Automatización del estudio de características lingüísticas del español escrito

Idalme López Oliva y Eduardo Aubert Vázquez.

Centro de Neurociencias de Cuba, Centro Nacional de Investigaciones Científicas, Avenida 25 y 158, Cubanacán, Playa, Ciudad de La Habana, Cuba.

Recibido: 25 de marzo de 1998. Aceptado: 15 de septiembre de 1998.

Palabras clave: psicolingüística, análisis de textos computadorizados, base de datos léxica, frecuencia de uso de las palabras, diccionarios de frecuencia.
Key words: psycholinguistic, computerized text analysis, lexical database, word frequency, frequency dictionaries.

RESUMEN. El uso de palabras como estímulos es muy frecuente en las ciencias de la conducta, por tanto, resulta de gran interés caracterizar las variables lingüísticas de los vocablos en el idioma que se trata. En el español existen algunos estudios de esta índole, pero presentan problemas que los limitan en cuanto a su representatividad, ya sea porque la cantidad de información que se analiza es muy pequeña, la procedencia geográfica del estudio difiere del lugar donde se van a utilizar sus resultados o porque han perdido vigencia a causa del envejecimiento del material usado para su confección. El empleo de los medios de computación permite analizar un gran volumen de información en tiempo breve, confiere mayor confiabilidad al estudio al disminuir los errores humanos y facilita la actualización constante de la información. En el presente trabajo, se expone un sistema para extraer automáticamente, a partir de un conjunto de textos, variables lingüísticas pertinentes a palabras y sílabas, tales como frecuencia de aparición, longitud, irregularidad grafema-fonema, cantidad de sílabas, acentuación ortográfica, patrón de división silábica y sucesión de consonantes y vocales. La salida del sistema consiste en una base de datos, lo que facilita la selección automática de palabras para la creación de listas según criterios lingüísticos prefijados por el investigador. Con este sistema, se analizó un conjunto de textos publicados en Cuba, creándose diccionarios de frecuencia de uso de vocablos y sílabas que incluyen otras variables lingüísticas. Este sistema se ha empleado también para realizar otros estudios linguísticos regionales en España y Colombia.

ABSTRACT. The use of words as stimuli is very frequent in behavioral research. For this reason, the study of linguistic characteristics of a language is very important. There exist some previous studies in Spanish language, but they have some bias due to three main factors: the amount of information analyzed is very small, the geographical origin of the study differs from the place where their results will be used, or they have lost validity because of the aging of the material used in its development. The use of computers allows to analyze a great volume of information in short time, it confers more reliability to the studies because it prevents human errors, and it facilitates the constant updating of the information. This paper describes a set of computer programs, which allows to extract several linguistic variables from texts automatically, such as frequency of use, length, grapheme-phoneme irregularities, number of syllables, orthographic stress, syllabic division pattern and consonant-vowel pattern. The output of the system is a database which facilitates the search of words and syllables to make lists according to linguistic criteria given by the researcher. Dictionaries of frequency of use of words and syllables that include other linguistic variables were created analyzing with this system a group of texts published in Cuba. The system has also been used to carry out other regional linguistic studies in Spain and Colombia.

INTRODUCCION

La evaluación de procesos psicológicos complejos tales como los relacionados con el lenguaje, el pensamiento, la lectura y la comprensión, resultan de gran interés tanto en el plano pedagógico como en el de la salud. Sin embargo, una adecuada evaluación de ellos depende en gran medida de un diseño apropiado de las tareas involucradas, lo que implica tener en cuenta para su construcción, las variables que afectan el proceso que se está midiendo.

El uso de material verbal para la evaluación de procesos psicológicos es algo muy frecuente en las investigaciones en el campo de las neurociencias, así como en el diagnóstico y la rehabilitación de procesos psicológicos dañados por diferentes causas. Es por ello que se requiere del conocimiento de las características lingüísticas de los vocablos en el idioma de que se trata, tales como la longitud, el tipo sintáctico, la concordancia grafema-fonema, el patrón consonante-vocal, la frecuencia de uso y la estructura interna, es decir, las características de las sílabas y letras que componen las palabras. Todas estas variables tienen un probado efecto sobre el reconocimiento verbal, por lo que es muy importante tenerlas en cuenta al diseñar una prueba de diagnóstico o una tarea de rehabilitación.

Una de las características cuyo estudio reviste mayor importancia, es la frecuencia de uso, pues se ha demostrado ampliamente el papel decisivo que su manipulación desempeña en los resultados de la in-

vestigación neuropsicológica y psicolingüística.[1-9]

Existen múltiples trabajos que reportan el estudio de la frecuencia de uso de las palabras en otros idiomas como en inglés,[10-14] francés,[15] chino,[16] etcétera. Sin embargo, aunque la lengua española es hablada por más de trescientos millones de personas en el mundo como primera lengua, para este idioma son escasos los estudios de frecuencia, y los que existen presentan limitaciones para ser usados en otro país que no sea el de procedencia. El trabajo más completo que existe en el español es el de Juilland y Chang-Rodríguez, publicado en 1962 y realizado sobre una muestra literaria española de los años treinta.[17] Sin embargo, por la antigüedad de la muestra tomada para el estudio, el resultado es poco representativo de las actuales tendencias del idioma, pues muchas palabras han modificado su frecuencia de uso, e incluso otras se han incorporado al repertorio común.

Aún más escasos resultan los estudios que abordan las características de la estructura interna de las palabras, es decir de las sílabas y letras que las componen. Algunos autores, como De Vega y colaboradores, han encontrado que la frecuencia de la sílaba en el idioma español influye notablemente en el reconocimiento de las palabras.[18]

Estudios de características lingüísticas del idioma español realizados por algunos investigadores como respuesta a sus necesidades investigativas, presentan diversos problemas que los limitan en cuanto a su representatividad, y por ende, en su utilidad para otro medio.[18-20] De estos problemas, los más importantes son, entre otros, los siguientes:

- La cantidad de información que analizan es muy pequeña. Esto puede provocar un sesgo en los datos y por tanto, falsear los resultados que se obtienen al seleccionar el material verbal para una determinada tarea.
- La diferencia entre la procedencia geográfica de los materiales usados para realizar el estudio y el lugar donde se van a aplicar sus resultados. Por esta causa pueden ser incluidos como frecuentes regionalismos propios del lugar de donde procede el estudio que no sean frecuentes en otro sitio, y por el contrario, pueden no aparecer considerados otros vocablos de uso frecuente para tal región.
- Resulta extremadamente difícil tener acceso al resultado de estos trabajos, pues no suele publicarse su texto íntegro.

La solución óptima para estos problemas sería la realización *in situ* de estudios amplios de las características lingüísticas del idioma. Realizar esto manualmente sería un trabajo sumamente engorroso debido al gran volumen de información que se debe procesar para obtener los datos necesarios. Sin embargo, la introducción de los medios de computación permite analizar una cantidad mucho mayor de información en un tiempo considerablemente menor y con gran confiabilidad, ya que reduce significativamente el riesgo de error humano debido a la fatiga, a la vez que no pone límites a las posibilidades de mantener actualizada la información.

El objetivo principal del presente trabajo fue desarrollar un sistema automatizado para el procesamiento de textos en español, que incluyera el análisis de la frecuencia de uso, así como de otras características lingüísticas de las palabras y las sílabas que componen las palabras. Adicionalmente, se propuso corroborar en la práctica la utilidad del sistema para la creación de diccionarios de frecuencia de uso de las palabras escritas en el idioma español.

MATERIALES Y METODOS

Elaboración del Sistema Automatizado para la Confección de Diccionarios de Frecuencia (SACDI)

Se elaboró un sistema de programas desarrollados en lenguaje Borland Pascal, el cual incluye un conjunto de algoritmos basados en la grámatica y la ortografía de la lengua española. Esto permite extraer, de manera automática, información lingüística a partir de una entrada de ficheros en formato ASCII que contengan textos escritos en español.[21, 22]

Estos programas, en una primera etapa de procesamiento, realizan una secuencia de operaciones sobre el texto para crear el Diccionario de Frecuencias de Palabras, que son las siguientes:

- Eliminación de los caracteres que no tienen validez para el estudio (números, signos de puntuación y caracteres de control del editor de textos que se usó originalmente en el proceso de composición).
- Conversión de las palabras en mayúsculas a minúsculas para garantizar la homogeneidad del material a procesar. Esta operación puede realizarse ya sea de manera completamente automática o supervisada por el usuario. Esta última opción se incluye para eliminar los nombres propios, que no son considerados en estudios de esta índole.
- Ordenamiento alfabético de las palabras teniendo en cuenta las características del español (incluyendo las letras **ll** y **ch**, y considerando iguales las vocales con y sin acentuación).
- Conteo de la cantidad de veces que cada palabra aparece en los textos (Frecuencia de uso).
- Conteo de la cantidad de letras que tiene cada una de las palabras (longitud ortográfica).
- División automática en sílabas, a partir de un algoritmo que emplea las reglas ortográficas del español. Esta operación permite extraer las variables siguientes:
  a) Cantidad de sílabas que tiene cada una de las palabras (longitud fonológica).
  b) Clasificación que le corresponde a cada palabra según el tipo de acentuación.
  c) Patrón de división silábica.
  d) Frecuencia silábica media absoluta y frecuencia silábica media posicional de cada palabra.
- Asignación a cada palabra del correspondiente patrón fonológico vocal-consonante, es decir, de la secuencia de vocales y consonantes que tiene la palabra.
- Asignación a los verbos en infinitivo de una etiqueta que los identifica como tales.
- Asignación de la etiqueta correspondiente a las palabras que presentan algún tipo de irregularidad grafofonemática. Se clasificaron como irregulares aquellas palabras que presentan letras con las ambigüedades grafema-fonema siguientes: **c** cuando tiene sonido /s/, **g** cuando tiene sonido /j/, y seguida de vocal, **r** entre dos vocales y presencia de las letras **v** o **h**.
- Creación y manejo de un banco de datos (en DBase III) donde se guarda toda la información extraída del material procesado.
- Análisis de la composición de las frecuencias que se obtienen del procesamiento de los textos y su graficación en forma de histograma.

En este punto del procesamiento, la salida de SACDI está constituida por una base de datos en

DBase III (Tabla 1) que constituye el cuerpo del Diccionario de Frecuencias de Palabras.

Una vez que se tiene el Diccionario de Frecuencia de Palabras y a partir del patrón de división silábica, el sistema de programas realiza un nuevo conjunto de operaciones sobre las palabras para crear el Diccionario de Frecuencias de Sílabas. Las operaciones que realiza entonces son las siguientes:

- Ordenamiento alfabético de las sílabas teniendo en cuenta las características del español (incluyendo las letras **ll** y **ch**, y considerando como iguales las vocales con y sin acentuación).
- Conteo de la cantidad de veces que aparece cada una en las palabras (Frecuencia de uso), así como la sumatoria de las frecuencias de las palabras de procedencia (Frecuencia total).
- Conteo independiente de la cantidad de veces que aparece en cada una de las posiciones posibles dentro de las palabras (Frecuencia posicional), así como de las veces que aparece en la última posición de la palabra.
- Conteo de la cantidad de letras que tiene cada una de ellas (longitud ortográfica).
- Asignación a cada sílaba del correspondiente patrón fonológico vocal-consonante, es decir, de la secuencia de vocales y consonantes.
- Creación y manejo de un banco de datos (en DBase III) donde se guarda toda la información extraída del material procesado.
- Análisis de la composición de las frecuencias que se obtienen del procesamiento de los textos y su graficación en forma de histograma.

En este punto del procesamiento, la salida de SACDI es una base de datos en DBase III (Tabla 2) que constituye el cuerpo del Diccionario de Frecuencias de Sílabas.

Una vez calculada la frecuencia de las sílabas, SACDI permite realizar operaciones que extraen nueva información de las palabras contenidas en la base de datos del Diccionario de Frecuencias de Palabras como la siguiente:

- Cálculo del valor promedio de la frecuencia de las sílabas que componen cada palabra (frecuencia silábica media absoluta).
- Cálculo del valor promedio de la frecuencia posicional de las sílabas que componen cada palabra (frecuencia silábica media posicional).

Así, en este momento del procesamiento de la información, se agregan, en la estructura de la base de datos que compone el Diccionario de Frecuencias de Palabras, dos campos más (Tabla 3).

Independientemente de que los resultados más importantes que se obtienen al aplicar SACDI a un conjunto de textos se almacenan en bases de datos, lo que facilita el manejo de la información a través de programas comerciales como el DBase III, el sistema incluye otros programas secundarios que permiten realizar procesamientos posteriores de los datos recopilados. Tal es el caso del cálculo y la graficación de los histogramas de las frecuencias contenidas en las bases de datos, del cálculo de la frecuencia de cada tipo diferente de patrón consonante vocal y de la búsqueda de las características lingüísticas de una determinada lista de palabras contenida en la base de datos, teniendo como entrada y salida ficheros de texto.

Tabla 1. Estructura de la base de datos donde se almacena la información referente a las palabras extraídas de los textos.

No.	Nombre del campo	Tipo	Número de bytes	Descripción
1	PALABRA	C	25	Nombre de la palabra.
2	FRECUENCIA	N	8	Veces que aparece en el texto.
3	NLETRAS	N	2	Cantidad de letras.
4	FUNCION	C	10	Clase gramatical.
5	REGULARID	C	1	Irregularidad grafema-fonema.
6	NSILABAS	N	2	Cantidad de sílabas.
7	ACENTO	C	1	Clasificación por su acentuación.
8	DIVSIL	C	40	Patrón de división silábica.
9	PATRONCV	C	25	Patrón fonológico consonante-vocal.

Tipo: se refiere a la naturaleza de los datos, que pueden ser en forma de caracteres (C) o de números (N).
Número de bytes: espacio que ocupa cada campo en la base de datos, según las definiciones del Dbase III.

Tabla 2. Estructura de la base de datos donde se almacena la información referente a las sílabas procedentes de las palabras.

No.	Nombre del campo	Tipo	Número de bytes	Descripción
1	SILABA	C	11	Nombre de la sílaba.
2	FRECUENCIA	N	8	Veces que aparece en el texto.
3	NLETRAS	N	2	Cantidad de letras.
4	FRECPAL	N	5	Suma de la frecuencia de las palabras.
5	PATRONCV	C	11	Patrón consonante-vocal.
6 . . . 15	FRECPOS 1 hasta FRECPOS11	N	4	Frecuencia posicional de la sílaba (para cada una de las 11 posiciones posibles).
16	FRECPOSULT	N	5	Frecuencia posicional de la sílaba cuando aparece en la útima posición de la palabra.

Tabla 3. Campos que se agregan en la estructura de la base de datos que compone el Diccionario de Frecuencias de Palabras.

No.	Nombre del campo	Tipo	Número de bytes	Descripción
10	FRECABS	N	5	Frecuencia silábica media absoluta.
11	FRECPOS	N	5	Frecuencia silábica media posicional.

Empleo de SACDI para la creación de Diccionarios de Frecuencia

Se empleó el sistema SACDI para la realización de estudios de las características del idioma español en textos tanto para adultos como para niños, dentro y fuera del país. Así, repondiendo a las necesidades de la investigación, se creó un Diccionario

rio de frecuencia de uso del español escrito en Cuba para adultos, un Diccionario Infantil de la frecuencia de uso del español escrito en España y otro similar en Colombia. A continuación se describe la utilización del sistema para la confección de los Diccionarios de frecuencias de uso del español escrito en Cuba.

Para la obtención del material a analizar, fueron utilizados como fuentes los textos publicados en el periódico "Granma" durante el primer trimestre del año 1992. Fue escogida esta publicación porque resulta el medio de comunicación escrito de mayor cobertura nacional, frecuencia de salida y hábito de consumo, lo que hace que el resultado obtenido de su análisis tenga validez para un amplio sector de la población cubana.

Con el objetivo de evitar que la frecuencia de uso obtenida estuviera sesgada por la reiteración de la información que se produce cuando se muestrea un período corto de tiempo y para incluir por tanto una mayor cantidad de vocablos diferentes, se analizó la información publicada diariamente durante tres meses, la cual fue recopilada en forma de ficheros ASCII.

RESULTADOS

Del procesamiento de los textos escogidos a través del sistema creado para ello, se obtuvo un Diccionario de Frecuencia de uso del español escrito en Cuba (DICFREC), cuyas características se describen a continuación.

La cantidad total de palabras extraídas de los textos fue 494 911, de las cuales 26 860 son diferentes. Esta cifra representa el 32,16 % de los vocablos aceptados en 1992 por la Real Academia Española en la edición 21 de su Diccionario, es decir su tercera parte, lo que constituye una cantidad considerable y hace que este estudio sea representativo del lenguaje español.[23]

El intervalo de frecuencias de uso encontrado (Tabla 4) oscila entre 1 y 42 877, sobre el total de palabras analizadas (494 911), lo que representa una frecuencia entre 2,02 y 86 611,54 expresada sobre un millón.

La figura 1 muestra la composición de frecuencias del diccionario, después de realizada una transformación logarítmica de los datos. Obsérvese que el intervalo de frecuencias en unidades logarítmicas que abarca el diccionario está comprendido entre 0 y 10,66.

Se plantea que el tiempo de procesamiento de las palabras disminuye entre 50 y 75 ms por cada unidad logarítmica de frecuencia.[24,25] El análisis de la composición de frecuencias del diccionario a partir del histograma de frecuencias en forma de unidades logarítmicas permite tener en cuenta estos criterios para seleccionar las palabras que se van a usar en la investigación.

El diccionario está compuesto por un gran número de palabras de frecuencias muy bajas, la mayor parte de ellas de clase abierta (sustantivos, adjetivos, verbos, adverbios) entre las que se encuentran, adjetivos como **abúlico**, sustantivos como **conucos** y verbos como **incautar**, todos con frecuencia 1. La suma de estas palabras que tienen una frecuencia igual a 1 representa el 35,74 % del total de palabras. En el otro extremo del diccionario, se encuentran unas pocas palabras de muy alta frecuencia, casi en su totalidad de clase cerrada. Entre las 20 más frecuentes, se encuentran seis preposiciones, cuatro artículos, tres adverbios, dos pronombres, dos contracciones, una conjunción, una forma verbal conjugada y forma pronominal de un verbo. Las palabras más frecuentes son la preposición **de** (42 877), el artículo **la** (25 395) y la preposición **en** (19 427). La Tabla 5 muestra el valor que para estas 20 palabras, toman las variables numéricas de la base de datos, antes analizadas de manera general para todo el diccionario.

Tabla 4. Comportamiento de las variables numéricas que se recogen en la base de datos que compone el Diccionario de Frecuencias de Palabras.

Variables numéricas	Mínimo	Máximo	$\bar{X}$	DE
Frecuencia	1	42 877	18,42	382,74
Cantidad de letras	1	22	8,63	2,52
Cantidad de sílabas	1	11	3,57	1,08
Frecuencia silábica media absoluta	1	3 089	673,016	380,789
Frecuencia silábica media posicional	1	2 169	253,906	193,596

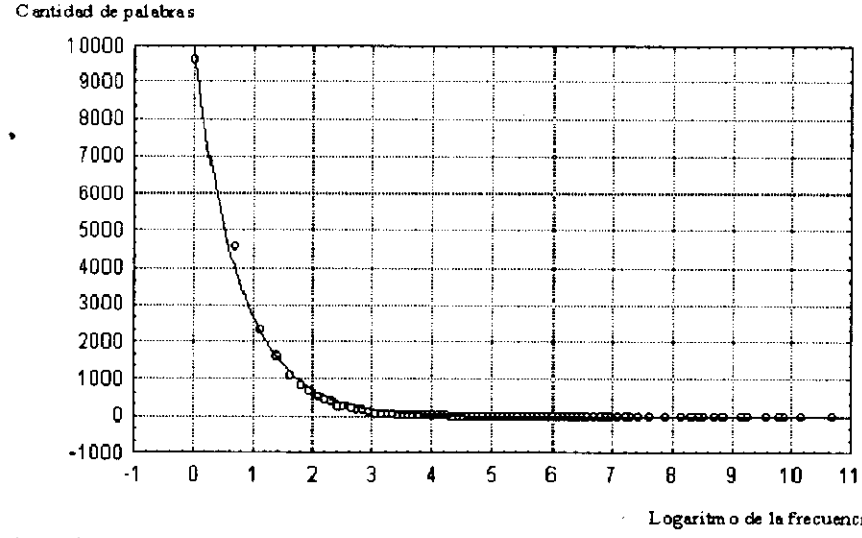

Fig. 1. Distribución de las palabras en el diccionario según su frecuencia. Para una mejor representación de los datos, y por ser frecuente su empleo en la literatura, se realizó una transformación logarítmica de la frecuencia de las palabras.

Tabla 5. Comportamiento de las variables numéricas de la base de datos que compone en Diccionario de Frecuencia de Palabras para las 20 palabras más frecuentes.

Variables numéricas	Mínimo	Máximo	$\bar{X}$	DE
Frecuencia	2 011	45 877	9 845,100	10 131,76
Número de letras	1	4	2,450	0,83
Número de sílabas	1	2	1,150	0,37
Frecuencia silábica media absoluta	2	3 089	612,300	727,53
Frecuencia silábica media posicional	1	2 169	321,200	512,02

Del total de palabras se encontraron 1 191 verbos en infinitivo diferentes, lo que representa el 4,434 % sobre el total de palabras diferentes.

Con los criterios de irregularidad grafofonemática usados, quedaron clasificadas como irregulares 12 885 palabras, para un 47,97 % del total de palabras diferentes, es decir, casi la mitad de ellas. Según el tipo de acentuación ortográfica, 5 251 palabras se clasificaron como agudas, 19 795 como llanas y 1 814 como esdrújulas, para un 19,25; 73,65 y 6,75 % respectivamente sobre el total de palabras diferentes.

De las palabras extraídas de los textos se encontraron 96 199 sílabas, de las cuales 1 911 son diferentes. En la Tabla 6 se muestran los valores tomados por las variables numéricas que se recogen en la base de datos que compone el Diccionario de Frecuencias de Sílabas (DICSIL).

La figura 2 presenta gráficamente la composición de frecuencias de las sílabas. Puede apreciarse que el intervalo de frecuencias que abarca el diccionario de sílabas, expresado en unidades logarítmicas, está entre 0 y 8,03.

Se encontraron 37 patrones consonante-vocal diferentes. De ellos, los tres más frecuente fueron: CV, CVC y CVV, mientras que los tres más infrecuentes fueron: VVCV, VCVVC y CVVCV.

Tabla 6. Comportamiento de las variables numéricas de la base de datos que compone el Diccionario de Frecuencias de Sílabas.

Variables numéricas	Mínimo	Máximo	$\bar{X}$	DE
Frecuencia	1	3 089	50,363	182,701
Suma de la frecuencia de las palabras	1	54 904	563,676	2 408,125
Cantidad de letras	1	7	3,261	0,774
Frecuencia de la sílaba en la posición:				
1	0	2 169	14,060	78,694
2	0	775	13,968	48,038
3	0	940	11,930	48,845
4	0	966	7,134	39,820
5	0	353	2,524	16,300
6	0	116	0,591	4,500
7	0	53	0,128	1,499
8	0	9	0,021	0,268
9	0	2	0,003	0,064
10	0	1	0,000 5	0,022
11	0	1	0,000 5	0,022
Ultima	0	1 601	14,061	71,769

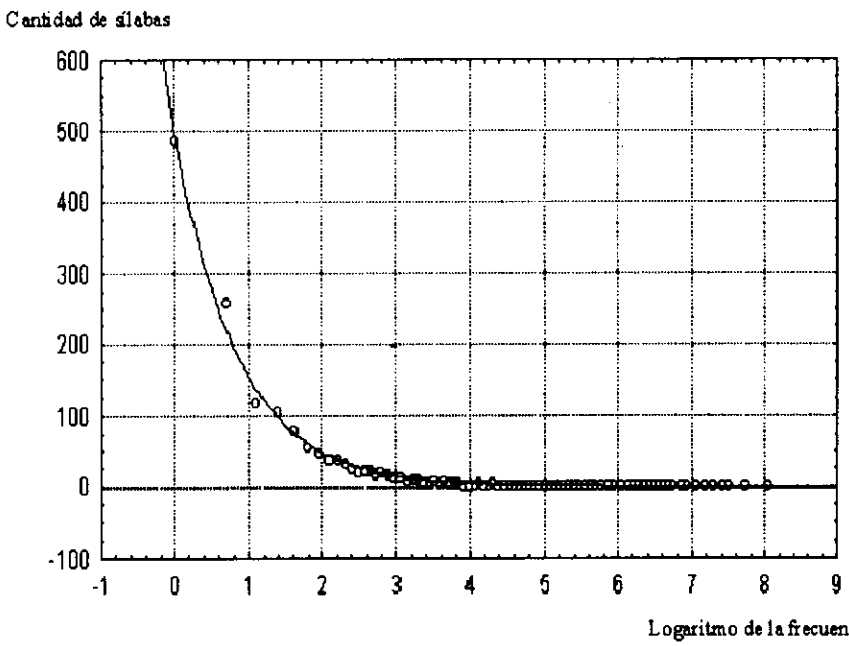

Fig. 2. Distribución de las sílalabas en el diccionario según su frecuencia. Para una mejor representación de los datos, y por ser frecuente su empleo en la literatura, se realizó una transformación logarítmica de la frecuencia de las palabras.

CONCLUSIONES

Se desarrolló un sistema automatizado para el procesamiento de las características lingüísticas de las palabras en el idioma español (SACDI). El uso de este sistema posibilita realizar en tiempo breve estudios que tengan en cuenta fuentes de información de diferentes grados de especificidad y que por tanto, involucren a diferentes sectores poblacionales, con la consiguiente repercusión que esto tiene para la investigación.

SACDI se empleó para realizar un estudio de las características linguisticas del español escrito en Cuba, que culminó con la creación de Diccionarios de frecuencias de uso de las palabras y sílabas. Estos instrumentos han sido ampliamente usados para el diseño del material de estimulación empleado en múltiples pruebas psicológicas y pedagógicas utilizadas en diversas investigaciones

El material resultante de este trabajo, constituye una valiosa fuénte para estudios posteriores, acerca del uso y abuso que se hace del léxico en los medios de difusión, para realizar propuestas a la Real Academia de la Lengua Española acerca de la inclusión de vocablos consagrados por su uso en Cuba, y otros.

BIBLIOGRAFIA

1. Howes D. On the interpretation of word frequency as a variable affecting speed of recognition: **J. exp. Psychol., 48**, 106, 1954.
2. Howes D. On the relationship between intelligibility and frequency word occurrence of English words. **J. Acoust. Soc. Am., 29**, 296, 1957.
3. García-Albea J. E., Sánchez-Casas R. M. y del Viso-Pabón S. Efectos de la frecuencia de uso en el reconocimiento de palabras. **Investigaciones Psicológicas, 1**, 24, 1982.
4. Gordon B. and Caramazza M. Lexical decision for open- and closed-class words: Failure to replicate differential frequency sensitivity. **Brain and Language, 19**, 335, 1982.
5. Segui J., Mehler J., Frauenfelder U.and Morton J. The word frequency effect and lexical access. **Neuropsicologia, 20**, 615, 1982.

6. Elliot L. L., Clifton L. A. B. and Servi D.G. Word frequency effects for a closed-set word identification task. **Audiology, 22**, 229, 1983.
7. Balota D.A. and Chumbley J.I. Are lexical decision a good measure of lexical access? The role of word frequency in the neglected decision stage. **Journal of Experimental Psychology, Human Perception and Performance, 10**, 340, 1984.
8. Rugg M. Event related potentials dissociate repetition effects of high and low frequency words. **Memory and Cognition, 18**, 367, 1985.
9. Tainturier M.J., Tremblay M. and Roch A. Eduaction level and the word frequency effect: A lexical decision investigation. **Brain and Language, 43**, 3, 1992.
10. Thorndike E. and Lorge I. The teacher's word book of 30 000 words, New York, Columbia University Press, 1944.
11. Carroll J. B., Davies P. and Richman B. The American Heritage World Frequency Book, Houghton Mifflin, Boston, 1971.
12. Wepman J. and Hass W. A spoken word count (children-ages 5, 6, and 7), Language Research Associates, Chicago, 1969.
13. Kucera H. and Francis W. N. Computational Analysis of Present-Day American English, Brown University Press, Providence, 1967.
14. Francis W. N.and Kucera H. English Frequency Analysis of English usage: Lexicon And Grammar, Houghton Mifflin, Boston, 1982.
15. Imbs P. Etudes statistiques sur le vocabulaire francais. Dictionnaire des fréquences. Vocabulaire littéraire des XIXe et XXe siecles, Centre de Recherche pour un Trésor de la Langue Francaise, Paris, 1971.
16. Beijing Language College. Modern Chinesse frequency dictionary, Beijing Language College Press, 1986.
17. Juilland A. and Chang-Rodríguez E. Frequency Dictionary of Spanish Words, Mouton, La Haya, 1964.
18. De Vega M., Carreiras M., Gutiérrez-Calvo M. y Alonso-Quecuty M. Lectura y Comprensión. Una perspectiva cognitiva, Alianza Psicología, Madrid, 1990.
19. Corrales I. Consideraciones sobre la confección de diccionarios de frecuencia. **Revista de Filología de la Universidad de la Laguna, 0**, 93, 1981.
20. Lecuona M. P. El lenguaje en la educación infantil, Ins. Univ. de Ciencias de la Educación Ed., Universidad de Salamanca, España, 1991.
21. Montrose M. Lo esencial del lenguaje castellano. Edit. Silver, Burdett y Co., Nueva York, 1900.
22. Real Academia de la Lengua Española. Esbozo de una Nueva Gramática de la Lengua Española. Espasa-Calpe, S.A., Madrid, España, 1973.
23. Diccionario de la Academia: Palabras, Palabras, Palabras ..., Revista Bohemia (Cuba), 1993.
24. Rubenstein H., Garfield L. and Millikan J. Homographic entries in the internal lexicon. **Journal of Verbal Learning and Verbal Behavior, 9**, 487, 1970.
25. Scarborough D.L., Cortese C. and Scarborough H.S. Frequency and repetition effects in lexical memory. **Journal of Experimental Psychology: Human Perception and Performance, 3**, 1, 1977.