

Morphophonemic and Orthographic Rules in a Multi-Dialectal Morphological Analyzer and Generator for Arabic Verbs

Nizar Habash & Owen Rambow

Columbia University

Center for Computational Learning Systems

Columbia Arabic Dialect Modeling Group (CADIM)

`{habash,rambow}@cs.columbia.edu`

Research Context

- Arabic-speaking world
 - Modern Standard Arabic = MSA (العربية الفصحى)
 - Focus of most natural language processing (NLP) research
 - Continuum of many spoken dialects
 - No standard orthography
 - Limited NLP tools
 - Ubiquitous switching between MSA and dialects
 - Need NLP tools that models both
- *Arabic variants* have similar, but different morphology
 - Want one morphological analyzer to handle all variants
 - Want to be able to leverage similarities among variants

MAGEAD

- Morphological Analysis and Generation for Arabic and its Dialects

- Deep representations

- lexeme and features

Aizdahar + per:3 gen:m
num:sg asp:perf

- root and pattern

[zhr,1tV2V3,iaa]+a

- Surface representations

- phonology

/izdahara/

- Orthography

Aizdahara (إِزْدَاهَر)

MAGEAD: Processing Steps

- Lexeme and features → abstract morphemes
 - Abstract morphemes are variant-independent
 - [FUT] = Future tense particle
- Order abstract morphemes
- Abstract morphemes → concrete morphemes

• [FUT] →

MSA	LEV	EGY	MOR
sa+ سـا	Ha++ حا	ha+ ها	ya+ يا

- Morphophonemic rules
- Orthographic rules

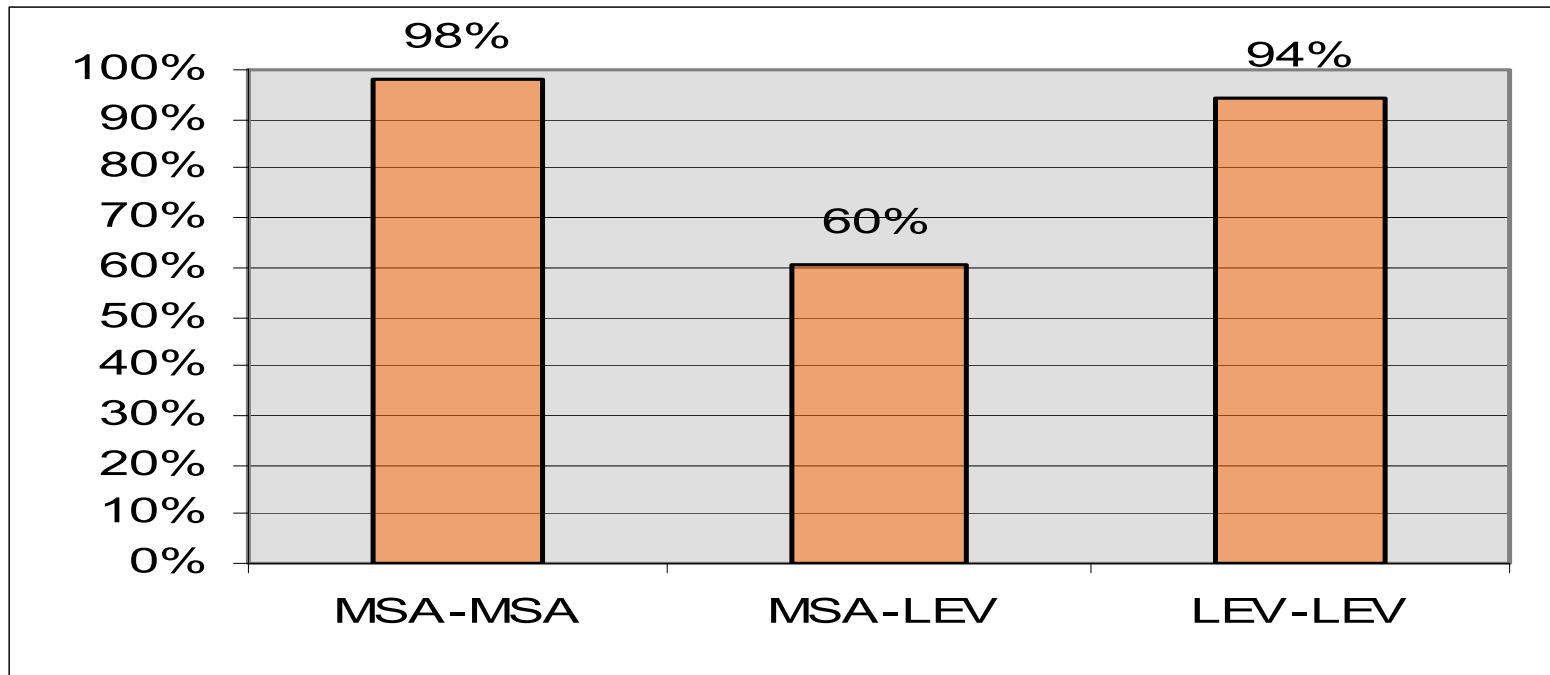
Morphological Rules

- Morphophonemic Rules
 - 69 MSA rules
 - $V1tV2V3, 1 \in \{/z/, /d/, /\delta/\} \rightarrow V1dV2V3$
 $/iztahara/ \rightarrow /izdahara/$
- Orthographic Rules
 - 53 MSA rules
 - $\emptyset \rightarrow A / \# _ \{i,u,a\}$
 $izdahara \rightarrow Aizdahara$

From MSA to Levantine

- **We use the Linguistic Data Consortium MSA-like orthography (Maamouri et al, 2006)**
- **New Abstract Morphemes**
 - The aspectual particle +ب *b+*
 - The postfix negation marker +ش *+š*.
- **Mapping Abstract to Concrete Morphemes**
 - Map to zero: dual number, and subjunctive and jussive moods
 - 2nd person masculine plural maps to +كم *+kum* and كوا *+kuwA*
 - Subject suffix changes: +ين *+iyana* in MSA → +ي *+iy* in Levantine
 - Future particle: +س *sa+* in MSA → +ح *Ha+* in Levantine
- **Rules: One rule change and one rule addition**
 - *madad+t* → مديت *maddayt* (not the MSA-like مددت *madadt*)
 - *b+A+aktub* → بكتب *baktub* not باكتب *bAaktub*

Levantine Evaluation



- Gold standard
 - MSA: Arabic Treebank (Maamouri et al. 2004)
 - LEV: Levantine Treebank (Maamouri et al. 2006)
- Context Token Recall = the percentage of words with one analysis matching the gold standard

Current State

- Implemented MSA verbal forms
- Implemented Levantine verbal forms
- Developed notion of lexeme compatible with root-pattern analysis
- Developed representation of morphological lexicons for dialects that can exploit commonalities
- Evaluated against test corpus

Future Work

- Extend system to handle nouns
- Extend system to other dialects
- Incorporate sound and pattern change rules
- Extend system to handle morphological code switching

شكراً

Thank you

Backup Slides

Contribution of this Paper

Description of *morphophonemic and orthographic rules* for morphological analysis and generation of Arabic and Levantine Verbs in MAGEAD

Previous Work

- Much existing work on Arabic morphology (Al-Sughaiyer & Al-Kharashi, 2004)
- Habash and Rambow, 2006 & Habash et al, 2005
 - MAGEAD: Morphological Analysis and Generation for Arabic and its Dialects
- Kiraz 2000
 - multi-tape automata
- Buckwalter 2002, 2004
 - concatenative morphology
- Beesley, Buckwalter & Newton 1989
 - two-level morphology with “detouring”
- Beesley & Karttunen 2000
 - compile-replace

MAGEAD: Levels of Morphological Representation

- Lexeme+Features Level

- Lexeme is an abstraction of all inflectional variants of a word

- Aizdahar₁ PER:3 GEN:m NUM:sg ASPECT:perf

- Morpheme Level

- [zhr, 1tV2V3,iaa] +a

- Surface Level

- Phonology: /izdahara/

- Orthography: Aizdahara (إِزْدَاهَرَ)

Arabic Morphology

- Types of morphemes
 - Templatic
 - Root: ktb (ك ت ب)
 - Pattern: 1V2V3
 - Vocalism: aa
 - Non-templatic word stem (NTWS)
 - نيويورك nyuwyrk
 - Affixes
 - +و wa+, +ي++uwna, +ها +hA
- Morphophonemic and orthographic rules

Morphophonemic Default Rules

- **Pattern tier has radical number**
 - Copy root tier letter to phonology tier
- **Pattern tier has ‘V’**
 - Copy vocalism tier letter to phonology tier
- **Pattern tier has anything else**
 - Copy pattern tier letter to phonology tier

V	1	t	Pattern
	z		Root
i			Vocalism
i	z	t	Phonology
			Orthography

Morphophonemic Lexical Rules

- **Verb Form VIII (افتعل):**
 - *iztahar+a → izdahar+a*
- **Weak C1:**
 - *y+awSil+u → y+aSil+u*
- **Weak C2 and C3:**
 - *daṣaw+a → daṣā*
 - *bayaṣ+a → bāṣa*
 - *quwil+a → qīla*
- **Geminates (C2=C3):**
 - *madad+a → madd+a,*
 - *madad+tu → madad+tu,*
 - *y+amdud+u → y+amudd+u*

Orthographic Default Rules

- **Copy phonology tier to orthography tier**
 - **Not all symbols on orthography tier are part of Arabic alphabet at this point**
 - Long vowels

v	l	t	Pattern
	z		Root
i			Vocalism
i	z	d	Phonology
i	z	d	Orthography

Orthographic Lexical Rules

- **Alif Maqsura Rules:**

- $ramay+a \rightarrow ram+a\acute{y}$
- $ramay+at \rightarrow ram+at$

- **First Person Singular Hamza Rule:**

- always realized as \acute{A}

- **Hamza of Form IV (أفعل) Rule:**

- always realized as \acute{A}

- **+wA Suffix Rules:**

- $qAl+\bar{u} \rightarrow qAl+uwA$
- $qAl+\bar{u}+hA \rightarrow qAl+uw+hA$

Orthographic Non-Lexical Rules

- **Clean up Rules:**
 - *mad0d+a* → *madda*
- **Long Vowel Spelling Rules:**
 - *ī* → *iy*, *ū* → *uw*
- **Shadda Rule:**
 - *madda* → *mad~a*.
- **Hamza Rules:**
 - *sa'ala* → *saĀala*
 - *su'ila* → *suŷila*
- **Vowel initial Spelling Rule:**
 - *izdahara* → *Aizdahara*

MAGEAD: Processing Steps

- Lexeme and features → abstract morphemes
 - Abstract morphemes are variant-independent

- Future tense particle:

MSA	LEV	EGY	MOR
sa+ +سـ	Ha+++حـ	ha+ +هـ	ya+ +غـ

- Lexicon defines lexeme as root + morphological behavior class (MBC)
- MBC maps features to abstract morphemes (patterns, vocalisms and affixes)
- Order abstract morphemes
- Abstract morphemes → concrete morphemes
- Morphophonemic rules
- Orthographic rules

Rules: Multitape Automata

- Following work of Kiraz (1996,2000)
- 5 tapes:

	v	1	t	v	2	v	3	+	a	Pattern
		z			h		r			Root
	i			a		a				Vocalism
	i	z	d	a	h	a	r	+	a	Phonology
A	i	z	d	a	h	a	r		a	Orthography

Rule Types

- **Morphophonemic Rules**
 - **Default rules** simply copy symbols to phonology tier
 - **Lexical rules** refer to specific morphemes
- **Orthographic Rules**
 - **Default rules** simply copy symbols to orthography tier
 - **Lexical rules** refer to specific morphemes
 - **Non-lexical rules** only refer to phonemes or graphemes (letters).

Rules: Example

Step 1: Enter data

Morphemes: zhr (ج ه ز) + V1tV2V3 + iaa + a

v	1	t	v	2	v	3	+	a
	z			h		r		
i			a		a			

Pattern

Root

Vocalism

Phonology

Orthography

Rules: Example

Step 4: Fill in orthography using orthographic default rules

v	1	t	v	2	v	3	+	a	Pattern
	z			h		r			Root
i			a		a				Vocalism
i	z	d	a	h	a	r	+	a	Phonology
i	z	d	a	h	a	r	+	a	Orthography

Rules: Example

Step 5: Apply orthographic lexical and non-lexical rules

	v	1	t	v	2	v	3	+	a	Pattern
		z			h		r			Root
	i			a		a				Vocalism
	i	z	d	a	h	a	r	+	a	Phonology
A	i	z	d	a	h	a	r		a	Orthography