

Building a Lexicon Database for Arabic Dialects

David Graff

Lead Programmer/Analyst, LDC

LDC Institute Presentation

Dec. 8, 2005


Overview

- Goals:
 - Desired properties of a lexicon and transcripts
- Common hurdles
- Hurdles particular to Arabic
- Some methods of approach:
 - Transcripts as plain-text or structured data files
 - Lexicon as plain-text or “tool-based” flat table
 - Lexicon as relational database (RDB)
- The next step -- now in progress:
 - Putting transcripts into the same RDB with the lexicon

What makes a good lexicon

- Full coverage for an adequately-sized corpus
 - List and describe all “word” tokens that are valid
 - Possibly add some common words not in the corpus
 - Token-count coverage of related corpora should be \geq 95%
- Consistent application of annotation conventions
 - Maximize the use of closed-set categorizations
 - Minimize the use of variant forms that “mean the same thing”
 - Centralize quality control activities among a small group of experts, working together closely
- Tight coupling between lexicon QC and transcripts
 - Error correction in the lexicon should propagate back to the transcripts

Common obstacles

- Axioms of manual transcription and annotation:
 - When a given task is done by N different people, there will be N different interpretations and techniques for the task.
 - The more often something must be done manually, the more mistakes will be made. (Typical minimum error rate is 5%)
 - For annotations involving unconstrained keyboard input, each person will create variant forms of a given annotation, and no two people will create the same form for it.
 - Most closed-set categorizations require a “miscellaneous” category. (Exception (?): “Is/Has X” vs. “Isn’t/Lacks X”)
 - For any particular annotation, a measurable percentage of instances within a corpus are indeterminate or ambiguous.
 - (“this guy” / ”the sky”) 

Common obstacles (cont.)

- Dependencies across layers of annotation
 - First layer: time segmentation of audio signal
 - Second layer: transcription of audio segments
 - Might catch segmentation errors
 - Third layer: building a lexicon from transcripts
 - Catches inconsistent spellings in transcription
 - Fourth layer: disfluency annotation
 - Catches segmentation and transcription errors
 - Fifth layer: treebanking
 - Potentially revises all previous layers
- **Worst case:** “Independent” layers create divergent versions of the same corpus.

Some Basic Problems with Arabic...

- The Arabic script-based writing system poses significant challenges for computational treatment.
 - Bidirectionality -- hard to render, harder to navigate and edit
 - Complex font with context-dependent rules for ligatures, glyph shape and character width
- Standard orthographic conventions represent an archaic form of the language.
 - Not “native” to any speaker of a current colloquial dialect
 - Colloquial (native) dialects have no standard orthography
- The absence of short vowels increases the difficulty.
 - Syntactic knowledge needed for correct word identification
 - Multiple meanings/pronunciations for a single written form

... and Special Problems with Colloquials

- Native speakers receive no formal instruction about their language -- no externalized grammar/analysis.
- Inherent phonological variability is unconstrained by - does not compete with - orthographic conventions or “correct speech”, so variants can have equal footing.
- Selection of consistent, appropriate spellings and morphological analyses for words entails deliberate and speculative choices.
- A systematic assessment of similarities and differences among colloquial dialects has yet to be done.

A Brief History of Implementation Details

Transcripts:

- Plain-text files and plain-text editors (bad old days)
 - Diverse information types are stored together “in-line”
 - Machine interpretation of content is difficult and brittle
 - Every corpus builder creates a new format
 - All data is manually editable -- nothing is safe
- Structured data files (XML) and specialized editors
 - Language content is always distinguishable from annotation
 - Many tools available for easy, reliable processing of data
 - Scope of format variation is constrained but not limiting
 - Scope of annotator effort is focused on appropriate tasks, and unrelated data is protected

Implementation Details (cont.)

Lexicons:

- Tab-delimited files and common (unix) shell tools
 - Tasks that are programmatic are fast, efficient, reliable; but:
 - Manual tasks are painful, and can break programmatic steps
 - Multi-stage manual work is especially risky, even with specialized tools for annotators/lexicographers
- Flat-table data and simple table-structured tools
 - Spreadsheets: easy to use and very capable, if entries are divided into reasonable sub-groups for handling/storage
 - Shoebox: has special attributes for linguists/lexicographers, but imposes its own set of limitations on what is possible
 - Data transfer across researchers/tools is simple and safe

Implementation Details: Lexicons (cont.)

- Relational Database (RDB)
 - Freely available servers are stable, easy to install, well documented, and can readily be made network-accessible.
 - Scalable to any size of lexical inventory with little or no effect on performance (speed)
 - Supports any appropriate conceptual model for lexicon creation, with configurable access permissions for users
 - Supports a wide range of “sanity constraints” on input data (uniqueness, data type, string length, numeric min/max, ...)
 - Provides lots of flexibility at the initial design stage and at any time thereafter (tables/fields can be added, modified)
 - Structured Query Language (SQL) provides a standardized, stable user interface for inserting, updating, retrieving data.

RDB Caveats

- Building a lexicon is always a complex process. RDB and SQL **do not** make it simpler (just more stable), and **do** require more technical expertise.
- Validating 50 K lexical entries is an inescapably long process. RDB will eliminate some delays and setbacks, but cannot reduce the basic effort required.
- Errors and failures are still possible -- on any scale.

Examples of Earlier Lexicons

- Callhome: tab-delimited, from transcripts & dictionaries

```
$aGGAlaB $@GG%l@//@$GG%lit 010 $aGGAlaB:noun+fem-sg//$aGGAlaB:adj+fem-sg
```

- Nahuatl (Jon Amith): based on Shoebox and fieldwork
 - Includes multiple dialects in a single DB, with indexed audio
 - Web enabled for maintenance, expansion and pedagogy
- CELEX: relational tables for lemmas vs. word forms, covering frequency, morphology, pronunciation, syntax
 - Distributed as a set of cross-referenced flat tables
 - Includes comprehensive documentation (~150 pages)

The Next Step (where we are now)

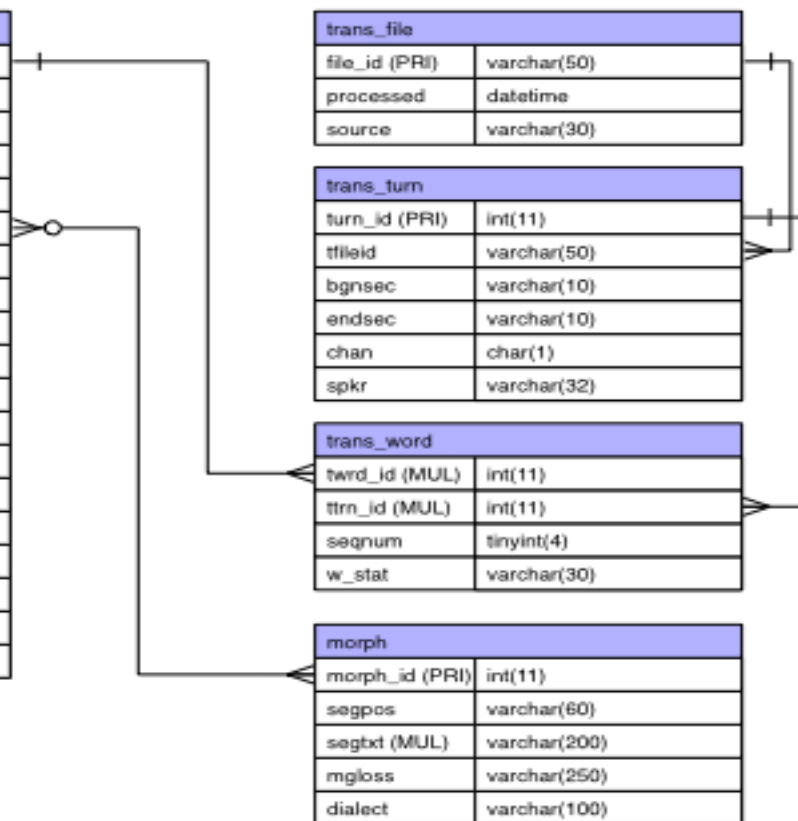
- Take one colloquial Arabic dialect at a time
- Create or acquire conversational transcripts
 - Time-stamped “turns” that index the associated audio
 - “Skeletal” orthography (no short vowels)
 - Can also include pronunciations (short vowels) as a separate layer, but this is not essential
- Load the transcripts into database tables
- Add morphology/POS/gloss annotations
- Review, revise and refine, then dump tables into the publishable lexicon and transcripts

Database Table Structure

lex_revision	
rwrld_id	int(11)
rgroup	varchar(250)
rgroupsiz	int(11)
rdate	datetime
rby	varchar(30)
rorth1	varchar(100)
rorth2	varchar(200)
rsegorth	varchar(250)
rseglbls	varchar(250)
rcanon	varchar(250)
rlgloss	varchar(250)
rdialect	varchar(100)
rword_stat	varchar(90)
rorth1_stat	varchar(90)
rorth2_stat	varchar(90)
rsegorth_stat	varchar(90)
rseglbls_stat	varchar(90)
rcanon_stat	varchar(90)
rlgloss_stat	varchar(90)
rdialect_stat	varchar(90)

(usage undefined)

lex	
word_id (PRI)	int(11)
orth1 (MUL)	varchar(100)
orth2	varchar(200)
segorth	varchar(250)
seglbls	varchar(250)
segmorph	varchar(200)
canon	varchar(250)
lgloss	varchar(250)
dialect	varchar(100)
rawfreq	int(11)
docfreq	int(11)
word_stat	varchar(90)
orth1_stat	varchar(90)
orth2_stat	varchar(90)
segorth_stat	varchar(90)
seglbls_stat	varchar(90)
canon_stat	varchar(90)
lgloss_stat	varchar(90)
dialect_stat	varchar(90)



Loading the tables

- For each transcript file:
 - Check for entry in trans_file, insert or update as needed
 - For each turn:
 - Check for entry in trans_turn, insert or update as needed
 - Delete entries (if any) from trans_word for this turn
 - For each word token:
 - Check for entry in lex, insert if needed
 - Add new entry to trans_word, citing turn-id, word-id, seq.number
 - Set “special feature” field in trans_word if token was uncertain “((this)) ((guy))“ or mispronounced “*nuclear”

Adding Morphology/POS/Gloss (MPG) Annotations

- Pull distinct words (skeletal “green” orthography) from lex table, sorted by frequency of occurrence in trans_word table (highest frequency first).
- Present one word at a time to an annotator, showing:
 - Skeletal (“green”) orthography
 - All associated vocalized (“yellow”) forms
 - Concordance drawn from token occurrences in turns
- Annotator provides:
 - “Canonical” vowelization
 - Segmentation into morphemes
 - Association of POS label to each morpheme
 - English gloss for each morpheme (and for word as a whole)

ABUMORPH Annotation Interface

created by Hubert Jin

AbuMorph.py

File Word

word done

\$nw Y

<A Y

Endy Y

hnA Y

lAzm Y

lk Y

zyn Y

GREEN: Endy

1 Einduy

1 Einduy

363 Einduy

POS Tags Gloss Comment Note

Eind/PREP+iy/PRON_1S at/with/near/by+me

Change Gloss Change Comment

Tag-1	Index	FullTag	Sel	Post	Word	Prev	POS Assigned	Yellow
ABBREV	1	ABBREV						
ACT_PART	2	ACT_PART						
ADJ	3	ADJ						
ADV	4	ADV						
CONJ	5	CONJ						
DEM_PRON_F	6	DEM_PRON_F			كل شي	عندي	ماكو اي	Eind/PREP+iy/PRON_1S Eindiy
DEM_PRON_FS	7	DEM_PRON_FS			كل شي	عندي	تفضلوا	Eind/PREP+iy/PRON_1S Eindiy
DEM_PRON_M	8	DEM_PRON_M			أشي من هاي النوع	عندي	لا لا أي	Eind/PREP+iy/PRON_1S Eindiy
DEM_PRON_MP	9	DEM_PRON_MP			كل شي بس عظام وشي ك	عندي	ما	Eind/PREP+iy/PRON_1S Eindiy
DEM_PRON_MS	10	DEM_PRON_MS			كل شي لا سلاح ولا ش	عندي	لا ما	Eind/PREP+iy/PRON_1S Eindiy
DET	11	DET			أشي بس أريد خلقتك	عندي	ما	Eind/PREP+iy/PRON_1S Eindiy
DISFL	12	DISFL			كل شي بس عندي معلوم	عندي	ما	Eind/PREP+iy/PRON_1S Eindiy
EXCEPT_PART	13	EXCEPT_PART			أشي إلك	عندي	ما	Eind/PREP+iy/PRON_1S Eindiy
FOCUS_PART	14	FOCUS_PART			أشي خطر	عندي	ما	Eind/PREP+iy/PRON_1S Eindiy
FUT	15	FUT			كل شي أي	عندي	نذك أي سلاح داخل	Eind/PREP+iy/PRON_1S Eindiy
INTERJ	16	INTERJ			كل شي غلط	عندي	دي كل شي أي غلط	Eind/PREP+iy/PRON_1S Eindiy
INTERROG_PART	17	INTERROG_PART			أشي أي (؟)	عندي	تراص علمود هالشي	Eind/PREP+iy/PRON_1S Eindiy
NEG_PART	18	NEG_PART			كل شي أي بس جاي أو	عندي	ما	Eind/PREP+iy/PRON_1S Eindiy
NOUN	19	NOUN			كل شي أنا ما سوي أي	عندي	ما (؟)	Eind/PREP+iy/PRON_1S Eindiy
NOUN_PROP	20	NOUN_PROP			أشي أنا في طريقي	عندي	أه أنا	Eind/PREP+iy/PRON_1S Eindiy
NUM	21	NUM			كل شي أي غلط	عندي	ما	Eind/PREP+iy/PRON_1S Eindiy
PART	22	PART			أشي أي كلش خوش و	عندي	ما	Eind/PREP+iy/PRON_1S Eindiy
PARTIALWORD	23	PARTIALWORD			كل شي انا عتاد ولا	عندي	بالسيارة ما عندك	Eind/PREP+iy/PRON_1S Eindiy
PREP	24	PREP						
PRON_1P	25	PRON_1P						
PRON_1S	26	PRON_1S						
PRON_2FS	27	PRON_2FS						
PRON_2MP	28	PRON_2MP						
PRON_2MS	29	PRON_2MS						
PRON_3FP	30	PRON_3FP						
PRON_3FS	31	PRON_3FS						
PRON_3MP	32	PRON_3MP						
PRON_3MS	33	PRON_3MS						
PRON_3P	34	PRON_3P						
PUNC	35	PUNC						
REL_ADV	36	REL_ADV						
REL_PRON	37	REL_PRON						
SUB_CONJ	38	SUB_CONJ						
VERB_PART	39	VERB_PART						
Z	40	Z						

Create the POS and Insert it to the list above

Pick All Drop All Stop Audio Remove Assignment Assign POS

Select Highlight Drop Highlight Play Audio

Vetting / Validating MPG Annotations

- Summary reports of morph and lex entries:
 - Alternate sortings by POS labels and orthography
 - Including frequency of occurrence
- Web-based query tool with login access limited to lexicographers:
 - Generic query-generator for finding items and sets in either the lex table or the morph table
 - Listing of lex or morph table entries with links to listings of element occurrences, and links to “Entry Editor” form
 - Entry Editor supports modification in place, creation of new entry based on modified current entry, and merging of current entry into some other entry

Main Lexicon Search Page

Arabic Lexicon Database: Main Search

Arabic Lexicon Database: Main S...

Arabic Lexicon Database: Main Search

- Use the controls in this table to build a query for word forms in the "lex" table, based on a single condition, or based on a conjunction of two distinct conditions (i.e. "A and B", or "A or B"). The "Search lex" button will fetch the results.

Parameter A:	<input checked="" type="radio"/> Green	<input type="radio"/> Yellow	<input type="radio"/> L.gloss	<input type="radio"/> Morphol.	<input type="radio"/> POS
Condition A:		<input checked="" type="radio"/> equals	<input type="radio"/> like	<input type="radio"/> contains	
Value A:	<input type="text" value="y\$wfn"/>				
Optional:		<input type="radio"/> and	<input type="radio"/> or		
Parameter B:	<input type="radio"/> Green	<input type="radio"/> Yellow	<input type="radio"/> L.gloss	<input type="radio"/> Morphol.	<input type="radio"/> POS
Condition B:	<input type="checkbox"/> not	<input type="radio"/> equals	<input type="radio"/> like	<input type="radio"/> contains	
Value B:	<input type="text"/>				<input type="button" value="Search lex"/>

[Click here to clear the form.](#)

- OR: Use the controls in the following table to build a query for morpheme forms in the "morph" table. The "Search morph" button will fetch the results.

Parameter A:	<input type="radio"/> segtxt	<input checked="" type="radio"/> segpos	<input type="radio"/> morph gloss
Condition A:	<input checked="" type="radio"/> equals	<input type="radio"/> like	<input type="radio"/> contains
Value A:	<input type="text" value=" V3MP"/>		
Optional:	<input type="radio"/> and	<input type="radio"/> or	<input type="checkbox"/> not
Parameter A:	<input type="radio"/> segtxt	<input type="radio"/> segpos	<input type="radio"/> morph gloss
Condition A:	<input type="radio"/> equals	<input type="radio"/> like	<input type="radio"/> contains
Value A:	<input type="text"/>		<input type="button" value="Search morph"/>

Note: Use an empty "Value" field in order to specify a "null" value for a given field. Use just a space to specify a non-null "empty string" value for the field.

- Search either the lex table or the morph table
- Use exact match, SQL "like" or regular-expression match
- Use single search criterion or two conjoined criteria::
 - "A and B"
 - "A and not B"
 - "A or B"
- Each "Search" button brings a separate pop-up window of "Search Results"
- Each pop-up is re-used on subsequent searches

Lexicon Search Results Page

Arabic Lexicon Search Results

Lex table search results for: orth1 = y\$wfwfn

Click on an ID to see turns containing that entry. Click on a stat value to edit the entry.

ID	Agreen	Bgreen	Ayellow	Byellow	N	POS_stat	W.Gloss	M.Gloss	Asegorth	Bsegorth	POS	Morph-IDs
4925	يشوفون	y\$wfwfn	يشوفون	y\$uwfuw	0							
5288	يشوفون	y\$wfwfn	يشوفون	y\$uwfuwn	0							
70454	يشوفون	y\$wfwfn	يشوفون	y\$uwfuwn	1	pass1	they see [masc.pl.]	they [masc.pl.] see, look at, check, show [masc.pl.]	ي شوفون	yi \$uwf uwn	IV3MP IV IVSUFF_SUBJ:P	1879 1503 1535
70455	يشوفون	y\$wfwfn	يشوفون	yi\$uwfuwn	8	pass1	they see [masc.pl.]	they [masc.pl.] see, look at, check, show [masc.pl.]	ي شوفون	yi \$uwf uwn	IV3MP IV IVSUFF_SUBJ:P	1879 1503 1535

[Click HERE](#) to mark all these entries as VALID

- “ID” links produce pop-up of transcript concordance page
- “POS_stat” links produce pop-up of lex-entry editor page
- “N” shows current frequency of word occurrence in transcripts

Transcript Concordance Display Page

Arabic Transcripts

Transcript turns containing word_id: 70455

Click on a turn_id to get the audio for that turn. Use the appropriate 'update' checkbox and paste in a different (**known**) word_id to change the word in a given turn. To change all words to a single different word_id, use the 'update all' checkbox and word_id paste-in at the bottom of the page.

Turn-ID	Following context	70455	Preceding context	Action	New WORD_ID
2471	هاي الزوارق	يشوفون	يروجون علمود	<input type="checkbox"/> to:	<input type="text" value="70455"/>
3139	لفتك هاي إذا يتكلم واياك واحد انجليزي بيرمج له بالإنجليزي هسا	يشوفون	بيرمجون به هاي اللغات يعني هسا	<input type="checkbox"/> to:	<input type="text" value="70455"/>
9573	أشكال ألوان	يشوفون	إي إي تدرين جماعتنا منا	<input type="checkbox"/> to:	<input type="text" value="70455"/>
10192	السيارة يؤيدوها	يشوفون	وتخمنها تجي لجنة من البنك تفحصها رأساً ثانية ويجون يفحصون السيارة بس	<input type="checkbox"/> to:	<input type="text" value="70455"/>
15842	أخويا	يشوفون	خطية بعد مو بعد ما خابرتي مو راح هو علمود	<input type="checkbox"/> to:	<input type="text" value="70455"/>
15855	سوريا شكرو ماكو انزين ها بعدين بعدين بعدين أقول لك	يشوفون		<input type="checkbox"/> to:	<input type="text" value="70455"/>
23789	طوله وعرضه يظلمهم بياعون عليه يقولون هذا عمره	يشوفون	إي لا مو تدرين من بجي	<input type="checkbox"/> to:	<input type="text" value="70455"/>
27458		يشوفون		<input type="checkbox"/> to:	<input type="text" value="70455"/>

- “Turn-ID” link fetches audio segment for the turn
- Transcription errors involving the target word can be corrected (so far, only word replacement is supported)
- Separate interface will be needed for word deletion/insertion

Lexicon Entry Editor Page

Arabic lex entry editor

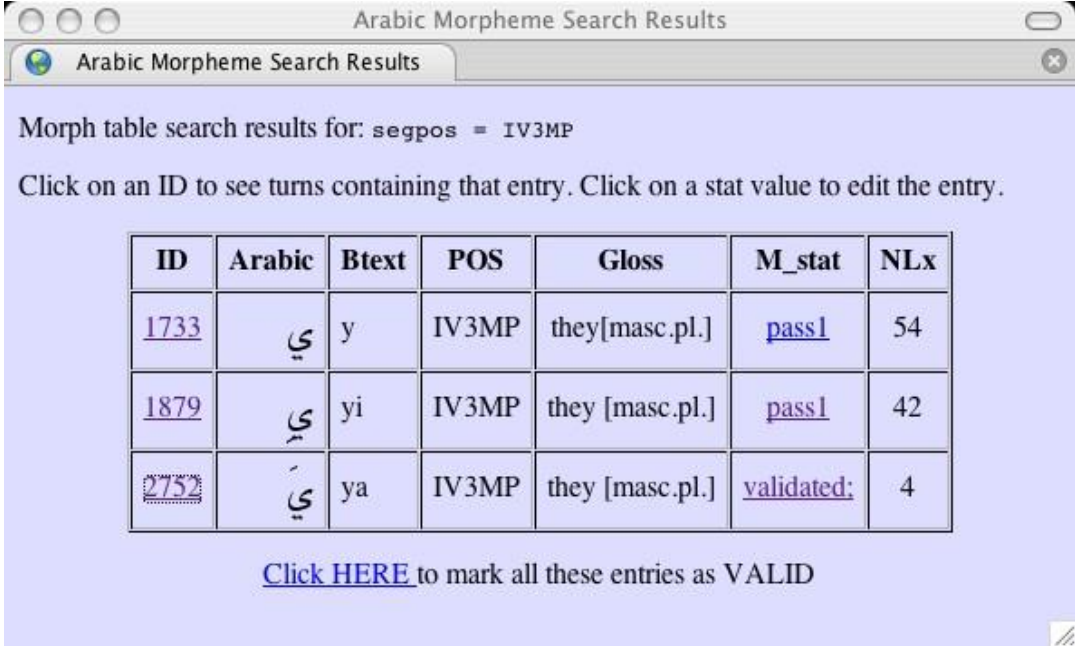
Arabic lex entry editor

Modify the lex table entry for word_id: 70454

Field	Arabic	Btext	Change	New Value
orth1	يشوفون	y\$wfwfn	<input type="checkbox"/> to:	y\$wfwfn
orth2	يشوفون	y\$uwfuwn	<input type="checkbox"/> to:	y\$uwfuwn
segtxt	ي + شُوف + وُن	yi + \$uwf + uwn	segpos	IV3MP + IV + IVSUFF_SUBJ:P
morph_id	1879 + 1503 + 1535		<input type="checkbox"/> to:	1879 + 1503 + 1535
lemma			<input type="checkbox"/> to:	
lgloss	they+see+[masc.pl.]		<input type="checkbox"/> to:	they+see+[masc.pl.]
word_id	70454	<input checked="" type="radio"/> merge into word_id:		70454
<input type="radio"/> Create new entry		<input checked="" type="radio"/> Update in place		<input type="button" value="Send to DB"/>

- Change Skeletal (“green”) or Pronunciation (“yellow”) spelling
- Change morphological composition and/or word gloss
- Update in place, or add as a new lex entry, or merge into some other existing lex entry (that is, render this entry obsolete)

Morph Search Results Page



Arabic Morpheme Search Results

Arabic Morpheme Search Results

Morph table search results for: `segpos = IV3MP`

Click on an ID to see turns containing that entry. Click on a stat value to edit the entry.

ID	Arabic	Btext	POS	Gloss	M_stat	NLx
1733	ي	y	IV3MP	they[masc.pl.]	pass1	54
1879	ي	yi	IV3MP	they [masc.pl.]	pass1	42
2752	ي	ya	IV3MP	they [masc.pl.]	validated:	4

[Click HERE](#) to mark all these entries as VALID

- “ID” links to Lex Search Results to show all lex entries containing this morph entry
- “M_stat” links to a pop-up morph entry editor page
- “NLx” = number of lex entries currently using this morpheme

Morpheme Entry Editor Page

Arabic morph entry editor

Arabic morph entry editor

Modify the morph table entry for morph_id: 1879

Field	Arabic	Btext	Change	New Value
segtxt	ي	yi	<input type="checkbox"/> to:	yi
segpos	IV3MP		<input type="checkbox"/> to:	IV3MP
mgloss	they [masc.pl.]		<input type="checkbox"/> to:	they [masc.pl.]
morph_id	1879		<input checked="" type="radio"/> merge into morph_id:	1879
<input checked="" type="radio"/> Create new entry		<input checked="" type="radio"/> Update in place		Send to DB

- Change the orthography, POS label and/or gloss
- Update in place, or create as a new entry, or merge all lex references to this entry so that they refer to some other morph entry instead (that is, render this entry obsolete)