Nell'ambito del Piano "Linguistica Computazionale: ricerche monolingui e multilingui" – Cluster C18 – Legge 488/1999,
Finanziato dal Ministero dell'Università e della ricerca Scientifica e Tecnologica (MURST), l'Istituto di Linguistica Computazionale ha collaborato a diversi progetti contenuti nel programma operativo del Piano.
In particolare per la realizzazione di strumenti finalizzati al trattamento della lingua araba l'Istituto ha partecipato a:

Progetto n. 6 "Corpus blingue italiano-arabo"

Capofila del progetto**: Istituto Universitario Orientale di Napoli**

Disegnare e creare un corpus bilingue italiano-arabo di testi comparabili, cioè disegnati, composti, strutturati, collegati e analizzati con gli stessi criteri. Una parte sostanziale di entrambi costituita da testi paralleli, cioè da testi arabi e dalla loro traduzione italiana o viceversa. Un sottoinsieme di entrambi i componenti prevede inoltre l'annotazione, cioè analisi morfosintattica semiautomatica e si avvale quindi delle risorse linguistiche e dei moduli software e sviluppati nel progetto 7.

Progetto n. 7 "Trattamento automatico della morfologia della lingua araba"

Capofila del progetto**: Università degli studi di Pisa – Dipartimento di Scienze Storiche del Mondo Antico**

Creazione del software necessario per l'annotazione morfosintattica semiautomatica dell'arabo. L'impegno prevede lo studio, il disegno, la realizzazione e la validazione di un motore morfologico della lingua araba, costituito da moduli per:

- la creazione di un lessico arabo di base corredato da un insieme di regole morfologiche;
- la fase di generazione, che da un lemma dato è in grado di produrre tutte le sue forme;
- la creazione di un analizzatore morfologico automatico che data una forma riesce a risalire al lemma cui la forma si riferisce.

Report Tecnico
Motore morfologico della lingua araba

Piano "Linguistica computazionale: ricerche monolingui e multilingui"
Progetto 7 "Trattamento automatico della morfologia della lingua araba " ed in particolare i Workpackages n. 1, 2, 3, del Progetto medesimo

Oggetto:

W1 - a1	Definizione del sistema di codifica da utilizzare per la rappresentazione dei dati lessicali e delle caratteristiche morfologiche per la lingua araba; definizione della composizione, della dimensione e della articolazione del "lemmario" arabo; definizione del sistema di codifica, della sintassi e della articolazione dell'archivio "regole morfologiche" di guida all'algoritmo morfologico.
W1 - a2	Definizione delle caratteristiche richieste dalla procedura informatica per la generazione e l'analisi automatica della morfologia per la lingua araba.
W2 – a3	Realizzazione del componente software per la generazione automatica delle forme e per l'analisi morfologica automatica
W3 - a1	Creazione del "lemmario" in formato leggibile dal calcolatore, con codifica e attribuzione delle informazioni grammaticali.
W3 - a2	Identificazione dei gruppi di lemmi in possesso di identici comportamenti morfologici e compilazione delle regole morfologiche secondo codifiche e sintassi definite.
W3 – a3	Inserimento nell'archivio "lemmario" dei codici di flessione appropriati.

Autori: per la parte informatica **Sassolini Eva**
per la parte linguistica **Nahli Ouafae**

Definizione dell'ambiente software

L'ambiente necessario al funzionamento del sistema è il seguente:

- Sistema operativo Windows 95-98-2000;
- Acquisizione del font arabo "Siddiqua.ttf" sotto la cartella c:\windows\fonts\ del computer su cui si installa il programma;

Non esistono indicazioni che ne richiedano l'installazione sotto directory specifiche.

Fase di generazione

La struttura dell'algoritmo necessario al trattamento dei dati si rifà, nella sua struttura più generale, a quella già collaudata utilizzata per la lingua italiana. Per la generazione delle forme di un lemma sono però necessarie le modifiche del sistema di codifica dei meccanismi che rappresentano le regole flessionali, per consentire il trattamento delle parole arabe .

Lo scopo dell'algoritmo è quello di generare tutte le forme di competenza di ogni lemma cioè il suo paradigma.
Congiuntamente il sistema morfologico dovrà fornire la classificazione morfosintattica (genere, numero, persona, ecc.) per ogni forma generata. La predisposizione del sistema per il progetto si compone di due tappe successive : una prima tappa da raggiungere prevede la preparazione di un software tale da permettere la creazione degli archivi (Lemmario e regole morfologiche), la seconda di consentire la verifica del meccanismo di flessione con possibilità di correzioni on-line.

L'algoritmo di generazione vero e proprio è composto dalle seguenti fasi:

- Analisi e segmentazione del lemma dato
- Consultazione del lemmario per la ricerca del lemma, della relativa radice, della categoria grammaticale e della regola di flessione.
  - ✓ Nel caso di un lemma verbale riconoscimento del tipo. Per quelli del primo tipo distinzione tra regolare o irregolare. Successive distinzioni di comportamento flessionale all'interno dei verbi regolari e di quelli irregolari.
- Generazione di tutte le forme sia per i lemmi verbali sia per quelli non verbali riconducibili al lemma dato. Per un lemma verbale è data tutta la coniugazione sia nella forma attiva che passiva. Per quello non verbale sono date tutte le forme comprensive di particelle determinative e indeterminative.

Fase di analisi

Questa fase permette di associare ad ogni forma analizzata il lemma di competenza. Chiaramente più di un lemma può essere riconosciuto come associato ad una singola parola del testo: in questo caso tutti i lemmi possibili vengono assunti con identica priorità.
La fase di disambiguazione, quella cioè che tenta di individuare il lemma corretto di una forma in un determinato contesto, è una fase di analisi linguistica successiva che dovrà aver come punto di partenza l'analisi morfologica che il presente lavoro descrive.
Per la fase di analisi l'algoritmo prevede di utilizzare lo strumento di generazione realizzato. Infatti le componenti utilizzate per la prima fase sono necessarie per verificare l'esattezza delle ipotesi formulate.

L'algoritmo si compone delle seguenti fasi:

- Possibili segmentazioni della forma utilizzando le combinazioni possibili tra radici ammissibili e prefissi, suffissi e infissi noti, in particolare sono sfruttate le risorse linguistiche create per la fase di generazione.
- Per ogni ipotesi di segmentazione identificazione della radice quindi del lemma e generazione di tutte le forme relative.
  - ✓ Per ogni radice irregolare, applicazione delle regole specifiche per i casi irregolari, unitamente ad una ricerca in un apposito dizionario dell'esistenza della corrispondente radice irregolare.
- Nel caso il lemma non appartenga al Lemmario, ma siano state individuate con la "analisi delle particelle" caratteristiche tali da ipotizzare con certezza la natura del lemma, il sistema fornisce comunque il riconoscimento del lemma.

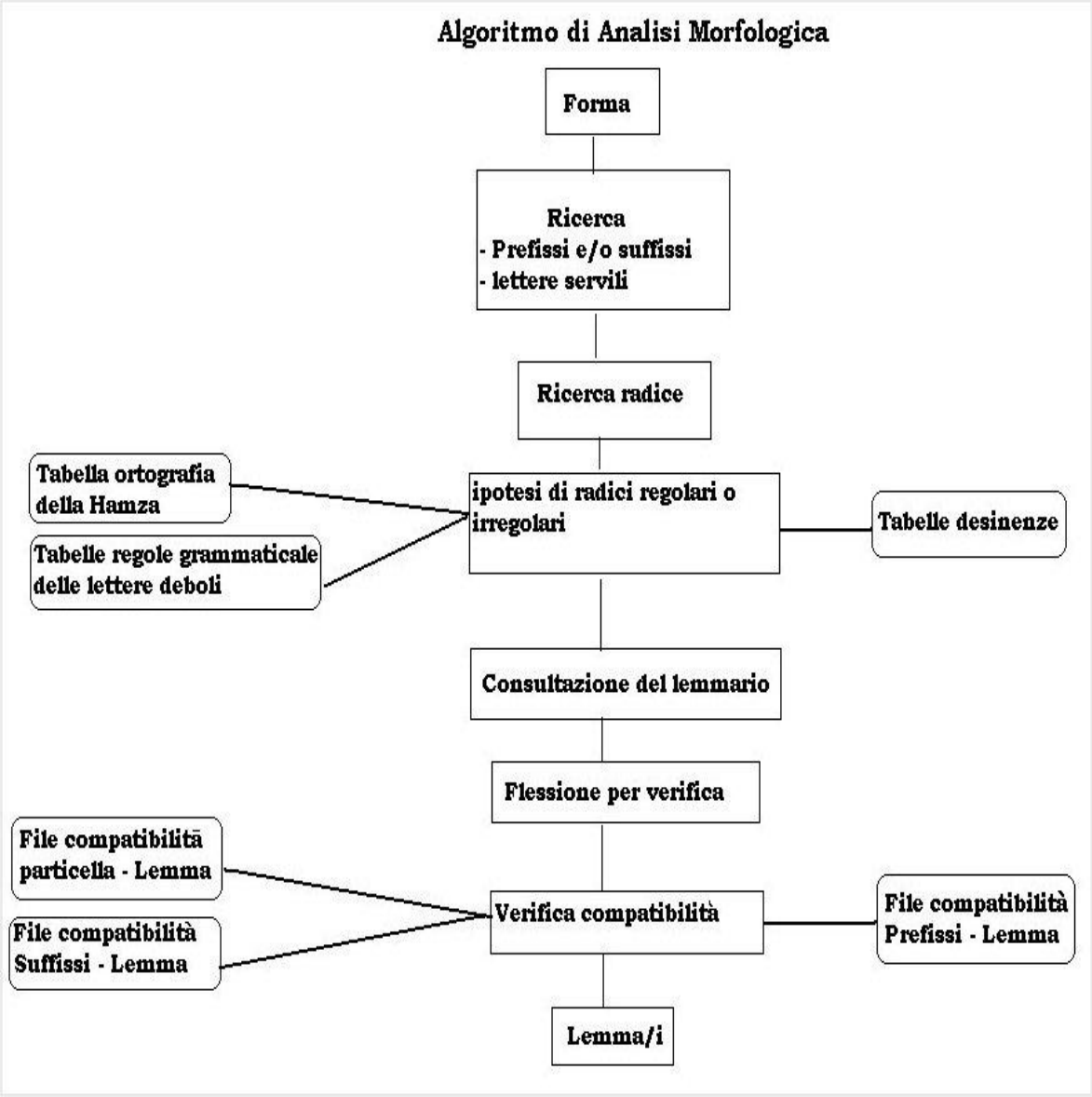

Figura 1: schema dell'algoritmo di analisi

Risorse necessarie

Le risorse indispensabili al corretto funzionamento del Sistema Morfologico sono:

- Un lemmario della lingua. L'archivio si stima dovrà contenere almeno 80.000 lemmi.
- Una tabella contenente la codifica dei caratteri per la scrittura a video delle parole arabe. Questa tabella è utilizzata per permettere la rappresentazione simultanea in più codifiche del lemma.
- Un elenco di regole di riconoscimento che contengono informazioni sulla generazione delle forme. Ogni regola flessionale viene identificata da un codice ed è rappresentata da una sequenza di caratteri che definiscono una particolare operazione. Ognuna di esse è interpretata dal sistema software del motore morfologico e trasformata in una serie di passi da eseguire sulla radice del lemma dato.
- Un elenco di tabelle contenenti le sequenze dei caratteri delle desinenze dei lemmi verbali secondo il loro gruppo di appartenenza (verbi trilitteri, quadrilitteri, regolari, irregolari, ecc.), dei lemmi non verbali (anch'essi raggruppati secondo un criterio basato su identità di comportamento). Elenco comprensivo altresì di tabelle con suffissi e prefissi per i casi che

richiedono questa modalità. Tra queste tabelle sono presenti tutti i passi per la formazione dei diversi plurali fratti, e dei masdar.

- Un archivio contenente un elenco di radici irregolari, cioè una risorsa che contiene una lista di casi in cui l'algoritmo non è in grado di trovare una soluzione coerente. In modo da ottenere comunque un comportamento omogeneo del sistema.

Archivio Lemmario

Struttura dell'Archivio

Questo archivio è strutturato come un Database ad accesso diretto; Con la particolarità di avere come chiave di accesso non il lemma completo ma una sua parte significativa.
L'archivio può avere chiavi di accesso duplicate, in quanto possono esistere lemmi uguali ma con codici grammaticali diversi.
Questo fa si che l'ordinamento del Lemmario sulla sua chiave di accesso non corrisponda all'ordinamento alfabetico stretto del Lemmario

Convenzioni adottate nel lemmario

Lemma:
Il lemma è presente nel Lemmario come una chiave costituita di due elementi:

1) la radice del lemma ottenuta togliendo al lemma le eventuali desinenze e vocali;
2) la sequenza delle vocali presenti nel lemma con l'informazione sulla loro posizione all'interno del lemma completo.

Per i lemmi verbali il lemma su cui costruire la chiave è rappresentato dalla 3° persona maschile singolare del compiuto.

Categoria Grammaticale:
Le principali categorie grammaticali previste sono due:

1: lemma verbale identificato dal carattere V:
all'interno della categoria dei lemmi verbali si fa distinzione tra verbo trilittero, identificato dalla lettera T e verbo quadrilittero denominato dal carattere Q.
Si può inoltre distinguere in tutti i verbi la loro forma primitiva (P) da quella derivata (D).

2: lemma non verbale identificato dal carattere S:
Sempre per i lemmi non verbali si distingue tra nomi (S) e particelle (P).

Gruppo:
Per un lemma verbale che rappresenta un verbo trilittero sono previste più categorie:

1: verbi forti (F) che sono formati da radicali forti.
Questi a loro volta si distinguono in tre sottogruppi:
gruppo di verbi trilitteri che utilizzano la vocale "*a*" per la formazione dell'imperativo (R1);
gruppo di verbi trilitteri che utilizzano invece la "*u*" (R2);
" " " " "*i*" (R3);

2: verbi hamzati (H) che rispettano la stessa suddivisione in sottogruppi dei forti:
R1;
R2;
R3.

3: verbi geminati (G):
R1;
R2;

R3.

4: verbi deboli (D):
R1;
R2;
R3.

5: verbi doppiamente irregolari (IR);
R1;
R2;
R3.

Per un lemma non verbale esiste invece una ulteriore suddivisione in:

1: sostantivi derivati da un verbo (SV), o derivati da un nome (SN)

2: lemmi maschili (M), che possono essere suddivisi in diptoti (P) e triptoti (T), anormale (R), indeclinabile (B)

3: lemmi femminili (F) " " " " " " "

4: lemmi sia femminili che maschili detti comuni (C) " " "

5: lemmi che presentano una forma maschile e una forma femminile, detti mobili (N)

6: particelle (P)

Codice di Flessione:

Numero di tre cifre che indica la regola flessionale da adottare per la generazione delle forme di un lemma sia verbale che non verbale.

Le regole di flessione comprendono sia i passi necessari per il recupero dei prefissi e suffissi opportuni, sia quelli per le modifiche della radice affinché sia possibile generare tutte le forme. Nel caso la regola corrisponda ad un numero di meno di tre cifre occorre anteporre tanti zeri quante sono le cifre che mancano alla formazione di un numero di tre cifre (ad es. per la regola 1 è previsto il codice flessionale 001).

Radice:

Radice del lemma corrispondente alla famiglia di lemmi cui il suddetto si riferisce, molto importante nella lingua araba perché permette la generazione di tutti i lemmi che hanno un comune significato; sia verbali che non verbali. La radice è strumento di guida per la flessione dei verbi geminati e derivati, e per la formazione dei plurali fratti nel caso di lemmi non verbali.

Uso del Lemmario

Questa risorsa è utilizzata dal programma che implementa l'algoritmo di generazione, per ottenere la chiave (radice + vocali) e le informazione ad essa connesse.

Il programma una volta ottenuta la possibile radice del lemma digitato, attraverso l'uso di opportune procedure, ne fa ricerca all'interno del Lemmario. Nel caso che la radice appartenga al Lemmario il programma la utilizza insieme alle informazioni ad essa associate per la flessione del lemma stesso in tutte le sue forme.

Anche i moduli software relativi alla fase di analisi utilizzano la risorsa Lemmario per recuperare il lemma, fletterlo e verificare se la parola del testo è una forma di quel particolare lemma ipotizzato.

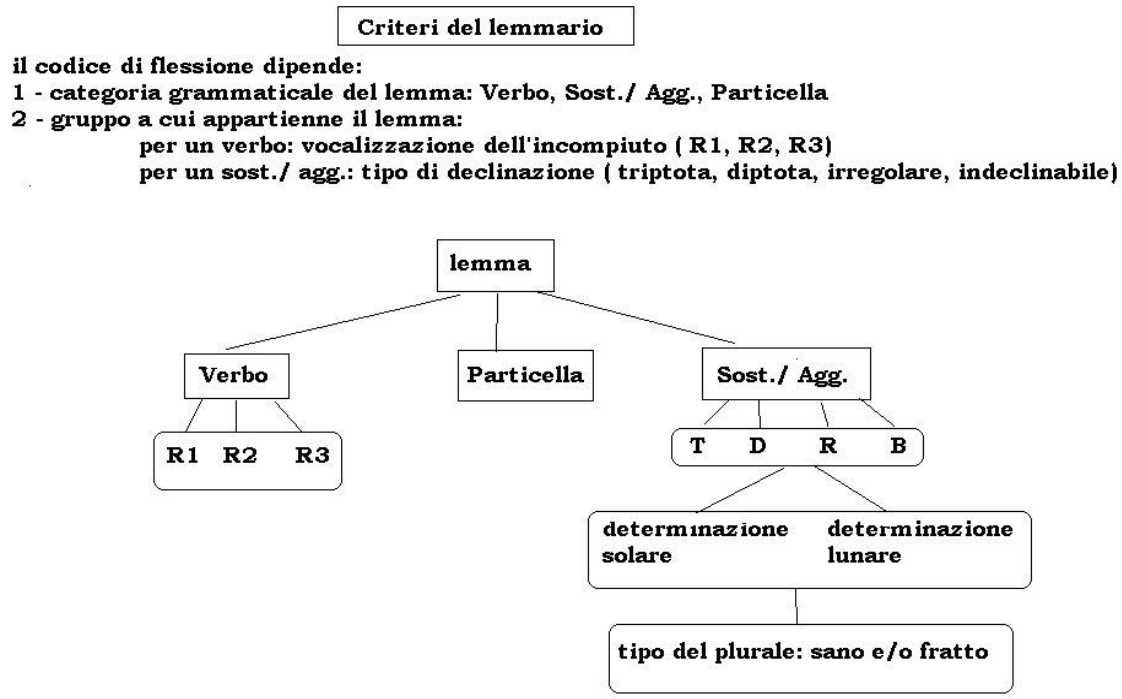

Figura 4: definizione criteri di inserimento di un lemma nel lemmario

Il motore morfologico

Descrizione dei principali moduli software

Le componenti software sono al momento descritte, nelle loro funzionalità, solo sommariamente in quanto prima della loro implementazione definitiva è necessario un periodo di test e verifica.
Uno schema generale della struttura e delle interazioni delle diverse componenti può comunque essere così riassunto:

- Programma per la creazione del Lemmario. Il programma gestisce il meccanismo di inserimento, nell'archivio "Lemmario", di un lemma comprensivo di vocali e di tutte le informazioni per il trattamento informatico dello stesso: categoria grammaticale, codice di flessione, ecc..
- Componenti per la gestione dell'interfaccia utente (inserimento dei lemmi e restituzione dei risultati), sia per quanto riguarda la parte di generazione che per la fase di analisi. Tale interfaccia sarà integrata con le componenti che gestiscono i meccanismi del motore morfologico vero e proprio.

Fase di generazione

- Componenti per la segmentazione del lemma restituito dall'interfaccia utente e per la fase di interrogazione del lemmario con il relativo recupero delle informazioni su di esso.
- Funzioni che in base al riconoscimento del lemma vanno a recuperare le regole necessarie per la generazione delle forme, in modo che eseguendo su di un lemma tutte e sole le operazioni indicate dalla regola si ottengano tutte e sole le forme flesse del lemma dato. Le regole al loro interno contengono anche degli operatori che servono a guidare la generazione.

- Componenti grafici che realizzano a video, in una forma tabellare, il prodotto della flessione.

Fase di analisi

- Procedure di supporto per il riconoscimento dei suffissi e prefissi che costituiscono la fase di "analisi delle particelle". Queste procedure servono ad escludere ipotesi che vadano in conflitto con le caratteristiche delle particelle individuate.
- Moduli software per la segmentazione del lemma utilizzando le possibili desinenze e le eventuali regole per la ricerca di possibili candidati da trovare nel lemmario
- Componenti che nel caso il sistema si trovi davanti ad una radice irregolare evitino comportamenti anomali del sistema stesso e uniformino il trattamento dei dati. Anche nel caso di lemmi non verbali, grazie alla "analisi delle particelle" è possibile ottimizzare il comportamento del sistema.

Modulo per la creazione del lemmario

La gestione di testi in alfabeto non latino presenta delle caratteristiche che complicano la fase di immissione, di gestione ed interrogazione, con l'ulteriore aggravante, per la lingua araba, della lettura del testo da destra verso sinistra. In particolare è necessario porre l'attenzione a:

come interpretare il testo in input;
come visualizzarlo a video;
come gestirlo anche per un suo corretto ordinamento alfabetico.

Caratteristiche del programma

Il programma gestisce la creazione di un lemmario della lingua araba e consente, attraverso una interfaccia grafica, l'immissione dei lemmi arabi verbali o non verbali da parte dell'utente utilizzando le seguenti funzionalità:

1. Inserimento di un nuovo lemma.
   L'utente deve inserire tutti i caratteri che formano la parola (consonanti e vocali), in modo che per ognuno sia possibile determinarne la posizione per eventuali operazioni sulla radice.
   L'interfaccia consente di inserire sia il lemma che la radice (scheletro consonantico); la sua categoria grammaticale; il gruppo di appartenenza e il codice di flessione che permette il trattamento corretto del lemma. Il lemma è inserito secondo l'ordine alfabetico che è utilizzato da DBT (codifica ASCII).

2. Veduta di tutti i lemmi, verbali o non verbali, che precedono o seguono il lemma digitato
   L'elenco dei lemmi a video è ordinata alfabeticamente e non per radice.

3. Per ognuno dei lemmi dell'elenco è possibile modificare sia la radice sia la categoria grammaticale che il gruppo di appartenenza e salvare le modifiche sul file lemmario. Le modifiche possono riguardare anche uno solo dei campi di immissione.

4. Ogni lemma inserito può essere eliminato dal lemmario con una opportuna funzione prevista dal programma ripristinandone l'ordine alfabetico.

5. Il programma è in grado di offrire all'utente la possibilità di visualizzare i lemmi inseriti ordinandoli per categoria grammaticale o per codice di flessione e all'occorrenza di stamparli.

6. Il programma al fine di permettere la visualizzazione di tutto il lemmario è dotato di una procedura di conversione che permette di tradurre i lemmi dalla codifica ASCII nella norma ISO 8859-6 e da all'utente la possibilità di stampare tutto il lemmario o parti di esso.

La tabella di riferimento

Per completezza di informazione alleghiamo di seguito la tabella di riferimento ricavata dalla norma ISO 8859-6 per i caratteri arabi. Il programma si basa su tale norma per il trattamento a video delle lettere arabe.

Charset ISO 8859-6 (Arabic)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0			32	0 48 ٠	@ 64	P 80	‘ 96	p 112			160			ذ 208	- 224	ـً 240
1			! 33	1 49 ١	A 65	Q 81	a 97	q 113					ء 193	ر 209	ف 225	ً 241
2			" 34	2 50 ٢	B 66	R 82	b 98	r 114					آ 194	ز 210	ق 226	ٌ 242
3			# 35	3 51 ٣	C 67	S 83	c 99	s 115					أ 195	س 211	ك 227	
4			$ 36	4 52 ٤	D 68	T 84	d 100	t 116			$ 164		ؤ 196	ش 212	ل 228	
5			% 37	5 53 ٥	E 69	U 85	e 101	u 117					إ 197	ص 213	م 229	
6			& 38	6 54 ٦	F 70	V 86	f 102	v 118					ئ 198	ض 214	ن 230	
7			’ 39	7 55 ٧	G 71	W 87	g 103	w 119					ا 199	ط 215	ه 231	
8			(40	8 56 ٨	H 72	X 88	h 104	x 120					ب 200	ظ 216	و 232	
9			) 41	9 57 ٩	I 73	Y 89	i 105	y 121					ة 201	ع 217	ى 233	
A			* 42	: 58	J 74	Z 90	j 106	z 122					ت 202	غ 218	ي 234	
B			+ 43	; 59	K 75	[91	k 107	{ 123				؛ 187	ث 203		ً 235	
C			, 44	< 60	L 76	\ 92	l 108	| 124			، 172		ج 204		ٌ 236	
D			- 45	= 61	M 77	] 93	m 109	} 125					ح 205		ٍ 237	
E			. 46	> 62	N 78	^ 94	n 110	~ 126					خ 206		َ 238	
F			/ 47	? 63	O 79	_ 95	o 111					؟ 191	د 207		ُ 239	

Figura 5: la tabella dei caratteri arabi utilizzata

La tastiera

Essendo l'immissione dei caratteri gestita dal programma è necessario adottare una convenzione sui caratteri da digitare sulla tastiera:

Convenzione da qui adottata è quella di indicare i lemmi e le forme della lingua araba con la stessa codifica utilizzata per la tastiera.

q = ق	w = و	E = ا E = ء	R= ع r = ر	T = ط t = ت	Y = ى y = ي	U = ؤ u = ے	I = ئ i = ِ	O= ۓ o = ہ
A = ا a = ۓ	S= ث s = س	D = ظ D = د	F = ِ f = ف	g = ج	H = ح h = ھ	J = ع j = ذ	K = خ k = ك	l = ل
Z = ظ z = ر	x = ش	C = ۓ C = ص	V = ا v = إ	B = ۓ b = ب	n = ن	M = ہ m = م		

Interfaccia utente

Questa che vediamo è la schermata che l'utente ha davanti a se quando ha inserito un lemma e scorre il lemmario nelle posizioni immediatamente successive o precedenti.

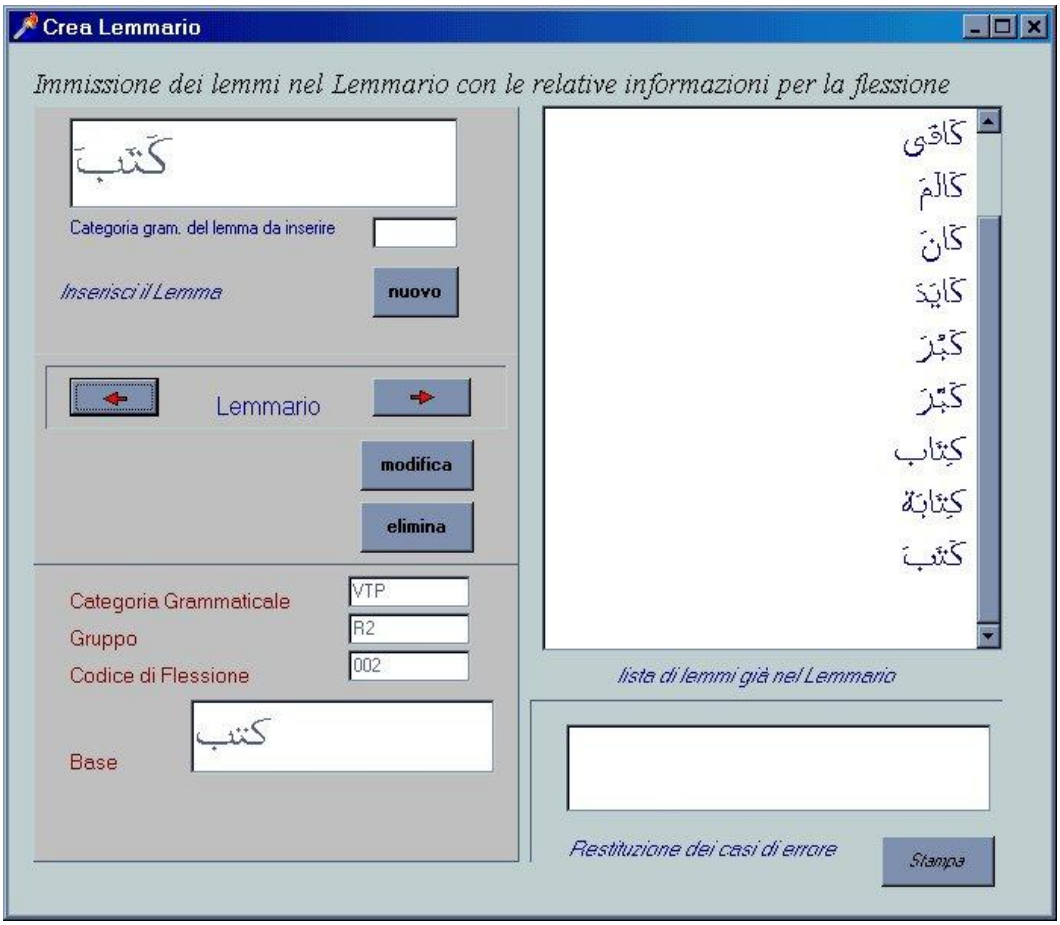

Figura 6: interfaccia d'uso del programma

Come è visibile in figura1, nell'interfaccia rivolta all'utente sono presenti più caselle di testo : Casella di testo 1 (con etichetta "inserisci il lemma"). In questa casella deve essere immesso il lemma da inserire nel lemmario, seguendo le specifiche suddette sui caratteri da digitare[1].

[1] In presenza del carattere LAM – ALIF poiché rispetto alle convenzioni assunte con la tastiera sopra indicata si ottiene digitando le coppie di caratteri '*lA*', '*lE*', '*lV*', '*lv*', se esistono vocali all'interno di queste coppie il programma non riconosce più la forma composta ma considera i caratteri singolarmente. La presenza di questa coppia è comunque salvata nella base e consentirà il suo reinserimento nel lemma nel caso di restituzione a video senza vocali.

Casella di testo 2 (con etichetta "categoria grammaticale")

All'interno della casella vanno inserite le lettere (secondo l'alfabeto latino) che identificano il lemma verbale o non verbale come da convenzioni stabilite al paragrafo 'file lemmario: convenzioni adottate'.

Casella di testo 3 (con etichetta "gruppo")

Qui vanno inserite le informazioni sulle caratteristiche del lemma sia verbale che non verbale sempre secondo le specifiche dello stesso paragrafo.

Casella di testo 4 (con etichetta "Codice di flessione")

Questa casella di testo prevede la conoscenza delle regole di flessione di ogni lemma e l'inserimento del numero corrispondente.

Casella di testo 5 (con etichetta "Base")

In questa casella deve essere inserita:
la radice del lemma (solo scheletro consonantico) che servirà sia in fase di generazione delle forme che in quella di analisi;
la terza persona maschile singolare del compiuto del verbo trilittero, nel caso di un verbo derivato (questa variazione si è resa necessaria per poter risalire al giusto verbo trilittero cui si riferisce il derivato, la presenza dei soli tre radicali non sarebbe stata sufficiente)

Esistono poi un'area di testo dove appare a richiesta l'elenco dei lemmi presenti nel lemmario nell'ordine alfabetico previsto dal lemmario stesso.
Una seconda area di testo dove il programma comunica il buon fine di un'operazione chiesta dall'utente o i problemi riscontrati nell'esecuzione della medesima.

<u>Modulo per la flessione dei lemmi</u>

Il software riguardante la flessione della lingua Araba è in continua revisione a causa del progressivo incremento del numero di lemmi inseriti nel lemmario e del conseguente adeguamento e ampliamento di regole e tabelle.
Il meccanismo per l'inserimento dei caratteri è quello utilizzato nel modulo software che crea il Lemmario.
L'algoritmo utilizzato dal programma si può così sintetizzare:

1. Una prima fase gestisce l'immissione dati e successivamente si occupa di verificare se il lemma inserito è già presente nel lemmario, dandone informazione all'utente;
2. La successiva in base alla regola flessionale associata al lemma digitato, recupera i prefissi (se necessari) ed i suffissi dalle tabelle opportune, seguendo i passi contenuti nella regola.
3. Una volta ottenuti i dati il programma crea una lista di forme che sono poi trattate, con procedure appropriate, per la restituzione a video. Sono inoltre restituite anche le informazioni che associate al lemma nel Lemmario.

<u>Caratteristiche del programma</u>

Esempio: il lemma ottenuto digitando i caratteri '*mlV*' è considerato diverso dal lemma ottenuto digitando i caratteri '*malaVa*'.
Il carattere '*B*' corrispondente tanuin <an> è sempre inserito al termine di un lemma e prende come sostegno '*A*' a parte alcune eccezioni: Il lemma che termina digitando i caratteri '*BA*' è considerato diverso dal sistema, da quello che finisce con '*AB*'.

Il programma prevede attualmente l'inserimento di un lemma comprensivo di vocali, la forma è quella che si trova come voce all'interno del lemmario (ad es: per flettere un verbo è necessario inserire la terza persona maschile del compiuto).
In questo momento sono pronte le funzionalità base che riguardano:

- ✓ I verbi forti (di tipo R1, R2, R3);
- ✓ I verbi geminati (tutti e tre i tipi);
- ✓ I verbi hamzati: di prima radicale hamzata (di tipo R1, R2, R3); di media radicale hamzata (R1, R2, R3); di ultima radicale hamzata (R1, R2, R3);
- ✓ I verbi deboli: di prima 'ya' e di prima 'waw' (R1, R2, R3); di media 'ya' e di media 'waw' (anch'essi R1, R2, R3); di ultima radicale debole 'ya' e 'waw' (R1, R2, R3);
- ✓ I verbi doppiamente irregolari in tutte le combinazioni di prima, media, ultima hamza e prima, media, ultima debole;
- ✓ I verbi derivati forti, i verbi derivati geminati, quelli hamzati e quelli deboli nelle loro rispettive dieci forme.

Per quanto riguarda i lemmi nominali[2] sono stati affrontati:

- ✓ I sostantivi verbali maschili;
- ✓ I sostantivi verbali femminili;
- ✓ I sostantivi verbali mobili.

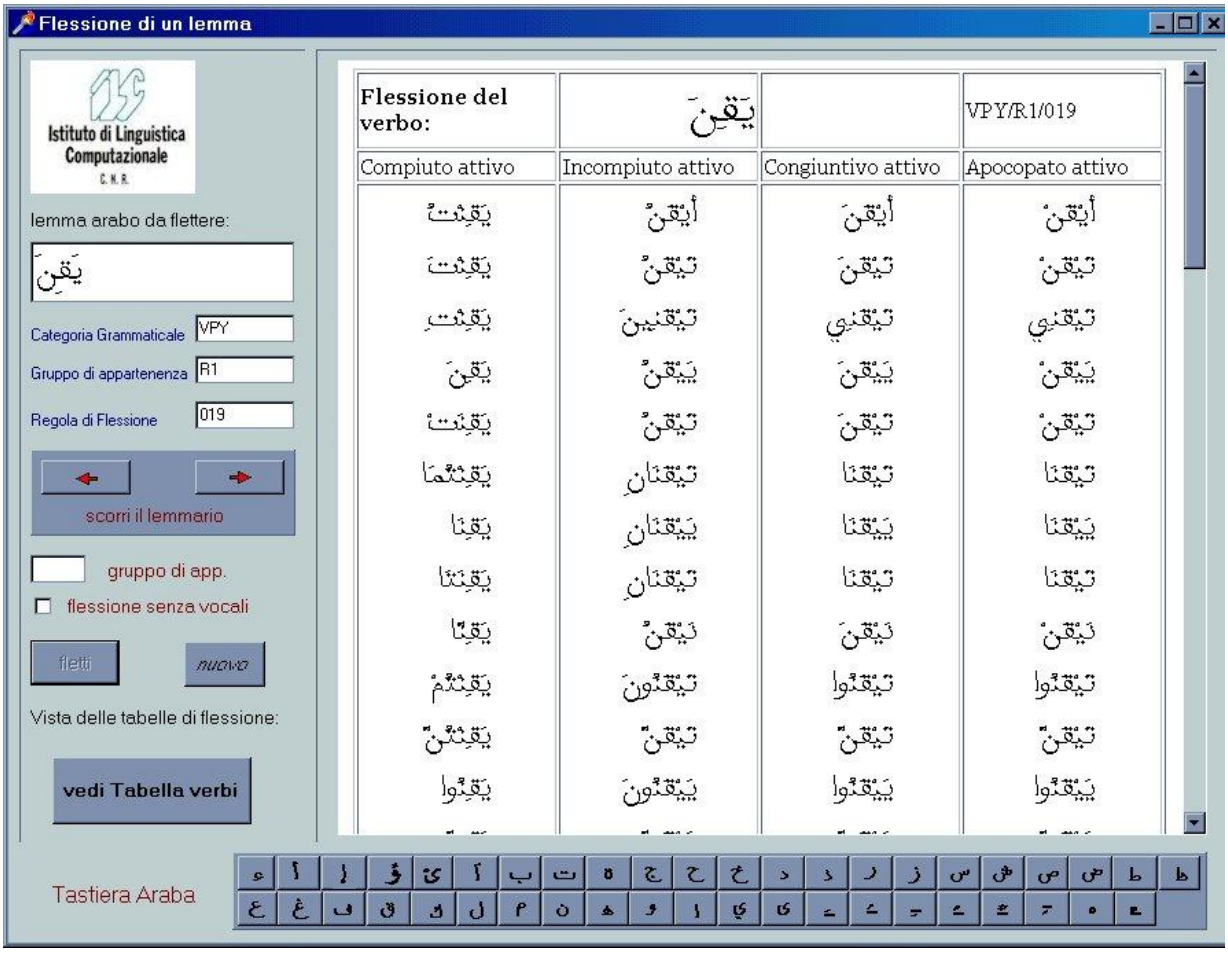

Figura 7: la flessione di un verbo debole

[2] I nomi hanno una flessione che comprende i diversi casi (nominativo, accusativo, obliquo), la eventuale determinazione e il numero (singolare, duale, plurale e dove esiste il plurale fratto).

Per i lemmi verbali è possibile ottenere la flessione senza vocali attivando l'opzione prevista nell'interfaccia, in ogni caso la flessione è legata al pulsante "fletti".

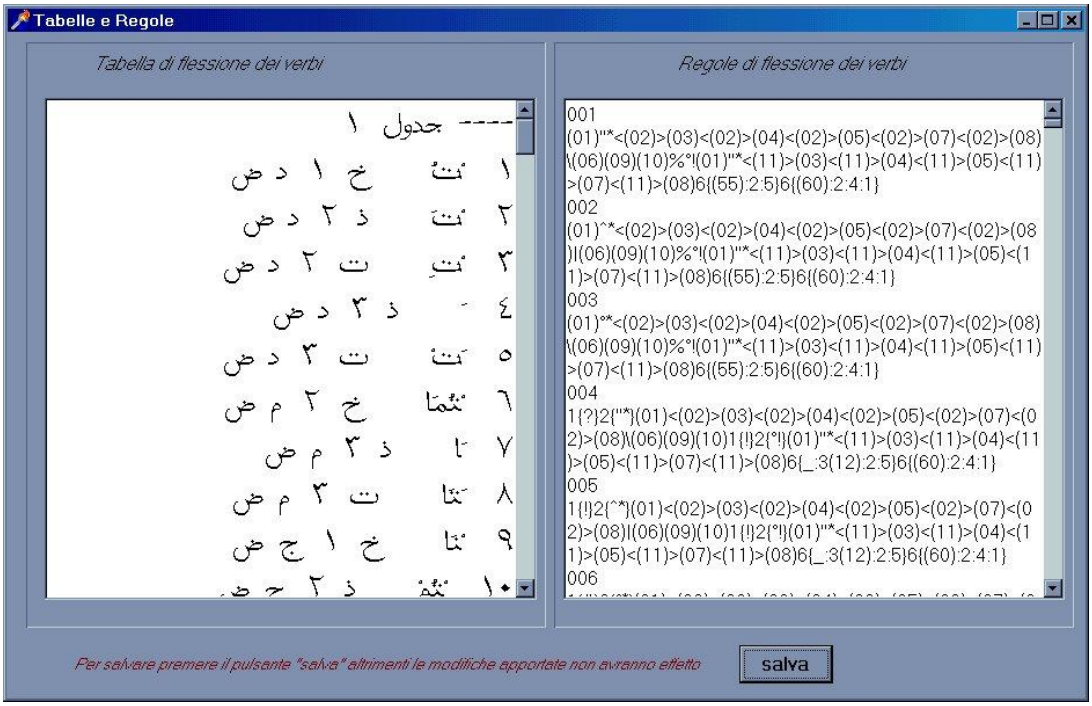

Figura 8: schermata delle tabelle delle desinenze e delle regole

Un secondo pulsante "vedi Tabella" consente di vedere le tabelle dei prefissi e dei suffissi che permettono la flessione e le regole che guidano il programma. Per i lemmi non verbali l'opzione non è operativa.
E' possibile modificare sia le regole che le tabelle ma, mentre per le prime è sufficiente agire come in un qualsiasi file di testo, per le tabelle contenenti le desinenze occorre evidenziare la sequenza da correggere poi premere invio. Apparirà allora una nuova finestra dove andrà inserita la nuova sequenza e quindi salvata nel file delle tabelle.

Modulo analizzatore

Il software riguardante l'analizzatore morfologico è pronto nella sua funzionalità di: riconoscimento dell'appartenenza delle forme a particolari lemmi anche nel caso in cui date forme contengano prefissi, suffissi, particelle o pronomi. E' previsto che il programma sia in grado di analizzare forme indipendentemente dalla presenza di vocali.
Il meccanismo del programma prevede il upload di un testo arabo opportunamente codificato, secondo le specifiche DBT per la codifica dei testi, all'interno di un'interfaccia che visualizzi i caratteri arabi (vengono utilizzate in parte le procedure già in uso per il modulo di flessione).
Una volta visualizzato il testo è possibile scorrere sequenzialmente le parole e richiederne l'analisi.
L'algoritmo utilizzato dal programma si può così sintetizzare:

1. Una prima fase gestisce la messa a video del testo arabo che si intende analizzare;
2. La successiva segmenta la forma flessa in modo da ipotizzare una possibile base e produce la flessione del lemma per la necessaria verifica del corretto riconoscimento.

3. Una volta ottenuta conferma dell'avvenuto riconoscimento, il programma crea un elenco di possibili lemmi cui la forma può appartenere. Le ipotesi restituite sono tanto maggiori nel caso di assenza delle vocali chiave.

Figura 9: l'analisi di un testo prova

<u>Definizione del protocollo di codifica delle regole flessionali</u>

Per il corretto utilizzo del sistema occorre predisporre:

TABELLE dove siano contenuti tutti i prefissi, i suffissi, le desinenze che occorrono per la flessione di un qualsiasi lemma, verbale o non verbale. I caratteri da inserire nella tabella devono essere comprensivi di vocali e di tutti i segni fonetici presenti e scritti in arabo.

REGOLE serie di passi da effettuare sulla radice per ottenere tutta la coniugazione di un verbo, se si tratta di un lemma verbale, oppure la flessione di un nome nel caso di lemma non verbale.

Come abbiamo detto le regole di flessione servono alla generazione di tutte le forme di un lemma e sono studiate in modo che tutti i lemmi che sviluppano la flessione delle loro forme secondo un identico meccanismo procedurale possano avere lo stesso codice di flessione (codice che identifica la regola di flessione da applicare). Nella prospettiva dell'uso di regole di flessione associate a ciascun lemma si è cercato di eliminare il concetto di irregolarità che è invece presente nelle regole

di flessione: non solamente ai comportamenti morfologici di generazione regolari è stato associato un proprio codice di flessione ma anche a quelli irregolari riconducendo l'intero sistema morfologico ad una gestione omogenea.
Le regole contengono operatori convenzionali che guidano la generazione delle forme, qui di seguito si da l'elenco dei maggiori operatori utilizzati. Le regole sono organizzate in un database sempre a disposizione dei diversi moduli software.

Operatori semplici per la costruzione delle regole

Del : modifica la base ottenuta fino a quel momento togliendo l'ultimo carattere;
DelX : " " " " " togliendo il carattere nella posizione X (= un qualsiasi numero)
Del0W : modifica la base ottenuta togliendo il carattere alfabetico W (= un qualsiasi carattere) all'inizio della base
DelGrZ(c1…cn)X : modifica la base ottenuta togliendo il gruppo c1..cn di Z caratteri nella posizione X a partire dal primo carattere della base.
TbX : Utilizzando la tabella di desinenze identificata da X, viene generata una nuova forma per ogni desinenza presente nella tabella. La forma è ottenuta giustapponendo la desinenza e la classificazione grammaticale associata alla base ottenuta
AddW : modifica la base ottenuta fino a quel punto aggiungendo il carattere alfabetico W in fondo alla base
AddGrZ(c1…cn): modifica la base ottenuta aggiungendo il gruppo c1..cn di Z caratteri in fondo alla stessa
PrefW : modifica la base ottenuta fino a quel punto aggiungendo il carattere alfabetico W all'inizio della base
AddXW : modifica la base ottenuta aggiungendo il carattere alfabetico W nella posizione X calcolata a partire dall'inizio della base
AddGrZ(c1…cn)X : modifica la base ottenuta aggiungendo il gruppo c1..cn di Z caratteri nella posizione X a partire dal primo carattere della base.
SubstXW : modifica la base ottenuta fino a quel momento aggiungendo il carattere alfabetico W nella posizione X, calcolata a partire dall'inizio della forma
SubstGr(c1....cn)(d1....dn)X : modifica la base sostituendo il gruppo (c1...cn) con il gruppo (d1....dn) nella posizione X, calcolata a partire dall'inizio della base
ChnZ(c1…cn)X
: sostituisce il carattere alla posizione X con il gruppo di Z caratteri (c1…cn)
ChnZ(c1…cn)X(g1…gn)
: sostituisce il gruppo di (g1…gn) alla posizione X con il gruppo di Z caratteri (c1…cn)

Operatori composti per la costruzione delle regole

Regola = «sostituisci carattere1»: operatore composto che modifica la radice del lemma verbale a seconda che appartenga al sottogruppo R1, R2, R3

if G=1 Subst4*a* if G=2 Subst4*u* if G=3 Subst4*i* End

Regola = «sostituisci carattere2»: operatore composto che modifica la radice del lemma verbale a seconda che appartenga al sottogruppo R1, R2, R3

if G=1 Subst2*a* if G=2 Subst2*u* if G=3 Subst2*i* End

Regola = «metti prefisso»: operatore composto che modifica la radice del lemma verbale

if G=2 Pref *Vu* if G=1 or G=3 Pref *vi* End

Codifica compatta degli operatori semplici e composti

Il protocollo studiato per la redazione delle regole flessionali, prevede anche un formato compatto degli operatori sopra descritti utilizzabile dai diversi moduli software, principalmente per la fase di flessione. Le operazioni sulla radice necessarie alla corretta flessione che sono comuni a più lemmi e che quindi ricorrono più frequentemente, sono stati compressi in operatori costituiti da un solo carattere. Questo per ottimizzare le funzioni di accesso alla risorsa e permettere, in caso, una successiva composizione in formule più complesse[3].

'-' : codifica l'operatore Del; cancella

'^' : codifica l'operatore:
Subst4u; Sostituisci '*u*' alla quarta posizione
Se è seguito da':': inserisce il carattere '*u*' nella posizione indicata dal numero che segue ':'

'*' : codifica l'operatore
per un lemma verbale: Subst2o; sostituisci '*o*' alla quarta posizione
per un lemma non verbale: cerca se c'è chadda e raddoppia il carattere che precede

'"' : codifica l'operatore
Subst4a; Sostituisci '*a*' alla quarta posizione
Se è seguito da':': inserisce il carattere '*a*' nella posizione indicata dal numero che segue ':'

'°' : codifica l'operatore
Subst4i; Sostituisci '*i*' alla quarta posizione
Se è seguito da':': inserisce il carattere 'i' nella posizione indicata dal numero che segue ':'

'#' : codifica l'operatore
per un lemma verbale: Subst2i; Sostituisci '*i*' alla seconda posizione
per un lemma non verbale: aggiunge il plurale determinato e indeterminato

'!' : codifica l'operatore Subst2u; Sostituisci '*u*' alla seconda posizione

'?' : codifica l'operatore Subst2a; Sostituisci '*a*' alla seconda posizione

'$' : codifica l'operatore
per un lemma verbale: DelX cancella un carattere dalla posizione indicata dal numero che segue ':'
es.: $:2: cancella il carattere in posizione 2
$:2:3 cancella tre caratteri in posizione 2

@ : codifica l'operatore
per un lemma non verbale: permette di controllare se il nome inizia con un prefisso (t o m o V o v)

'%' :codifica l'operatore

[3] es.: 5{-:2:3+[650104]:7}
la regola 5 identifica un tipo di regola composta che sta ad indicare:
togliere 3 caratteri alla posizione 2, aggiungere, in posizione 7, il gruppo di caratteri dati dalla riga 1 alla riga 4 della tabella 65

per un lemma verbale: Azzera le basi verbali per la formazione della voce passiva;
per un lemma non verbale: indica che ha un femminile dopo il maschile

'|' : codifica l'operatore
per un lemma verbale: AddGr(*Vu*)1; Inserimento del prefisso '*Vu*'
per un lemma non verbale: controlla, dopo la formazione del plurale fratto, se questo ultimo inizia con '*v*', la cambia in '*V*'

'\' : codifica l'operatore
per un lemma verbale: AddGr(*vi*)1; Inserimento del prefisso '*vi*'
per un lemma non verbale: Del(*aM*); toglie il gruppo (*aM*) se è alla fine del lemma

'+' :codifica l'operatore
per un lemma verbale: AddGr(*Va*)1; Inserimento del prefisso '*Va*'

'£' : codifica l'operatore
per un lemma verbale: SubstGr(*Vu*)1; sostituisci il gruppo di due caratteri alla prima posizione con '*Vu*'

'{' : indica l'inizio di un operatore composto;

'}' : indica la fine di un operatore composto;

'[' : indica che si deve prendere solo alcune righe di una tabella

Sono mostrati di seguito alcuni esempi di flessione a partire dall'interpretazione delle regole di flessione.

Coniugazione del verbo trilittero forte

Le regole sono associate all'utilizzo di tabelle in cui sono inseriti i prefissi ed i suffissi necessari alla flessione del lemma. E' mostrato di seguito un esempio di tabella dove accanto ad ogni desinenza è presente l'informazione sul tempo, il modo e la persona che si ottiene giustapponendo il suffisso alla radice. Ad esempio nella sequenza MFC1S:
MF sta per maschile/femminile ;
C sta per compiuto;
1 sta per prima persona;
S sta per singolare.

Tb1 Tabella suffissi Compiuto

001	*otu*	MFC1S
002	*ota*	MC2S
003	*oti*	FC2S
004	*a*	MC3S
005	*ato*	MC3S
006	*otumaA*	MFC2D
007	*aA*	MC3D
008	*ataA*	FC3D
009	*onaA*	MFC1P
010	*otum*	MC2P
011	*otunOa*	FC2P
012	*uWA*	MC3P

013 *na* FC3P

Vediamo adesso un esempio di regola per la flessione della voce attiva di un lemma verbale appartenente alla categoria dei verbi trilitteri forti (F).

Voce attiva

Formazione del compiuto: **Del,Tb1**

- estrazione dal Lemmario del lemma verbale
- eliminazione dell'ultimo carattere
- aggiunta dei suffissi contenuti nella tabella Tb1

Formazione dell'incompiuto indicativo**:**

Del, Subst2o, »sostituisci carattere1» , Tb2,Tb3

- estrazione dal Lemmario del lemma verbale
- sostituzione del carattere nella posizione 2 con "*o*"
- sostituzione del carattere in posizione 4 con la vocale indicata dal codice associato al lemma e contenuto nel Lemmario (vocale del tema)
- eliminazione del ultimo carattere
- aggiunta del prefisso (contenuto nella tabella Tb2) e del suffisso (nella tabella Tb3)

Formazione dell'incompiuto congiuntivo:

Del, Subst2o, «sostituisci carattere1», Tb2, Tb4

- estrazione dal Lemmario del lemma verbale
- sostituzione del carattere nella posizione 2 con "*o*"
- sostituzione del carattere in posizione 4 con la vocale indicata dal codice associato al lemma e contenuto nel Lemmario (vocale del tema)
- eliminazione del ultimo carattere
- aggiunta del prefisso (contenuto nella tabella Tb2) e del suffisso (nella tabella Tb4)

Formazione dell'incompiuto apocopato:

Del, Subst2o, «sostituisci carattere1»,Tb2, Tb5

- estrazione dal Lemmario del lemma verbale
- sostituzione del carattere nella posizione 2 con "*o*"
- sostituzione del carattere in posizione 4 con la vocale indicata dal codice associato al lemma e contenuto nel Lemmario (vocale del tema)
- eliminazione del ultimo carattere
- aggiunta del prefisso (contenuto nella tabella Tb2) e del suffisso (nella tabella Tb5)

Formazione dell'imperativo:

Del, Subst2o, «sostituisci carattere1», «metti prefisso», Tb6

- estrazione dal Lemmario del lemma verbale
- sostituzione del carattere nella posizione 2 con "*o*"
- sostituzione del carattere in posizione 4 con la vocale indicata dal codice associato al lemma e contenuto nel Lemmario (vocale del tema)
- analisi della vocale del tema e aggiunta davanti al lemma sin qui ottenuto di un nuovo prefisso
- eliminazione del ultimo carattere
- con questa nuova base aggiunta dei suffissi contenuti nella Tb6

Formazione dell'incompiuto energico 1:

Del, Subst2, "sostituisci carattere1",Tb2, Tb7

- estrazione dal lemmario del lemma verbale
- sostituzione del carattere nella posizione 2 con "*o*"
- sostituzione del carattere in posizione 4 con la vocale indicata dal codice associato al lemma e contenuto nel Lemmario (vocale del tema)
- eliminazione del ultimo carattere
- aggiunta del prefisso (contenuto nella tabella Tb2) e del suffisso (nella tabella Tb7)

Formazione dell'incompiuto energico 2:

Del, Subst2, "sostituisci carattere1",Tb2, Tb8

- estrazione dal lemmario del lemma verbale
- sostituzione con la vocale "*o*" nella posizione 2
- sostituzione del carattere in posizione 4 con la vocale indicata dal codice associato al lemma e contenuto nel Lemmario (vocale del tema)
- eliminazione del ultimo carattere
- aggiunta del prefisso (contenuto nella tabella Tb2) e del suffisso (nella tabella Tb8)

Formazione dell'imperativo energico 1:

Del, Subst2o, «sostituisci carattere1», »metti prefisso», Tb9

- estrazione dal Lemmario del lemma verbale
- sostituzione del carattere nella posizione 2 con "*o*"
- sostituzione del carattere in posizione 4 con la vocale indicata dal codice associato al lemma e contenuto nel Lemmario (vocale del tema)
- analisi della vocale del tema e aggiunta davanti al lemma sin qui ottenuto di un nuovo prefisso
- eliminazione del ultimo carattere
- con questa nuova base aggiunta dei suffissi contenuti nella Tb9

Formazione dell'imperativo energico 2:

Del, Subst2o, «sostituisci carattere1», »metti prefisso», Tb10

- estrazione dal Lemmario del lemma verbale
- sostituzione del carattere nella posizione 2 con "*o*"
- sostituzione del carattere in posizione 4 con la vocale indicata dal codice associato al lemma e contenuto nel Lemmario (vocale del tema)
- analisi della vocale del tema e aggiunta davanti al lemma sin qui ottenuto di un nuovo prefisso
- eliminazione del ultimo carattere
- con questa nuova base aggiunta dei suffissi contenuti nella Tb10

Esempio di flessione nominale

tabella 1: prefisso della determinazione

1a- Agg(*Val*)0, Agg(*O*)4 se il nome inizia con una lettera solare
Le lettere solari sono *t, S, d, j, r, z, s, x, c, D, T, Z, l, n*

1b- Agg(*Valo*) se il nome inizia con una lettera lunare
Le lettere lunari sono: *e, E, v, V, b, g, H, K, J, R, f, q, k, m, h, w, y*

1c- in due casi (solare e lunare) Agg(*Al*)0

tabella 2a: declinazione triptota d'un nome che termina con una lettera forte

Caso	Indeterminato	Determinato in stato costrutto	determinato in stato assoluto

Nominativo	Agg (*C*)	Agg(*u*)	Agg(*u*)
Accusativo	Agg(*AB*)	Agg(*a*)	Agg(*a*)
Obliquo	Agg*(F)*	Agg(*i*)	Agg(*i*)

tabella 2b: declinazione triptota d'un nome che termina con "*M"* o "*e"* o "*V"*

caso	Indeterminato	Determinato in stato costruto	Determinato in stato assoluto
nominativo	Agg (*C*)	Agg(*u*)	Agg(*u*)
accusativo	Agg(*B*)	Agg(*a*)	Agg(*a*)
obliquo	Agg*(F)*	Agg(*i*)	Agg(*i*)

tabella 4: suffissi del femminile del aggettivo
Agg(*aM*)

tabella 5: suffissi e declinazione del duale maschile:

caso	Indeterminato in stato assoluto	determinato in stato costrutto	determinato con l'articolo: (+Tb1)
nominativo	Agg(*aAni*):	Agg(*aA*):	Agg(*aAni*):
accusativo | obliquo	Agg(*ayoni*):	Agg(*ayo*):	Agg(*ayoni*):

Esempi di flessione dei sostantivi maschili verbali / nominali

Regola di flessione 300 -> SM con flessione triptota e determinazione lunare e ultimo radicale forte e plurale sano es.: *JaAlam*
la determinazione: Tb1b, il nome inizia con lettera lunare, quindi Add.(*Valo*)
la declinazione: il nome termina con una lettera forte Tb2
il duale: Tb5
il plurale sano: con declinazione Tb5

Regola di flessione 301-> SM con flessione triptota e determinazione solare e ultimo radicale forte e plurale sano
es: *taAbiJ*
la determinazione: Tb1a, il nome inizia con lettera solare, quindi Add(*Val*)0, Add.(*O*)4
la declinazione: il nome termina con una lettera normale Tb2
il duale: Tb5
il plurale sano: con declinazione Tb5

Regola di flessione 302 -> SM con flessione triptota e determinazione lunare e ultimo radicale forte e plurale fratto triptota (*VafoJaAl*):
es.: *qufol*, *Jinab*,
la determinazione: Tb1b, il nome inizia con lettera lunare, quindi Add(*Valo*)
la declinazione: il nome termina con una lettera forte Tb2a
il duale: Tb5
formazione del plurale fratto VafoJaAl triptota : Add(*o*)2, Add(*aA*)4, Add(*Va*)0
declinazione: Tb2a