la letra o grupo de letras que van después de la VT o entre la VT y la última V; *d)* la letra o letras que van después de la última V si no es ésta la VT.

Todos los requisitos que hemos enumerado aquí se han traducido a un programa en lenguaje Fortran con un arreglo de tabla, que consta de siete subrutinas. La primera se denomina TCHARD y es la que crea la tabla y las subtablas que la componen. En las subtablas encontramos traducidos a caracteres numéricos todos los elementos importantes para la determinación de la rima: acento, vocales, corchetes, paréntesis, blancos, etc. Las subrutinas restantes (llamadas ACENTO, PAREN, CORCHE, ERROR, DETACE y DETRIM) son de dos tipos diferentes; las cinco primeras se encargan de localizar las terminaciones y guardar todos sus componentes analizados; la última, con base en las anteriores, compara los componentes de las terminaciones de una copla y establece la existencia y tipo de rima que se da entre los versos.

El programa tiene como unidad de trabajo la copla, y funciona de la siguiente manera:

Una vez creada la tabla, se aplica a la última palabra de cada verso un algoritmo de lectura que funciona de derecha a izquierda; este algoritmo analiza carácter por carácter y cuando aparece un carácter significativo (acento, paréntesis, corchete o blanco) entra a una de las subrutinas siguientes:

1) Subrutina ACENTO. Entra en funcionamiento cuando aparece en la lectura un acento ortográfico (/); reconoce la vocal tónica y guarda la terminación completa, las vocales de que consta y su número para la comparación posterior.

2) Subrutina PAREN. Se entra a ella cuando aparece un paréntesis derecho [)]; analiza cada carácter hasta encontrarse el paréntesis izquierdo [(]; si lo encuentra, inicia el análisis de la palabra que lo precede, pues, como ya dijimos, las palabras entre paréntesis no se toman en cuenta; si no encuentra el paréntesis izquierdo, se traslada a la subrutina ERROR.

3) Subrutina ERROR. En ella se guardan los casos en que se han encontrado errores de codificación para marcarse en la impresión final y corregirse a mano.

4) Subrutina CORCHE. Se utiliza cuando aparece un corchete derecho (]); sigue el análisis ya sea para encontrar el acento ortográfico o para determinar el prosódico.

5) Subrutina DETACE. Entra en funcionamiento al reconocer el primer blanco, determinándose así el límite de la palabra. Al no haber encontrado el acento ortográfico, aplica las reglas de acentuación (cf. *supra,* pp. 502 *s.*), que en ellas están contenidas para localizar la vocal tónica. Una vez determinada, sigue los mismos pasos que la subrutina ACENTO.

6) Subrutina DETRIM. Una vez establecidas las terminaciones de una copla, en base a las cinco subrutinas precedentes, se entra a ésta para realizar sus comparaciones en todas las combinaciones posibles: terminación del primer verso con el segundo, tercero, etc., la del segundo con el tercero, cuarto, etc. Puede asignar en cada una de las comparaciones los siguientes valores:

0 no hay rima
1 rima asonante
2 rima intermedia
3 rima consonante

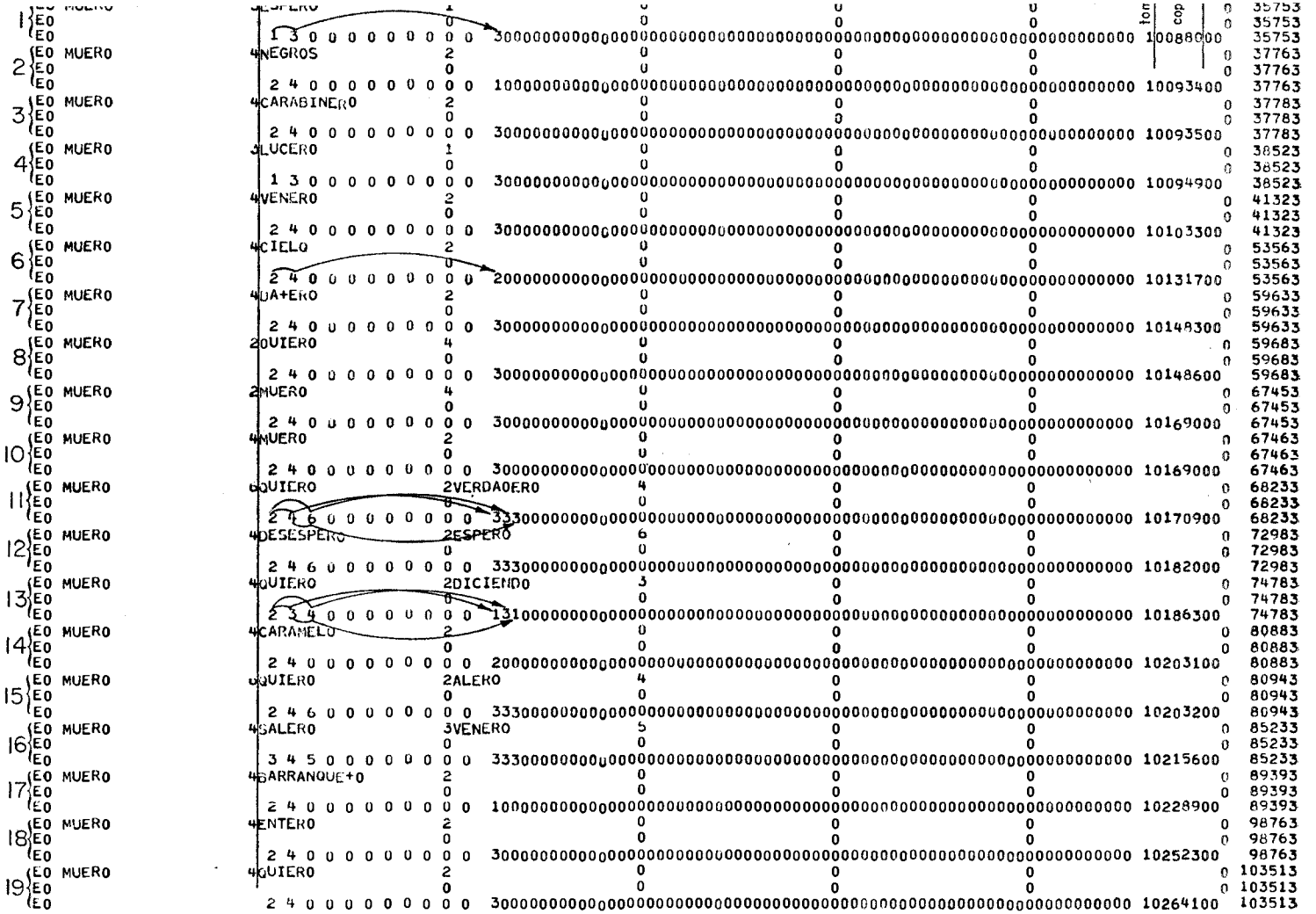
En el valor 3 se incluyen los casos de diferentes grafías para un mismo fonema.

Esta subrutina analiza las vocales de las terminaciones que compara; si no son iguales asigna el valor 0; si son iguales compara las consonantes (cuando las haya) y, si son iguales, asigna un valor 3. En el caso de que las consonantes no sean iguales las analiza para ver si se trata de un caso de diferente grafía o de rima intermedia; cuando sucede lo primero asigna un valor 3, cuando lo segundo, un valor 2, y si no se da ninguno de estos casos, asigna el valor 1.

Los datos producidos por el programa rastreador se ordenan, por medio de programas muy sencillos, de acuerdo con las terminaciones rimantes atendiendo principalmente a las vocales de la terminación y se obtienen grupos del tipo *á, á-a, á-e,...é, é-a, é-e,...í,* etc. Dentro de cada uno de los grupos se presentan las palabras rimantes de cada copla en orden alfabético y en todas las combinaciones posibles; por ejemplo, para una copla con rima *é,* cuyas palabras rimantes son *hallé* y *corté,* encontramos en la agrupación *é,* en la H: *hallé* y *corté* y en la C: *corté, hallé.*

Incluimos aquí una hoja de los listados, que se debe interpretar de la siguiente forma: en el extremo izquierdo aparece la asignación del grupo al que corresponden todos los datos (en este caso EO = é-o). Los datos de cada copla ocupan tres renglones consecutivos; en el primero se leen las palabras rimantes con un número a la derecha que indica el verso a que corresponde cada una [11]; en el segundo renglón vuelve a aparecer la palabra bajo la que se está alfa-

[11] El espacio entre la palabra y su número se debe a necesidades prácticas, pues puede haber palabras muy largas que lo ocupen casi en su totalidad.



betizando y espacios para el caso en que la copla tenga más de seis palabras rimantes, que son las que caben en el primer renglón; en el tercero encontramos varias agrupaciones de números: en primer lugar están los números de los versos entre los que se da esa rima (los ceros son espacios para los casos de coplas largas); la siguiente serie indica los valores que se obtienen como resultado de la comparación entre las terminaciones rimantes, o sea, el tipo de rima que existe entre cada combinación (se lee como está marcado con las flechas en algunos de los grupos); a continuación, vienen los datos de la copla (número de tomo y número de copla; los dos ceros finales son espacios para número de verso, que en este caso no es pertinente). Los números de la extrema derecha marcan el orden en que se trabajó la copla, antes de ordenarse como aparece en la hoja.

Así, por ejemplo, en el cuadro hay 19 coplas analizadas. La explicación que sigue se refiere a las coplas 1, 6, 13:

En esta esquina te espero;
1 dime que sí, chaparrita,
porque de pena me muero.

¡Ay de mí!, Llorona,
Llorona, llévame al cielo
6 a ver a las rezadoras (Llorona),
que digan cuándo me muero.

Preso estoy, vida mía,
porque te quiero,
13 y yo sigo diciendo
que por ti muero.

En este cuadro aparece sólo la rima en e-o; otras rimas que pueda haber dentro de las coplas aparecen en el grupo de la terminación que les corresponde. En el primer grupo de tres renglones tenemos las palabras rimantes *muero* en el verso 3 y *espero* en el 1; por lo tanto en el tercer renglón leemos 1 y 3, que nos indican los números de verso entre los que se da una rima con valor 3 (consonante), y al final de ese mismo renglón vemos que los datos pertenecen a la copla 880 del t. 1. En el grupo 13 leemos *muero* del verso 4, *quiero* del 2 y *diciendo* del 3; en el tercer renglón encontramos los números 2, 3 y 4 que indican los números de versos rimantes; entre el verso 2 y el 3 tenemos rima 1 (asonante), entre el 2 y el 4 rima 3 (consonante) y entre el 3 y el 4 rima 1 (asonante); los datos pertenecen a la copla 1863 del t. 1. También encontramos un caso de rima intermedia, en el grupo 6.

El programa rastreador de rimas, planeado y aplicado en la forma que aquí lo presentamos, funciona en un 100% de los casos y

no requiere trabajo de post-edición para utilizar los datos que proporciona; al presentarnos el juego completo de rimas de todo nuestro corpus se abre un campo bastante amplio para investigaciones posteriores sobre este tema. En base a esos datos se pueden obtener frecuencias de los tipos de rimas, obtener y analizar los diferentes tipos de estrofa que existen en relación a la rima y conocer, apoyándonos en su frecuencia, cuáles de estos tipos son los más característicos y cuáles los más raros. Si se combina este programa con el analizador gramatical del *DEM* [12], se podrían obtener y analizar también los tipos y frecuencia de las palabras rimantes, su número de sílabas, la categoría gramatical a la que pertenecen (sustantivos, verbos, pronombres, etc.), los sufijos y terminaciones verbales empleados para crear rimas, etc. También se podría pensar en una investigación en la que se compararan los datos resultantes con aquéllos obtenidos en corpus similares de otros países hispánicos.

MARÍA ÁNGELES SOLER DE LA CUEVA
SILVIA PONCE DE LEÓN

El Colegio de México.

[12] Para una explicación de lo que es y cómo funciona el analizador gramatical del *Diccionario del español de México*, véase L. F. LARA, "Méthode en lexicographie: valeur et modalité du dictionnaire de machine", *CLex*, 1976, núm. 2, 103-128.