ESTRUCTURA BÁSICA DEL LÉXICO MONOLINGÜE ÁRABE DE UN SISTEMA DE TA

INFORMACIÓN MORFOLÓGICA, SINTÁCTICA Y SEMÁNTICA

Nebot E.1, Alonso J.A2, Herrero I., Blanquer H.

Resumen

En este artículo se presenta una descripción del léxico monolingüe árabe del sistema de traducción automática árabe-español que se está desarrollando en los departamentos de Árabe y Lingüística de la Universitat de Barcelona como parte de un proyecto de colaboración con el Instituto Bourguiba de Túnez. El artículo presenta en primer lugar los requisitos, funciones y estructura del léxico. Más adelante, se centra en la organización interna del léxico, en el formato de las entradas léxicas y en los diferentes tipos de información que contiene cada una de ellas.

1. Marco

Exponemos a continuación los criterios básicos seguidos en la construcción de un Léxico Monolingüe Árabe (LMA) para el sistema de Traducción Automática Asistida Árabe-Español que se está desarrollando en la Universitat de Barcelona (Depto. Árabe y Sección de Lingüística General) en colaboración con la Universidad de Túnez (Instituto Bourguiba de Lenguas Vivas).

Este sistema está basado en la arquitectura de transfer y se ha intentado que el diccionario, tanto por su diseño como por su contenido, sea reutilizable para otras aplicaciones distintas de la traducción automática (TA). En otras palabras, se ha pretendido que pueda constituir un producto lingüístico en sí mismo.

El léxico se ha venido considerando el eje central sobre el que giran los diferentes componentes del lenguaje, a saber: la fonética, la morfología, la sintaxis, la semántica y la pragmática. Cumple la función de 'contenedor' del conocimiento e información

1 Universitat de Barcelona, Departamento de Árabe; e-mail: nebotlingua.fil.ub.es

2 INCYTA S.L.; e-mail: juanincyta.es

acerca del lenguaje y suele desempeñar un papel crucial en la resolución de las ambigüedades semánticas. En palabras de M.Y. Al-Hafez (cf. {1}): "The lexicon has the responsibility for the success of the NLP applications. Any failure in fulfilling a successful access to a lexicon subsystem stops the whole NLP system whatever the application is. The lexicon must provide to the NLP application as much evidence as possible in order to solve the ambiguous states and rectify the lack of knowledge that the computation experiences".

2. Requisitos del LMA

Los requisitos de diseño planteados pretendían que el módulo tuviera las siguientes características:

- *Compacto*: se debe tener el máximo de información en el mínimo de entradas léxicas. Como veremos más adelante, el hecho de basar el diccionario en raíces y no en lexemas nos ha permitido reducir muchísimo el tamaño del mismo y mantener las entradas agrupadas en familias léxicas.

- *Neutro*: se ha buscado que la estructura de representación del conocimiento léxico sea lo más neutra e independiente posible, de modo que no condicione las operaciones de importación, consulta y, en general, de manejo de datos del léxico desde los otros módulos.

- *Flexible:* se ha intentado que el léxico sea de fácil ampliación y mantenimiento, e incluso que, en caso de que sea necesaria alguna modificación en su estructura, ésta no suponga un rediseño y reimplementación drásticos.

3. Funciones del LMA

A continuación se indican las funciones más importantes del LMA en el sistema de traducción automática que nos ocupa:

- Desambiguación de la salida del analizador morfológico (AM): el AM proporciona todas las interpretaciones lingüísticamente plausibles de cada palabra árabe que analiza. No obstante, muchas de esas formas no existen en realidad. Por lo tanto, las interpretaciones dadas por el AM deben cotejarse con el léxico monolingüe árabe para ver cuáles de ellas existen realmente3.

- Guiar y restringir el proceso de análisis sintáctico (AS): parte de la información existente en el LMA (por ejemplo, la información de concordancia gramatical, de rección léxica y de subcategorización, entre otras) se utiliza por la gramática de análisis, por un lado para restringir la aplicación de las reglas de estructura de frase durante el proceso de análisis, y por otro para poner en los nodos del árbol que se va construyendo la información necesaria para sucesivas etapas del análisis y para las fases siguientes de transferencia y generación.

- Proporcionar información necesaria para la fase de transfer: el LMA contiene

3 Poniendo un ejemplo en castellano, el AM nos podría dar una interpretación de la palabra "pinar" como 'verbo en infinitivo'. Sin embargo, al acudir al léxico monolingüe castellano comprobaríamos que el verbo 'pinar' no existe, por lo que se desecharía esa interpetación.

también información léxica que se utiliza en la frase de transferencia léxica para desambiguar las traducciones.

```
drb   --> $(subj (np,hum, obl), dobj (np, inst, obl))$
      --> $(subj (np,_, obl), dobj (np,_,obl))$
```

Por ejemplo, el verbo "daraba" (raíz: **drb**) en árabe, subcategoriza por un lado un sintagma nominal ('np') que desempeña la función de sujeto y cuyo tipo semántico debe ser 'humano' y un objeto directo que es también un sintagma nominal ('np') con tipo semántico 'instrumento musical', en cuyo caso se traducirá como "tocar"; por otro lado, subcategoriza un sujeto y un objeto directo que son dos sintagmas nominales con tipo semántico sin especificar, traduciéndose entonces como "golpear".

⟨ Proporcionar la información necesaria para la fase de generación (Español-Árabe): en el supuesto de que se genere árabe, la información morfológica presente en el LMA (clase de declinación/conjugación, patrones de plural/femenino, etc.) es esencial para poder formar correctamente las palabras árabes correspondientes.

4. Estructura del LMA

El LMA está organizado en 2 módulos principales:

⟨ el ***léxico*** propiamente dicho
⟨ el ***módulo de gestión*** del mismo

4.1 *El léxico*

El ***léxico*** consiste a su vez de:

⟨ las bases de datos léxicas (BDLs)
⟨ las tablas de conocimiento lingüístico

4.1.1 BDLs

Las BDLs constituyen el núcleo del léxico; contienen las entradas léxicas con su información lingüística; hay 5 BDLs o subléxicos: BDNombres, BDVerbos, BDPartículas, BDNE (a saber: Nombres Extranjeros, Nombres Propios y Nombres Bilíteros y Quintilíteros, Demostrativos y Cuantificadores) y BDAuxiliares.

4.1.2 Tablas

En las tablas se encuentra estructurado todo el conocimiento lingüístico necesario para el procesamiento de los datos en algún momento u otro del proceso de traducción.

He aquí un ejemplo del tipo de información recogida en las tablas. Vemos la estructura del predicado "bd_rasg". Así, por ejemplo, el primer registro nos dice que este predicado tiene un posible rasgo "AFIJ", que indica si la "partícula" en cuestión

-"AFIJ" es un rasgo de la entrada del léxico de Partículas- puede ir afijada a un nombre, verbo a otra partícula o a varios de ellos a la vez, o si, por el contrario, no es afijable (los valores correspondientes a este rasgo estarán en la estructura del predicado "bd_rasval" (BD que relaciona los rasgos con sus posibles valores).

BD_RASG

```
info( bd_rasg, [$AFIJ$,$partícula afijada a n,v,p, o bien suelta$]).
info( bd_rasg, [$ARGS$,$frame de subcategorización$]).
info( bd_rasg, [$BAB$,$bab del verbo en formal$]).
info( bd_rasg, [$CAN$,$forma canónica$]).
info( bd_rasg, [$CF$,$código flexivo verbal (modelo)$]).
info( bd_rasg, [$CLASS$,$clase morfológica$]).
info( bd_rasg, [$COM$,$comentario, ejemplo uso ARGS$]).
info( bd_rasg, [$CTEM$,$código temático$]).
info( bd_rasg, [$DECL$,$declinación$]).
info( bd_rasg, [$FORM$,$forma del verbo$]).
info( bd_rasg, [$GEN$,$género$]).
info( bd_rasg, [$MOD$,$modelo flexivo del verbo$]).
info( bd_rasg, [$NUM$,$número$]).
info( bd_rasg, [$NV$,$nombre verbal$]).
info( bd_rasg, [$PLUPAT$,$patrón de plural$]).
info( bd_rasg, [$SGFEM$,$forma el fem. sing. añadiendo #$]).
info( bd_rasg, [$TADJ$,$tipo de adjetivo$]).
info( bd_rasg, [$TADV$,$tipo de adverbio$]).
info( bd_rasg, [$TAUX$,$tipo de verbo auxiliar$]).
```

4.2 *Módulo de gestión*

El ***módulo de gestión*** contiene:

- el interfaz de usuario (IU)
- el módulo gestor

4.2.1 IU

El IU contiene una serie de opciones que permiten manipular el léxico de una forma cómoda y eficaz. Mediante el IU se pueden dar de alta y de baja, consultar o modificar entradas léxicas.

4.2.2 Módulo gestor

El módulo Gestor se ocupa de la intercomunicación entre módulos. Por ejemplo:

- gestiona el acceso del AM al léxico.
- organiza la información que el léxico proporciona a las fases de análisis sintáctica, transferencia y generación.

5. Organización del LMA

5.1 *Raíces vs. lexemas*

Una de las primeras decisiones que hubo que tomar respecto al léxico, fue la de seguir un enfoque basado en raíces u otro basado en lexemas. Recordamos que las palabras árabes se forman a partir de una raíz, generalmente triconsonántica, a la que se aplica un conjunto de patrones de vocalización con el fin de formar unos lemas a los que, a su vez, se pueden añadir una serie de afijos. Naturalmente, se podría haber optado por el segundo enfoque, listando en el léxico monolingüe todos los posibles lexemas para una palabra árabe y olvidándonos de las raíces y los patrones de vocalización. Sin embargo, y entre otros inconvenientes, el tamaño del diccionario se habría disparado excesivamente y se habría desaprovechado la compacidad que la regularidad morfológica de la lengua aporta.

No obstante, pensamos que el enfoque basado en raíces es mucho más apropiado para la naturaleza léxica y morfológica del árabe. Contrariamente a lo que ocurre con muchas lenguas europeas, en árabe puede deducirse mucha información lingüística esencial a partir de la forma particular del propio lexema (es decir, del patrón de vocalización) y también, claro está, de sus afijos. Este aspecto es particularmente interesante para el tratamiento durante la fase de análisis de palabras que no se encuentren en el léxico monolingüe (resulta relativamente fácil deducir la categoría y cierto número de rasgos gramaticales).

Asimismo, la decisión de analizar la palabra árabe descomponiéndola en raíz y patrón (cf. {9}), estuvo fuertemente ligada a la de adoptar un enfoque basado en raíces para los léxicos del sistema.

Finalmente, un factor determinante en la adopción de este enfoque basado en raíces fue la práctica imposibilidad de averiguar el singular que corresponde a un patrón particular de plural fracto (es plural fracto el formado por derivación no lineal y, por lo tanto, de patrón impredecible, a partir de las mismas tres o cuatro consonantes del singular, es decir, de la raíz). Teniendo en cuenta, además, que cada plural fracto tiene un número bastante elevado de potenciales singulares a partir de los cuales pueda haberse formado, la única manera de encontrar en el léxico árabe la entrada léxica singular correspondiente es a través de la raíz y del patrón que nos proporciona el análisis morfológico y que constituyen la entrada léxica del diccionario.

5.2 *Formato de la entrada léxica*

- forma de citación
  - raíz (3ms Perfecto), "forma" y "bab" para los verbos plenos y verbos auxiliares.
  - raíz y "patrón de vocalización" del singular para los nombres y adjetivos que no flexionan en género.
  - raíz y "patrón de vocalización" del masculino singular para los nombres y adjetivos que flexionan en género.
  - forma canónica (forma sin flexión) y forma base (forma canónica sin vocalizar) para las partículas, nombres extranjeros, nombres propios, nombres bilíteros y quintilíteros, demostrativos y cuantificadores.
- variante de la raíz y/o del patrón
- categoría sintáctica

- información morfológica
- información sintáctica
- información semántica
- información adicional de ayuda al lexicógrafo (comentarios, forma canónica...)

5.3 *Criterios lingüísticos: categorías gramaticales*

5.3.1 Categorías léxicas mayores y categorías léxicas menores

Hemos respetado la tradición gramatical árabe que divide las palabras en nombres, verbos y partículas, grupos que constituyen nuestras "categorías léxicas mayores"; éstas se subdividen a su vez en una serie de "categorías léxicas menores" (a las que hemos dado el nombre de "subcategorías"). Estas "subcategorías" se clasifican ulteriormente en tipos específicos: conjunción copulativa, pronombre relativo, adjetivo demostrativo, etc...

A continuación damos en detalle la clasificación arriba mencionada:

CATEGORÍA MAYOR	SUBCAT	TIPO	DESCRIPCIÓN
n (nombres)	nom	np	nombre propio
		nc	nombre común
		nclas	nombre clasificador
		nnumc	nombre numeral cardinal
		nnumo	nombre numeral ordinal
	adj	adjc	adjetivo común
		acomp	adjetivo comparativo
		aquant	adjetivo cuantificador
		adem	adjetivo demostrativo
		anisba	adjetivo nisba
		anumc	adjetivo numeral cardinal
		anumo	adjetivo numeral ordinal
v (verbos)	vb	rec	verbo recíproco
		pas	verbo siempre en pasiva
		nopas	verbo nunca en pasiva
		ref_lex	verbo reflexivo léxicamente
		ref_pas	verbo que expresan pasiva refleja o ergatividad
		ref_nafs	verbo que puede formar reflexivo con 'nafs'
	vaux	temporal	verbo auxiliar temporal
		modal	verbo auxiliar modal
		aspectual	verbo auxiliar aspectual
p (partículas)	prep		preposición
	pro	ppers	pronombre personal
		ppos	pronombre posesivo
		prel	pronombre relativo
		pind	pronombre indefinido
		pint	pronombre interrogativo
		pdem	pronombre demostrativo
		pquant	pronombre cuantificador
		pnumc	pronombre numeral cardinal

		pnumo	pronombre numeral ordinal
	adv	avdeic	adverbio deíctico
		avloc	adverbio locativo
		avtemp	adverbio temporal
		avneg	adverbio de negación
		avint	adverbio interrogativo
		avafir	adverbio de afirmación
		avman	adverbio de manera
		avrel	adverbio relativo
		avfut	adverbio de "futuro"
		avcor	adverbio corroborativo
		avqad	adverbio 'qad'
	intj		interjección
	conj	ccop	conjunción copulativa
		cdis	conjunción disyuntiva
		cadvs	conjunción adversativa
		cexpl	conjunción explicativa
		cexc	conjunción exceptiva
		cscond	conjunción subordinada condicional
		cstemp	conjunción subordinada temporal
		csloc	conjunción subordinada locativa
		csfin	conjunción subordinada final
		cscau	conjunción subordinada causal
		csconc	conjunción subordinada concesiva
		cscons	conjunción subordinada consecutiva
		cscompl	conjunción subordinada completiva
		csexc	conjunción subordinada exceptiva
		cscomp	conjunción subordinada exceptiva

5.3.2 **Categorías Abiertas y Categorías Cerradas**

Categorías abiertas son aquéllas que constituyen una lista abierta (ampliable y modificable por parte del usuario).

- nombres (nom)
- adjetivos (adj)
- verbos plenos (vb)

Categorías cerradas son las que podemos describir por enumeración puesto que se trata de listas cerradas (no ampliables) de elementos. Son categorías funcionales.

- preposiciones (prep)
- adverbios (adv)
- pronombres (pro)
- conjunciones (conj)
- interjecciones (intj)
- verbos auxiliares (vaux)

5.4 *Criterios de implementación*

La razón principal que nos ha empujado a dividir el LMA en varios léxicos específicos es el distinto tratamiento que reciben las palabras según sean o no *derivadas de raíz*.

En este sentido, decidimos separar en léxicos distintos las palabras derivadas de raíz de las que no lo son. Los léxicos de palabras derivadas de raíz están ordenados alfabéticamente por raíces y los de palabras no derivadas de raíz están ordenados alfabéticamente por lexemas.

La división en tres grandes categorías -*nombres, verbos y partículas*- se ve reflejada también en la estructura del LMA, que contiene (entre otros) un léxico específico de nombres derivados de raíz y con patrón reconocible (**BDN**), un léxico de verbos (**BDV**) -lógicamente derivados de raíz puesto que en árabe no existen verbos "primitivos" ni extranjeros y las raíces bilíteras y quintilíteras son eminentemente nominales - y finalmente un léxico de partículas (**BDP**), por definición no derivadas de raíz.

Las diferencias en la *información lingüística* presente en las "fichas" de nombres y en las de verbos (algunos atributos obligatorios son comunes pero muchos de ellos no lo son) han contribuido a la decisión de crear un léxico específico de nombres y otro de verbos.

Por otra parte, hemos querido separar de manera clara las palabras *funcionales* de las que no lo son. Así, aún cuando todas las "partículas" son no-derivadas, la razón que nos ha llevado a agruparlas en un léxico específico de partículas (BDP) ha sido la de pertenecer sus categorías a la clase funcional, cosa que no sucede con las palabras del léxico BDNEt.

Otro de los criterios adoptados ha sido el de procurar agrupar en un único léxico todas aquellas palabras pertenecientes a *categorías cerradas*. Recordamos que las categorías cerradas son las que no son susceptibles de ser aumentadas. La coherencia con la adopción de dicho criterio nos ha llevado a considerar los pronombres (que tradicionalmente pertenecen a la categoría "nombre") como "partículas", por cuanto consituyen un grupo cerrado. Además, el hecho de que los pronombres sean categorías funcionales y no derivadas de raíz refuerza esta decisión.

Hemos considerado a los *adverbios* como "partículas" - siguiendo de nuevo la tradición gramatical árabe - puesto que cumplen dos de los requisitos "no estrictos" arriba mencionados para pertencer a esta categoría: por un lado, constituyen una clase cerrada (nos referimos, claro está, a los adverbios formados léxicamente; de los que se forman sintácticamente se ocupa la sintaxis), y por otro lado, no derivan de raíz (salvo excepciones que no cuestionan la decisión adoptada).

Mención aparte merece el caso de *los demostrativos y cuantificadores*, que son entradas del léxico de nombres extranjeros, nombres propios, quintilíteros y bilíteros, demostrativos y cuantificadores (**BDNE**). Como se observará, este léxico engloba categorías muy heterogéneas, pero que tienen en común dos cosas:

- No son derivadas de raíz: Esto significa que, a pesar de ser casi todas ellas nombres y adjetivos (extranjeros, n. propios, n. quintilíteros y bilíteros, etc.), no pueden figurar en el léxico de nombres (BDN) como los demás nombres y adjetivos.

- La mayoría de ellas son nombres y adjetivos: Esto impide que pertenezcan al léxico de partículas (BDP). Como ya hemos apuntado, únicamente los pronombres demostrativos y cuantificadores podrían pertenecer al BDP. Los adjetivos

demostrativos y cuantificadores no podrían ser entradas de dicho léxico y, por consiguiente, se ha decidido considerar conjuntamente a estos pronombres y adjetivos como entradas léxicas de la BDNE.

A continuación, especificamos las categorías que pertenecen a cada uno de estos léxicos en función de las razones expuestas.

a) Son entradas léxicas de la **BDNE:**

-nombres extranjeros
-nombres propios
-nombres quintilíteros
-nombres bilíteros
-demostrativos (pro,adj)
-cuantificadores (pro,adj,n)

b) Son entradas léxicas de la **BDP**:

-preposiciones
-conjunciones
-pronombres (excepto demostrativos y cuantificadores)
-adverbios
-interjecciones

c) Son entradas léxicas de la **BDN**:

-nombres comunes
-nombres extranjeros con raíz y patrón (préstamos arabizados)
-nombres clasificadores
-nombres numerales cardinales y ordinales
-adjetivos numerales cardinales y ordinales
-adjetivos comunes
-adjetivos comparativos
-adjetivos nisba

d) Son entradas léxicas de la **BDV**:

-verbos plenos

e) Son entradas léxicas de la **BDVAUX**:

-verbos auxiliares

6. Formato de las entradas léxicas

En las siguientes tablas vemos la estructura de rasgos y valores para cada uno de los léxicos arriba descritos, ejemplificada con entradas reales.

RASGOS BDN	VALORES BDN		RASGOS BDV	VALORES BDV
RAIZ	ktb		RAIZ	jry
VAR			VAR	
CAT	n		CAT	v

PAT	1i2aa3	SUBCAT	vb
VARPAT		FORM	1
CAN	**kitaab**	BAB	a_u
VARCAN		NV	1u2uw3
SUBCAT	nom	GEN	m
GEN	m	ARGS	(subj(np,anim,obl), (pobj(prep("min"),np,inam,obl)
NUM		COM	
PLUPAT	1u2u3	TY	act
SGFEM		TV	
DECL	trip	CTEM	gral
COL		CF	V001
TY	doc		
TN	nc		
TADJ			
ARGS			
CTEM	gral		

RASGOS BDP	VALORES BDP	RASGOS BDNE	VALORES BDNE
CAN	**qad**	CAN	**bint**
BASE	**qd**	BASE	**bnt**
CAT	p	VARCAN	**ban**
SUBCAT	adv	VARBASE	**bn**
DECL	inv	CAT	n
AFIJ		SUBCAT	nom
TPRO		GEN	f,f
TCONJ		PLUR	psf
TADV	avqad	DECL	trip,dipp
GNP		TY	hum
TGRAM	i	TN	nc
PSUF		TADJ	
RCAS		TPRO	
RNUM			
TREAL	fut		
ASP	prog		
MODAL	pos		

RASGOS BDAUX	VALORES BDAUX
RAIZ	**kwn**
VAR	**k'n**
CAT	v
SUBCAT	vaux
FORM	1
BAB	a_u
TAUX	temp
TGRAM	i
TREAL	past
ASP	prog

MODAL	
RESTT	
RESTP	

7. Definición de los campos

7.1 *Información gráfica*

Estos campos contienen información sobre:

- Transliteración adoptada. Ausencia de vocales
- Necesidad de estandarización del texto a traducir
- Variantes ortográficas (morfografémicas en raíces, patrones y lemas):"tarwiih"=""taryiih"; defectivos en 'alif, waw,'alif maqsuura; cóncavos en 'alif, waw, yaa, etc...
- Homógrafos

Algunos de los rasgos que utilizamos para representar esta información son:

- CAN: forma canónica
- VAR: variante grafémica de la raíz
- VARCAN: variante de la forma canónica
- VARBASE: variante de la base
- BASE: forma canónica sin vocalizar

7.2 *Información morfológica*

Estos campos contienen información sobre:

- Clases morfológicas (conjugación y declinación).
- Patrones aplicados a las raíces.
- Formas verbales derivadas y 'bab'.
- Formas femeninas y tipos de plurales.
- Formación de los nombres verbales.
- Afijabilidad de partículas y posibilidad de que éstas tomen pronombres sufijados.
- Rección de caso y número de las partículas.
- Restricciones sobre verbos defectivos.

Algunos de los rasgos empleados son:

- RAÍZ: Raíz de la palabra
- FORM: Forma de conjugación verbal
- PAT: Patrón que aplicado a la raíz forma la palabra
- GEN: Género gramatical

- PLUPAT: Patrón de plural
- NUM: Número gramatical
- CF: Código flexivo verbal
- SGFEM: Formación del femenino
- DECL: Declinación
- PLUR: Tipo de plural
- COL: Colectivo y/o nombre de unidad
- TGRAM: Tiempo gramatical
- RCAS: Rección de caso
- RNUM: Rección de número
- AFIJ: Afijabilidad
- PSUF: Pronombre sufijado.

7.3 *Información sintáctica*

Estos campos contienen información sobre:

- categoría y subcategoría.
- subcategorización verbal, nominal y adjetival a nivel funcional, sintagmático y temático
- tipos de nombre, adjetivo, pronombre, adverbio y verbo.

Algunos de estos rasgos son:

- CAT: Categoría gramatical.
- SUBCAT: Subcategoría gramatical.
- ARGS: Estructura argumental de nombres verbales, verbos y de ciertos adjetivos.
- TN: Tipo de nombre.
- TADJ: Tipo de adjetivo.
- TPRO: Tipo de pronombre.
- TADV: Tipo de adverbio.
- TV: Tipo de verbo.
- TAUX: Tipo de auxiliar.

7.4 *Información semántica*

La información dada trata sobre:

- Tipos semánticos de nombres, verbos y adjetivos.
- Aspecto, temporalidad y modalidad.
- Subcategorización semántica.

Algunos de los rasgos utilizados son:

- TREAL: Tiempo real
- MODAL: Modalidad
- ASP: Aspecto
- ARGS: Estructura argumental (a nivel semántico)
- TY: Tipo semántico
- CTEM: Código temático.

8. Acceso al léxico

El acceso al léxico se realiza a través de una tabla de acceso léxico rápido que desempeña dos funciones:

- sirve de puntero a las entradas léxicas del diccionario correspondiente, agilizando de esta forma el acceso al léxico.
- filtra las formas base, raíces y/o patrones no presentes en el léxico.

Los registros de esta tabla (<raíz>, <rafz-patrón>, <forma-base>) se generan automáticamente al dar de alta una entrada léxica en cualquiera de los diccionarios (BDN, BDP, BDV,...). Por cada nueva entrada, la tabla se actualiza con la raíz en todo caso, con los patrones de singular y de plural si se trata de nombres o adjetivos (BDN), y con la forma base si se trata de una partícula (BDP) o entrada del léxico de nombres extranjeros, etc (BDNE).

Así por ejemplo, la palabra **'uluwm** -entrada del BDN- dará lugar en la tabla a una raíz **'lm** y a un patrón ***1u2uw3***: {**'lm**, ***1u2uw3***}. A este registro se le asigna un número que es el puntero al léxico: {**'lm**, ***1u2uw3***, 0037}. Pero resulta que el patrón en cuestión (***1u2uw3***) , además de ser el patrón de plural normal de la entrada léxica **'ilm**, es el patrón principal de otra entrada léxica **'uluwm** (que corresponde a un plural lexicalizado, como en español "padres"). En este caso, habrá tres registros en la tabla, a saber: uno correspondiente al patrón ***1i23*** apuntando a la entrada **'ilm** y dos correspondientes al patrón ***1u2uw3*** para la forma **'uluwm**; uno de estos dos registros corresponde al plural de la entrada léxica **'ilm** (y también apunta a esta entrada), y el otro apunta a la entrada léxica **'uluwm**.

RAÍZ: **'lm**
CAN: **'ilm**
PAT: ***1i23***
PLUPAT: ***1u2uw3***

'ilm = ciencia, conocimiento; **'uluwm** = ciencias, conocimientos

RAÍZ: **'lm**
CAN: **'uluwm**
PAT: ***1u2uw3***
PLUPAT: ---

'uluwm = Ciencias Naturales

9. Implementación

⟨ Requisitos de hardware:
  ⟨ PC Compatibles 386 o superiores.
  ⟨ Mínimo 4Mb RAM; disco duro 1Mb.

⟨ Requisitos de software:
  ⟨ MSDOS 5.0 o superior.
  ⟨ Arity Prolog versión 6.
  ⟨ Sistema operativo árabe SAKHR versión 2.01.

10. Bibliografía

. AlHafez, M.Y., Mrayati, M., Vella, A., Clarke, J.D. 1992. Design of an Arabic Language knowledge-base as a Lexicon for NLP, in: Proceedings of the 3rd International Conference and Exhibition on Multilingual Computing, University of Durham, 10-12 December, 1992.
2. AlHanash, M. 1992. Almu9jim alaaliyy lilluga# al9arabiyya#: qaa9ida# bayaanaat alta9aabiir almaskuuka#.
3. Alonso, M. 1993. Las Funciones Léxicas en el Modelo Lexicográfico de I. Mel'cuk. Tesis Doctoral presentada en el Dept. de Filología Española de la UNED. Madrid. Septiembre, 1993.
4. Brusset, J. & Abdelghani, S. 1989. Création d'une Base de Données Lexicales de l'Arabe Ecrit Utilisable par un Système Morpho-Syntaxique, in: Linguistique Arabe et Informatique: Actes du IV Colloque Internationale de Linguistique, 9-12 Novembre, Tunis, Série Linguistique N°7.
5. Ditters, E. 1994. The Basic Structure of a formal Arabic-English Verbal Lexicon, in: Proceedings of the 4th International Conference and Exhibition on Multilingual Computing (Arabic and Roman Script), London, 7-9 April. ICEMCO. University of Cambridge, Centre of Middle Eastern Studies.
6. Dorr, B. 1991. Conceptual Basis of the Lexicon in Machine Translation, in: Lexical Acquisition. Exploiting On-Line Resources to Build a Lexicon. Uri Zernik (ed.), Lawrence Erlbaum Associates Publishers, Hillside, New Jersey 1991.
7. Grishman, R., MacLeod, C., Meyers, A. 1994. Comlex Syntax: Building a Computational Lexicon, in: COLING 1994, vol.1.
8. Hamzaoui, R. 1975 L'Academie de la Langue Arabe du Caire: Histoire et Oeuvre, Publications de l'Université de Tunis.
. Herrero L., Nebot E., El analizador morfológico del árabe de un sistema de TA árabe-español, en las actas del X Congreso de la S.E.P.L.N., Córdoba, Julio 1994.
10. Jazaa, A. 1992. A Knowledge-Based System for Analizing Arabic Language, in: Proceedings of the 3rd International Conference and Exhibition on Multilingual Computing (Arabic and Roman Script), University of Durham, 10-12 December.
11. Nomura, N., Jones, D.A., Berwick, R.C. 1994. An Architecture for a Universal Lexicon: A case study on shared syntactic information in Japanese, Hindi, Bengali, Greek and English, in: COLING 1994, vol.1.
12. Perry, J.R. 1993. Early Arabic-Persian Lexicography: the "Asami" and "Masadir" genres, in: Proceedings of the Colloquium on Arabic Lexicology and Lexicography, The Arabist, Budapest Studies in Arabic 6-7, Part One.
13. Processing Arabic, Report N°1 (1986), Report N°2 (1987), Report N°3 (1988), Report N°4 (1989), Report N°5 (1990), Nijmegen: TCMO.
14. Saad, G.N. 1982. Transitivity, Causation and Passivization. A Semantic-Syntactic Study of the Verb in Classical Arabic. Library of Arabic Linguistics. Monograph N°4. Kegan Paul International, London, Boston and Melbourne, 1982. Muhammad Hasan Bakalla (Ed).
15. Stetkevych, J. 1975. The Modern Arabic Literary Language: Lexical and Stylistical Developments, The University of Chicago Press, William R. Polk (ed.), Publications of the Center for Middle Eastern Studies, Number 6.
16. Walker, D.E., 1989. Developing Lexical Resources.