RECURSOS INFORMÁTICOS PARA EL ANÁLISIS DEL LENGUAJE

M. JOSÉ NAVARRO PERALES
Dpto. de Didáctica, Organización y Métodos de Investigación.
Universidad de Salamanca

RESUMEN

Las investigaciones sobre lenguaje verbal requieren frecuentemente una serie de trabajos previos de clasificación, recuento, etc., lentos y laboriosos que generalmente son realizados manualmente por el investigador.

Con objeto de paliar este problema hemos construido un programa informático que resuelve de forma eficaz la descomposición de palabras en sílabas, su recuento y clasificación, así como su ubicación en determinados Tipos Silábicos y Estructuras Silábicas y su posición en las palabras.

La eficacia del programa se centra en ser apto para la descomposición silábica de cualquier texto escrito en castellano sin necesidad de manipular previamente el texto y es susceptible de su utilización en un ordenador personal. Además, posibilita una tercera ventaja más importante. Este programa admite también la descomposición silábica de listados de palabras con sus frecuencias asociadas.

El programa resultante puede considerarse de muy aceptable, obteniéndose una descomposición silábica correcta al 100% en textos cuyas palabras no presentan ni hiatos ni diptongos y del 95% en los casos que sí los presentan.

Por los resultados obtenidos con la utilización de este programa para una investigación sobre silabeo infantil consideramos que este programa puede ayudar no solamente a los investigadores dedicados al estudio del lenguaje, sino también a todas aquellas personas que se dediquen a la enseñanza del mismo, puesto que puede usarse como seleccionador de la dificultad de las palabras castellanas basadas en la dificultad silábica.

COMPUTER RESOURCES FOR ANALYSIS OF LANGUAGE

SUMMARY

Research on verbal language frequently requires a series of prior studies of classification, counting, etc., which are slow and laborious and generally carried out manually by the researcher.

In order to mitigate this problem we have designed a computer program which efficiently resolves the division of words into syllables, their counting and classification, as well as their situation in certain Syllabic Types and Syllabic Structures and their position in words.

The efficiency of the program lies in its capacity to divide into syllables any text written in Spanish without requiring prior manipulation of the text and it can be used on a personal computer. Moreover, it affords a third more important advantage: this program also admits the division into syllables of lists of words with their associated frequency.

The resulting program can be considered as hyghly acceptable, obtaining a 100% correct division into syllables in texts in which the words have neither hiatus nor diphthongs and 95% in cases where these are present.

Given the results obtained with this program in research on children's syllabication, we think that this program may help not only researchers dedicated to the study of language, but also those people dedicated to teaching language, since it can be used for selecting the difficulty of Spanish words based on syllabic difficulty.

RESSOURCES INFORMATIQUES POUR L'ANALYSE DU LANGAGE

RESUME

Les recherches menées sur le langage verbal ont fréquemment besoin d'une série de travaux préalables de classification, d'inventaire, etc. lents et laborieux que le chercheur doit généralement réaliser de forme manuelle.

Dans le but de faciliter cette tâche, nous avons élaboré un programme informatique qui résout de manière efficace la division des mots en syllabes, le dénombrement et classification des syllabes ainsi que leur classification dans certains types Syllabiques et Structres syllabiques, et leur position dans le mot.

L'efficacité du programme réside dans son aptitude à la division syllabique de n'importe quel texte écrit en espagnol sans manipulation préalable ainsi que dans son utilisation dans une P.C. Il possède un troisième avantage encore plus important: ce programme admet aussi la division syllabique de listes de mots avec leurs fréquences associées.

Les résultats obtenus peuvent être considérés comme très acceptables, obtenant une division sillabyque 100% correcte dans des textes dont les mots ne présentent ni hiatus ni diphtongues et 95% dans les textes où ils y sont présents.

Ces résultats obtenus dans le cadre d'une recherche sur la division syllabique infantile, nous font penser que le programme pourra être d'une grande utilité non seulement aux chercheurs intéressés à l'étude du langage, mais aussi à tous les enseignants de langue, puisqu'il pourra être employé comme indicateur de la difficulté des mots espagnols d'après leur difficulté syllabique.

1. INTRODUCCIÓN

Las investigaciones sobre lenguaje verbal, requieren, frecuentemente, una serie de trabajos previos, lentos y laboriosos de partición, clasificación, recuento etc., de los términos lingüísticos que se vayan a analizar. Estos trabajos, generalmente, el investigador tiene que realizarlos manualmente. El programa informático que a continuación describimos, surge como alternativa para reducir el tiempo y el coste de esos trabajos previos.

El programa informático que presentamos surge como respuesta al problema que plantea el recuento de sílabas. En el momento en que se lleva a cabo la elabora-

ción de este programa se estaba realizando una investigación sobre las características subléxicas del lenguaje infantil.

La elección del nivel subléxico como objeto de estudio se apoya en la intuición de que en castellano, la sílaba, es el escalón intermedio entre la representación de letras y la representación fonológica de las palabras. Si esto fuera así, la sílaba podría ser un elemento operativo para enseñar a leer. Nuestra intuición ha sido confirmada por de Vega (1991). Este autor confirma que el lector castellano, reagrupa las letras en sílabas pronunciables.

Esta confirmación nos alentó en nuestra investigación. Pero analizar el nivel silábico requiere un trabajo previo lento, laborioso y manual que consistía en separar las palabras en sílabas y proceder después a su recuento y agrupamiento. Estos trabajos suponían un excesivo tiempo y esfuerzo dado el voluminoso número de palabras con las que se contaba. Además, teníamos una cierta preocupación porque el recuento manual careciera de escaso rigor científico.

Nuestras indagaciones con respecto a la existencia de algún programa que permitiera estos trabajos fue totalmente infructuoso. Lo único que encontramos fue alguna aplicación que contenía una opción de rotura de palabras, pero presentaba múltiples errores en la separación silábica y además lo hacía de palabra en palabra.

Ante esa realidad abordamos la confección del programa informático que presentamos a continuación.

2. Alcance del programa

La primera etapa en el desarrollo del Programa, fue la definición de su alcance. Era necesario concretar que era lo que el Programa debería hacer. Esta etapa es determinante en el desarrollo del trabajo por dos razones.

En primer lugar, no se pueden olvidar las limitaciones de un programa informático. No se pueden pretender procesos que impliquen valoraciones cualitativas, del tipo "mejor que", o, "peor que". Tampoco pueden existir indeterminaciones ni ambigüedades, como "quizá", " tal vez" etc.

En segundo lugar debe conseguirse una relación aceptable entre los resultados obtenidos y el esfuerzo necesario para lograrlos. Esta regla, que se olvida en ocasiones, es básica en cualquier trabajo. En muchos procesos existen alternativas poco probables, cuyo impacto puede apreciarse separadamente, sin excesiva dificultad. Hay ocasiones en las que conseguir que un programa informático recoja y procese correctamente estas alternativas, supone aumentar extraordinariamente la complejidad y el tamaño del programa, y, por lo tanto, los requisitos del ordenador adecuado.

En el ejemplo que presentamos, se pretendía un programa que permitiera conocer determinadas características silábicas de un texto cualquiera escrito en castellano. Como condición complementaria se pretendía que pudiera procesarse en un ordenador personal Macintosh, de características medias, no debiendo superar 4 Mb de RAM.

El objetivo presentaba una amplitud y complejidad apreciables. Téngase en cuenta que, aunque el silabeo en castellano sigue unas reglas bastantes claras y precisas, es notoria la cantidad de excepciones que existen a esas reglas silábicas. Ello supuso obtener una extensa información lingüística que permitiera asumir las

excepciones a las reglas de partición silábicas y preparar todas las secuencias posibles contemplando las excepciones.

El programa tendría que ser capaz de identificar, clasificar y agrupar las sílabas, los tipos silábicos y las estructuras silábicas presentes en cualquier texto que se fuera a analizar. Como resultado el programa tendría que proporcionar las siguientes salidas:

a.- Listado ordenado de sílabas presentes en función de su tipo y en orden alfabético, junto con las palabras asociadas. Número de palabras diferentes en que aparece cada sílaba y número total de ocurrencias de dicha sílaba.

b.- Listado de sílabas presentes. Para cada Tipo se obtendrían: el Número de Sílabas Diferentes que pertenecen a ese Tipo, Número Total de Ocurrencias de ese Tipo de Sílaba y Frecuencia Media de las sílabas pertenecientes a ese Tipo.

c.- Listado de Estructuras Silábicas presentes ordenadas en función del número de sílabas de las palabras. Para cada Estructura Silábica se pedía: Número de palabras diferentes con esa estructura, Número total de Ocurrencias de dicha estructura y Frecuencia Media de las palabras correspondientes.

d.- Para cada subconjunto de palabras con diferente número de sílabas, se obtenía, listado ordenado alfabéticamente de las sílabas presentes, con las frecuencias de cada sílaba en las diferentes posiciones posibles y en su totalidad.

e.- El mismo listado descrito en el punto *d* pero para el conjunto total de palabras.

Con objeto de acotar las alternativas que habría que considerar, se decidió que el programa debería realizar todas las operaciones anteriormente descritas para las palabras existentes en castellano y escritas con ortografía correcta. Se prescindió de onomatopeyas, extranjerismos no adaptados al castellano, etc.

Como medida complementaria se consideró conveniente que algunas palabras, poco corrientes, que se apartan de las reglas generales de descomposición silábica, aunque estuvieran correctamente escritas, se excluirían del campo de aplicación del programa, almacenándolas para su posterior partición silábica de forma individual. Entre las excluidas se encuentran palabras como *sublingual* o *sublunar,* etc.

Las decisiones tomadas en relación con los diptongos e hiatos, fueron meditadas con especial atención. Se suponía que al incluir en el programa toda la casuística de los diptongos e hiatos se convertiría en un programa complejo con numerosos errores. Sin embargo no tener en cuenta ni hiatos ni diptongos suponía una pérdida de información que, a nuestro entender, podía ser valiosa en cualquier investigación de este ámbito. Finalmente se optó por incluir tanto diptongos como hiatos, aunque admitiendo que se podrían presentar algunas limitaciones.

Además se pensó que sería conveniente que el programa trabajara con textos escritos pero también con listados de palabras y sus respectivas frecuencias asociadas.

3. Desarrollo del Programa

El desarrollo del programa requirió la solución de numerosos problemas que se pueden agrupar en dos grandes áreas. Una de ellas engloba los problemas derivados de la propia descomposición silábica de las palabras. La segunda comprende todos los problemas relativos al tamaño del programa y la cantidad de memoria requerida.

La descomposición silábica de las palabras, al tener en cuenta las diferencias entre hiatos y diptongos, obligó a definir cuatro planos de análisis diferentes. Esos planos son: Palabras, Sílabas, Tipos Silábicos y Estructuras Silábicas.

El análisis de cada uno de estos planos, enlaza con los análisis correspondientes a los planos restantes. Como consecuencia se produce un entramado de decisiones lógicas que asocia a cada una de las diferentes palabras de un texto o de un listado de palabras, las Sílabas que la forman, los Tipos Silábicos de éstas y la Estructura Silábica de la palabra. En base a estos datos, posteriormente, el programa, clasifica y recuenta las Sílabas, Tipos Silábicos y Estructuras presentes.

La variada casuística que hay que tener en cuenta en la descomposición de una palabra en sílabas generó un programa con gran número de líneas. Además, los diversos listados solicitados, obligan a diferentes procesos de clasificación y recuento. Como resultado se obtuvo un programa de 203 K, que puede procesar textos de cerca de 80000 palabras. Para ello utiliza algo más de 3 Mb de memoria RAM.

4. Conclusiones

El programa resultante puede calificarse de muy aceptable. Las pruebas de contrastes realizadas permiten afirmar que el programa es altamente operativo. Para realizar las pruebas de contraste se utilizaron textos de publicaciones diarias, obteniéndose una descomposición silábica correcta en el 100% de las palabras que no presentaban diptongos o hiatos. En los casos de palabras que presentan la singularidad del diptongo o del hiato, la descomposición silábica resultó correcta en el 95% de los casos comprobados. Todo este contraste se realizó con palabras habituales.

Como limitación al programa señalamos el tiempo de proceso que es superior a la media hora en textos de varios miles de palabras. A pesar de esa limitación el programa es eficaz y ahorra al investigador tiempo y esfuerzo. El tiempo que invierte es mínimo si se le compara con el tiempo que tiene que dedicarse si la descomposición silábica se hace manualmente.

5. Ejemplificaciones de las Salidas que proporciona el Programa

Archivo estudiado:
Pepe vuela en su avioneta

LISTADO DE SÍLABAS CON LAS PALABRAS ASOCIADAS

		Nº S/P	Nº OP	Nº TOS
Grupo: 4				
Sílaba: la				
	vuela	1	1	1
Total sílaba		1	1	1
Sílaba: ne				
	avioneta	1	1	1
Total sílaba			1	1
Sílaba: pe				
	pepe	2	1	2
Total sílaba			1	2
Sílaba: su				
	su	1	1	1
Total sílaba			1	1
Grupo: 12				
Sílaba: vue				
	vuela	1	1	1
Total sílaba		1	1	1
Total tipo				1

LISTADO DE SÍLABAS PRESENTES

TIPO	SILABA	Nº Pal. Dif.	Nº Ocu.S.
1	a	1	1
1	o	1	1
2	en	1	1
4	la	1	1
4	ne	1	1
4	pe	1	2
4	su	1	1
4	ta	1	1
4	vi	1	1
12	vue	1	1

LISTADO DE TIPOS SILÁBICOS

TIPO	Nº Sil. Dif.	Nº Ocu. S.	Fr.M
1	2	2.0e+0	1.00
2	1	1.0e+0	1.00
3	0	0.0e+0	0.00
4	6	7.0e+0	1.17
12	1	1.0e+0	1.00

ESTRUCTURA SILÁBICA DE PALABRAS

PALABRA	Nº Pal. D.	Estructura
Palabras de 1 sílaba:		
en	1	(2)
su	1	(4)
Palabras de 2 sílabas:		
pepe	1	(4, 4)
vuela	1	(12, 4)

NÚMEROS DE PALABRAS SEGÚN SU ESTRUCTURA SILÁBICA

Estructura	Nº P. D.	Nº Oc.	Fr. M.
Palabras de 1 sílabas.			
(2)	1	1	1.00
(4)	1	1	1.00
Palabras de 2 sílabas.			
(12,4)	1	1	1.00
(4,4)	1	1	1.00
Palabras de 5 sílabas.			
(1,4,1,4,4)	1	1	1.00

POSICIÓN DE LAS SÍLABAS SEGÚN EL TIPO DE PALABRAS

Palabras de 2 sílabas.

SIL	Nº Tot. Oc.	Nº Oc. 1º Pos.	Nº Oc. 2º Pos.
la	1	0	1
pe	2	1	1
vue	1	1	0

POSICIONES DE LAS SÍLABAS EN EL TEXTO

Sil	1º Pos	2º Pos	3º Pos	4º Pos	5º Pos	6º Pos	7º Pos	Nº Tot.Oc.
a	1	0	0	0	0	0	0	1
en	1	0	0	0	0	0	0	1
la	0	1	0	0	0	0	0	1
ne	0	0	0	1	0	0	0	1

6. Campos de Aplicación del Programa

Este programa puede tener múltiples aplicaciones. Se nos ocurre que puede ser aprovechado tanto por investigadores, como por docentes, como por aprendices de una lengua.

Con respecto al ámbito investigador su utilización puede interesar a variadas ramas científicas. Así para la Lingüística, la Psicolingüística, la Tecnología Didáctica, Didáctica de la Lengua, etc., puede convertirse en una herramienta útil en el estudio de determinados aspectos léxicos. Además puede interesar a todos aquellos profesionales dedicados a la investigación sobre lenguajes, análisis de textos, fórmulas de lecturabilidad, etc.

A los docentes puede serles útil en el conocimiento de aspectos léxicos de sus alumnos, o bien porque pueden utilizarlo como medio y recurso tecnológico en la enseñanza de la lengua, de la ortografía, la lectura, etc.

Por último, interesa a los alumnos porque pueden contar con un instrumento útil para el aprendizaje de la Ortografía del Castellano, de la Métrica, de la lectura en los aprendices de lector, etc.

BIBLIOGRAFÍA

DE VEGA, M. (1991): *Lectura y comprensión. Una perspectiva cognitiva.* Madrid, Alianza Psicología.