A.Z.R.A.E.L

A-Z Reconocedor Automático de EspañoL

Elena Álvarez Mellado

Facultad de Filología
Lic. de Lingüística
Universidad Complutense de Madrid

Motivos

¿Es esto español? ¿Y esto?
Berenjena ¿Corulla?
Troglodita ¿Disfagia?
Dinosaurio ¿Alucón?
Clarinete ¿Clánide?
Pespunte ¿Sucinda?

Un individuo es capaz de, dada una palabra, reconocer su idioma (o al menos aventurarlo), aún cuando no conozca su significado.

¿Qué tienen en común las palabras de una lengua?

¿Y si en vez de *clánide* hubieses sido..?

... trock?
...shippet?
...whert?

¿En base a qué sabemos que una palabra pertenece a un idioma?

¿Es posible crear un reconocedor automático en base a ese criterio?

Las palabras ajenas: Existen palabras que son reconocidas como ajenas a la lengua a pesar de tener un significado asociado como la palabra *whisky* -> a pesar de tener significado la percibimos como una palabra ajena al español

Los extranjerismos que conservan su aspecto original producen vacilaciones a los hablantes en la escritura. Sufren un proceso de adaptación desde su forma en la lengua original a la de destino

El mecanismo por el que detectamos una lengua se basa fundamentalmente en el aspecto de las palabras, es un mecanismo formal, independiente del significado.

=> Es posible crear un programa informático que detecte, a través de la forma de las palabras, si una palabra es española

El reto será doble:

1) Conseguir llegar a un patrón que recoja cuál es la forma de las palabras en español (patrón que debe ser lo más general posible pero a la vez exhaustivo)
2) Implementar ese patrón en un programa informático

La sufijación y la prefijación permiten crear palabras que el oyente es capaz de identificar.

¿Existe un conjunto de posibilidades de inicio de palabra y otro conjunto de posibilidades de final de palabra?

Parece que necesitamos moldear las palabras, de forma semejante al proceso que sigue un canto rodado, para incorporarlo a la lengua.

La sílaba

* Primera gramática: Sánscrito (Panini V a.c.)
* Primeros trabajos comparativos:
Dante Alighieri: forma de la palabra *si.*
Joseph Justus Scaliger: La palabra *dios.*

Los detectores de idiomas funcionan en base a:
*Diccionario
*Conectores
*Alfabeto
*Frecuencia de aparición de letras
*Modelo de n-grama

La estructura silábica es característica de cada lengua

scanner -> escáner jerez -> sherry

Método para el estudio de la sílaba en español:
* Corpus de español de 600.000 palabras
* Observar las posibilidades de combinación de cada letra

Consonante + Vocal: Todas las consonantes excepto la Q
Consonantes intercaladas: R,L
Consonantes emparejadas: B,C,G,F y P (R,L) T y D (admiten R)
CH,LL,RR,QU
Consonantes finales:R,S,L N,D,Z

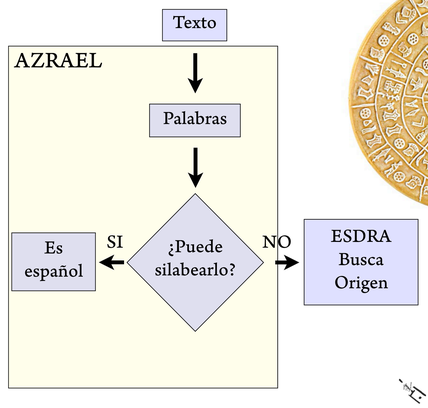
Final de sílaba	Contexto consonántico	Casos en el corpus	Ejemplos
C	t	8627	ac-tuar
	c	2189	ac-ce-der
	n	275	a-rác-ni-do
	m	28	drac-ma
	s	19	fuc-sia
F	t	332	dif-te-ria
	g	18	af-ga-no
	n	8	haf-nio
G	n	2487	ig-no-rar
	m	696	dia-frag-ma
	d	42	a-míg-da-las
M	p	21054	pom-pa
	b	16549	bom-ba
	n	564	a-lum-no
	m	52	gam-ma
	l	10	drum-lin
	h	6	brom-hí-dri-co
P	t	2972	a-cep-tar
	c	413	op-ción
	n	121	hip-no-sis
T	m	194	at-mós-fe-ra
	l	115	at-le-ta
	n	72	et-nia
B	s	1024	ob-ser-var
	t	413	sub-te-rráneo
	v	367	ob-vio
	d	327	ab-di-car
	j	269	ob-je-to
	c	250	ob-ce-car
	y	175	ab-yec-to
	n	89	ob-nu-bi-lar
	m	66	sub-ma-ri-no
X	p	2569	ex-pan-dir
	t	2459	ex-tra-er
	c	1573	ex-cep-ción
	h	397	ex-ha-lar
	q	53	ex-qui-si-to

C. pseudofinales | Extranjerismos | Prefijos

El programa

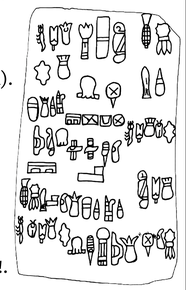
Para realizar este trabajo he creado tres programas:

- AZRAEL, el programa general de reconocimiento automático de español
- Silabeador, subprograma de AZRAEL que silabea palabras siguiendo la estructura de la sílaba del español
- ESDRA (Etimólogo Selectivo del Reconocedor Automático), programa encargado de buscar la lengua de origen de los extranjerismos que han sido rechazados por el Silabeador.



```
español(X,Restos,Origen):-
        español(X,Restos),
        origen_etimológico(Restos,Origen).

origen_etimológico([],[]).
origen_etimológico([Restos|MasRestos],[Origen|MasOrigen]):-
    atom_chars(Restos,Deletreo),
    combinaciones(Restos,Deletreo,Origen),
    !,
    origen_etimológico(MasRestos,MasOrigen).
silabear_aux([],[]).
silabear_aux([Z|T],[Sílaba|Sílaba_resto]):-
        posibilidades(Z,Silabeo,[Z|T],Resto),
        atom_chars(Sílaba,Silabeo),
        silabear_aux(Resto,Sílaba_resto).
posibilidades(s,[s,u,b],[s,u,b,Cons|W],Resto):-
        cons(Cons),
        append([s,u,b],Resto,[s,u,b,Cons|W]), !.
posibilidades(o,[o,b],[o,b,Cons|W],Resto):-
        cons(Cons),
        not(consPareja(b,[b,Cons])),
        append([o,b],Resto,[o,b,Cons|W]),  !.
posibilidades(a,[a,b],[a,b,Cons|W],Resto):-
        cons(Cons),
        not(consPareja(b,[b,Cons])),
        append([a,b],Resto,[a,b,Cons|W]), !.
posibilidades(e,[e,x],[e,x,Cons|W],Resto):-
        cons(Cons),
        append([e,x],Resto,[e,x,Cons|W]),  !.
```



Conclusiones

La sílaba como unidad de reconocimiento de una lengua.

La sílaba es un campo investigación inexplorado.

Tradicionalmente, los estudios descriptivos de una lengua saltan del nivel fonético (los sonidos de una lengua) al morfológico (afijos y mecanismos de derivación).

Existe una una lógica en la manera en que las letras se combinan para formar sílabas y es propia de cada lengua.

Ventajas de usar la sílaba como elemento caracterizador:

- Es un criterio universal: todas las lenguas pueden ser segmentadas en sílabas.
- Es un criterio tipológico: la estructura silábica es característica a cada lengua.
- Es un criterio formal: se fundamenta en la forma. Esto permite el tratamiento automático del lenguaje. Asimismo, la forma es el criterio por el que los hablantes reconocen el idioma de una palabra, lo que significa que AZRAEL reconoce el idioma de una palabra por el mismo criterio que un hablante.
- Es un criterio estable: un diccionario del siglo XVII no nos sirve para caracterizar español del siglo XX, ya que constantemente se crean palabras nuevas y se abandonan otras. Las palabras van y vienen sin que ello signifique que nos encontremos en un idioma distinto. Sin embargo, la estructura de la sílaba permanece inmutable durante los siglos, ya que es una característica de la estructura profunda de la lengua.

A través de la forma de la palabra podemos obtener mucha información:
* Época de entrada de la palabra
* Lengua de origen.
* Vía de ingreso.

Para los prefijos, puesto que aportan significado, las leyes de la sílaba son más flexibles. Se prefiere mantener el contenido léxico que adaptar la palabra a la lengua.

Aquellas palabras españolas que no responden al patrón general de la sílaba son palabras importadas desde otras lenguas, de entrada tardía y que pertenecen al ámbito de la lengua especializada: son cultismos provenientes del griego.

Origen etimológico		
Combinaciones	Origen	Ejemplos
TL	Griego o náhuatl	atleta, tlacoyo
TM	Griego	atmósfera
TN	Griego	etnia
PN	Griego	hipnosis
FT	Griego o árabe	difteria, muftí
GM	Griego	diafragma
GD	Griego	amígdalas
CN	Griego	arácnido
IT	Latinismo	déficit