



### Computational Morphology and Lexicography Modeling of Modern Standard Arabic Nominals

Christian Khairallah, Reham Marzouk, Salam Khalifa Mayar Nassar and <u>Nizar Habash</u>





- Motivations for Morphology Modeling
  - Additive value of morphological analyzers to neural models in several NLP tasks
    - Grammatical Error Correction (Alhafni et al., 2023)
    - Controlled Generation (Alhafni et al., 2022)
    - Morphological Disambiguation (Inoue et al., 2022)
    - Machine Translation (Oudah et al., 2019)
  - Help with low resource languages
  - Help with morphologically rich languages
  - Filter method, feature source, augmentation support, explainability





- Camel Morph Project (Habash et al., 2022)
  - Build open-source Arabic morphological models
  - Maximize coverage of Arabic linguistic phenomena
  - Include Standard, Classical and Dialectal Arabic
  - Cover many genres and domains
  - Use linguistically grounded representations
  - Built models are readily usable within an existing Python open-source suite for Arabic NLP, Camel Tools (Obeid et al., 2020)

#### جامعـة نيويورك أبوظـي NYU ABU DHABI



- Contributions in this paper ...
  - Defining the space of challenges in modeling Modern Standard Arabic (MSA) nominals
  - Developing an extendable large-scale implementation using Camel Morph
  - Benchmarking our models against publicly available analyzers
  - Our data and code are publicly available

# **Arabic Morphology Challenges**

#### • Morphological Richness

gender, number, person, aspect, mood, case, state and voice + many clitics

#### • Morphological Complexity

Allomorphs of many affixes and clitics

#### • Dialectal Variations

Many dialects with important differences

#### • Orthographic Ambiguity

- MSA: 12 readings/word due to optional diacritics
- Orthographic Inconsistency

#### • Morphological Challenges

- Rich & Complex Morphology
  - Templatic and Concatenative, many features & interactions
- Form-Function Mismatch
  - Broken plurals, irregular gender, case variants, syncretism
- Prefix-Stem-Suffix interactions
  - Prefix, Stem and Suffix Allomorphs

#### • Lexicographic Challenges

- Paradigm incompleteness
- Stem variants
- Inter-paradigm ambiguity

• Gender-Number-Case-State Discrepancies

	ni	nd	nc	gi	gd	gc	
ma	••	) C	~	Ç	Ģ	TMC	
ms	ũ	ι	l	ĩ		i	-1412
fa	ేం	รึ่ง	ó	ৄ	õ	ó	TE
15	aħũ	aħ	iu	aħĩ	al	<b>+r5</b>	
	ونَ	்	ۇر	ؠڹؘ	Ģ	ِي	
тр	uwi	na	uw	iyr	na	iy	TIVIP
fm	<u>َ</u> اتٌ	تُ	ló	<u>َ</u> اتٍ	تِ	ló	
тр	aAtũ	aA	tu	aAtĩ	aA	Ati	ŦĔ₽

Function → Form

- Syncretism: definite/indefinite/construct
- Default Function-Form mappings

Gender-Number-Case-State Discrepancies

					-		
	ni	nd	nc	gi	gd	gc	
ms	ំ ũ	ີ ເ	์ เ	្វ ĩ	Ç	> i	+MS
fs	َّة aħũ	ة at	í Iu	َةِ aħĩ	ة al	ó ħi	+FS
mp	ونَ uwi	் na	ُو uw	ینَ iyr	9 Na	ي iy	+MP
fp	َاتِّ aAtũ	تُ aA	ló Atu	َاتِ aAtĩ	تِ aA	ló Ati	+FP

Function  $\rightarrow$  Form

Lemma + Function → Stem + Form

			]	Featu	res	
Lemma	Gloss	Stem	ms	mp	fs	fp
muwaĎ~af	employee	muwaĎ~af	+MS	+MP	+FS	+FP
cofivr	ambassador	safiyr	+MS		+FS	+FP
safiyr	ambassador	sufaraA'		+MS		
nAr	fire	nAr			+MS	
IIAI	me	niyrAn				+MS
walivefat	aalinh	xaliyf	+FS			
xaliyiali	campii	xulafA'		+MS		

- Syncretism: definite/indefinite/construct
- Default and non-default Function-Form mappings
- Incomplete Paradigms

- Sound Plural vs Broken Plural
- Allomorphs
  - Stems
  - Buffers
  - Enclitics

			Features					
Lemma	Gloss	Stem	ms	mp	fs	fp		
safiyr	ambagadan	safiyr	+MS		+FS	+FP		
	ambassador	sufaraA'		+MS				

Word		(a) وَلِسَفِيرَاتِهِم walisafiyraAtihim `and for their ambassadors [f]'										
G (	Duo	alitias		Baseword								
Surface Sogmontation	Pro	cittics		Stem	Suffi	xes	Enclitic					
Segmentation	wa+	li+		safiyr	+aAt	+i	+him					

Word		(b) وَلِسُفَرَابِهِم walisufaraAŷihim `and for their ambassadors [m]'									
Surface	Pro	elitics		Baseword							
Surface	110	chues		Stem	Suffixes	Encitic					
Segmentation	wa+	li+		sufaraAŷ	+i	+him					

Sound Plural
 vs Broken Plural

			Features					
Lemma	Gloss	Stem	ms	mp	fs	fp		
<b>C</b>	ambagadar	safiyr	+MS		+FS	+FP		
sallyr	ambassador	sufaraA'		+MS				

Word		(a) وَلِسَفِيرَاتِهِم (walisafiyraAtihim `and for their ambassadors [f]'										
Correction of the second	Dwo	alitias		Baseword							En alidia	
Surface	Frochtics			Stem			Suffixes				Encirc	
Segmentation	wa+	li+			safiyr		+aAt	t	+i		+him	
Morpheme	prc2	prc2 prc1 prc0			root	pattern	gen	num	cas	stt	enc0	
& Features	wa+	li+	Ø	safiyr	s.f.r	1a2iy3	f	p	g	c	+hum	

Word		(b) وَلِسُفُرَائِهِم walisufaraAŷihim `and for their ambassadors [m]'										
	Due	alitian		Baseword							E	
Surface	Frochtics			Stem					Suffixes		Encitic	
Segmentation	wa+	li+				sufaraAŷ			+i		+him	
Morpheme	prc2	prc1	prc0	lex root pattern gen num cas					stt	enc0		
& Features	wa+	li+	Ø	safiyr	s.f.r	1u2a3aA'	m	p	g	c	+hum	

- Sound Plural vs Broken Plural
- Allomorphs
  - Stems
  - Buffers
  - Enclitics

sufaraA	'	а	
sufaraA	'	а	+hum
sufaraA	'	и	
sufaraA	Ŵ	и	+hum
sufaraA	'	i	
sufaraA	Ŷ	i	+him

			Features					
Lemma	Gloss	Stem	ms	mp	fs	fp		
safiyr	ambagadar	safiyr	+MS		+FS	+FP		
	ambassador	sufaraA'		+MS				

Word		(a)	وَلِسَفِيرَاتِ	walisafiyraAtihim `and for their ambassadors [f]'									
George	Dwo	alitias		Baseword							Englitig		
Surface	FTU	chucs		Stem		Suffixes				Encirc			
Segmentation	wa+	li+			safiyr		+aA	t	+i		+him		
Morpheme	prc2	prc1	prc0	lex	root	pattern	gen	num	cas	stt	enc0		
& Features	wa+	li+	Ø	safiyr	s.f.r	1a2iy3	f	p	g	c	+hum		
Buckwalter	DB	Prefix	D	BSter	m		D	BSuffix					
Database	w	ali+		safiyr +aAtihim					safiyr +aAtihim		+aAtihim		
Camel	[Conj]	[Prep]	[Art]	[Stem	]	[Buffer]	[Suff]				[Pron]		
Morph Specs	wa+	li+	Ø	safiyı	safiyr Ø			+aAt			+him		
Word		(b) (b)	وَلِسُفَرَائِ	walisufara	ıAŷih	im `and for	r their am	bassad	lors [m]'				
	Due	alitian		Baseword			seword				En alitia		
Surface Segmentation	Pro	cities				Stem			Suffixes		Encitic		
Segmentation	wa+	li+				sufaraAŷ			+i		+him		
Morpheme	prc2	prc1	prc0	lex	root	pattern	gen	num	cas	stt	enc0		
& Features	wa+	li+	Ø	safiyr	s.f.r	1u2a3aA'	m	p	g	c	+hum		
Buckwalter	DB	BPrefix				DBStem			DB	Suf	fix		
Database	w	ali+ sufaraAŷ +			ihin	1							
Camel	[Conj]	[Prep]	[Art]	[Stem	]	[Buffer]		[Suf	f]		[Pron]		
Mounh Succe		1:	a			A.	a				11.1		

- Sound Plural vs Broken Plural
- Allomorphs
  - Stems
  - Buffers
  - Enclitics
- Buckwalter
   DBPrefix
   DBStem
   DBSuffix
- Camel Morph Specifications

			Features						
Lemma	Gloss	Stem	ms	mp	fs	fp			
active	ambagadar	safiyr	+MS		+FS	+FP			
sallyr	ambassador	sufaraA'		+MS					

Word		(a) وَلِسَفِيرًا تَهم walisafiyraAtihim `and for their ambassadors [f]'											
Surface	Pro	clitics			Stem	Bas	seword	Suffi	xes (		Enclitic		
Segmentation	wa+	li+		safiyr			+aAt		+i		+him		
Morpheme	prc2	prc1	prc0	lex	root	pattern	gen	num	cas	stt	enc0		
& Features	wa+ li+ &		Ø	safiyr	s.f.r	1a2iy3	f	p	g	c	+hum		
Buckwalter	DB		D	BSter	n		D	BSuffix					
Database	W			safiyr			+;	aAtihim					
Camel	[Conj]	[Prep]	[Art]	[Stem	]	[Buffer]	[S		f]		[Pron]		
Morph Specs	wa+	li+	Ø	safiyr		Ø	+aAt	;	+i		+him		
Word		(b) (b)	وَلِسُفَرَائِ	walisufara	Aŷih	im `and for	r their aml	assaa	lors [m]'				
					_								
	Dwo	alitian				Bas	seword				Englitio		
Surface Segmentation	Pro	clitics				Bas Stem	seword		Suffixe	s	Enclitic		
Surface Segmentation	Pro wa+	clitics				Bas Stem sufaraAŷ	seword		Suffixe +i	s	Enclitic +him		
Surface Segmentation Morpheme	Pro wa+ prc2	litics li+ prc1	prc0	lex	root	Bas Stem sufaraAŷ pattern	seword gen	num	Suffixe +i cas	s stt	Enclitic +him enc0		
Surface Segmentation Morpheme & Features	Pro wa+ prc2 wa+	litics li+ prc1 li+	prc0 Ø	lex safiyr	root s.f.r	Bas Stem sufaraAŷ pattern 1u2a3aA'	eword gen m	num p	Suffixe +i cas g	s stt c	Enclitic +him enc0 +hum		
Surface Segmentation Morpheme & Features Buckwalter	Pro wa+ prc2 wa+ DB	clitics li+ prc1 li+ Prefix	prc0 Ø	lex safiyr	root s.f.r	Bas Stem sufaraAŷ pattern 1u2a3aA' DBStem	gen m	num p	Suffixe +i cas g DB	s stt c Suff	Enclitic +him enc0 +hum		
Surface Segmentation Morpheme & Features Buckwalter Database	Pro wa+ prc2 wa+ DB W	clitics li+ prc1 li+ Prefix ali+	prc0 Ø	lex safiyr	root s.f.r	Bas Stem sufaraAŷ pattern lu2a3aA' DBStem sufaraAŷ	gen m	num p	Suffixe +i cas g DB +i	s stt c Suff	Enclitic +him enc0 +hum fix		
Surface Segmentation Morpheme & Features Buckwalter Database Camel	Pro wa+ prc2 wa+ DB W (Conj)	litics li+ prc1 li+ Prefix ali+ [Prep]	prc0 Ø [Art]	lex safiyr [Stem	root s.f.r	Bas Stem sufaraAŷ pattern 1u2a3aA' DBStem sufaraAŷ [Buffer]	gen m	num p [Suf	Suffixe +i cas g DB +j	s stt c Suff	Enclitic +him enc0 +hum fix I [Pron]		

# **Camel Morph Approach**



- Camel Morph Specifications
- Camel Morph DB
- Camel Tools analysis and generation engines (Obeid et al., 2020)

# **Camel Morph Approach**



- In between two different approaches
  - Top down; linguistic representations; FSM; rules
  - Bottom up; lists of surface complex prefixes and suffixes, stems, and their compatibilities

					Morph Order												01
		DBPro	efix		DBS	em				DE	Suffix		saf	iyr+a/	At+i+h	im	02
	01	[Conj] []	Prep]		[NomStem]	Non	nBuff]			[Non	nSuff.IG]			sufa	araA+ý	γ+i+him	02
	O2	[Conj] []	Prep]		[NomStem]	Non	nBuff]			[NomSuff.	CG] [Pronoun]				Al+s	afiyr+aAt+i	03
	O3	[Conj] [Prep]	Determiner]		[NomStem]	Non	nBuff]			[NomSuff.DG]							
		Class	Lemma/ Morpheme	Form	Gloss	gen	num	stt	cas	Set Conds	Required Conds						
con	Lla	[NomStem]	safiyr	safiyr	ambassador	-	-	-	-		MS FS FP		1		1		
Lexi	L1b	[NomStem]	safiyr	sufaraA	ambassador	m	p	-	-	#A' #dip	MS	<ul><li>✓</li></ul>		<ul> <li>✓</li> </ul>			
rc	P1	[Determiner]															
<b></b>	P2	[Determiner]	Prc.Al	Al	the										$\checkmark$		
	B1	[NomBuff]									else						
fers	B2a	[NomBuff]		1							#A'	√	·				
Buf	B2b	[NomBuff]		ŷ							#A' obj suff-i			√			
	B2c	[NomBuff]		ŵ							#A' obj suff-u						
	S1a	[NomSuff.IG]	Suff.MSIG	ĩ		m	s	i	g	MS	else						
	S1b	[NomSuff.IG]	Suff.MSIG	a		m	s	i	g	MS	#dip	. ↓ √	·				
kes	S2	[NomSuff.IG]	Suff.FPIG	aAt+ĩ		f	p	i	g	FP							
lffi	S3	[NomSuff.CG]	Suff.MSCG	i		m	s	c	g	MS suff-i							
õ	S4	[NomSuff.CG]	Suff.FPCG	aAt+i		f	p	c	g	FP suff-i		ļ	√				
	S5	[NomSuff.DG]	Suff.MSDG	i		m	s	d	g	MS							
	S6	[NomSuff.DG]	Suff.FPDG	aAt+i		f	p	d	g	FP					√		
tics	C1	[Pronoun]					ļ		 					. <b> </b>			
nclit	C2a	[Pronoun]	Pron.3MP	hum	their					obj	else						
Ð	C2c	[Pronoun]	Pron.3MP	him	their					obj	suff-i		√	√			

• Morph order defines the full space of all morphemes that can co-occur by their class.

					sufa	raA+	'+a			01							
		DBPr	efix		DBS	tem				DB	Suffix		safi	yr+aA	t+i+h	im	02
	01	[Conj] []	Prep]		[NomStem]	[Non	nBuff]			[Non	nSuff.IG]			sufa	raA+ý	}+i+him	02
	02	[Conj] []	Prep]		[NomStem]	[Non	nBuff]			[NomSuff.				Al+s	afiyr+aAt+i	03	
	03	[Conj] [Prep]	Determiner]		[NomStem]	m] [NomBuff]				[NomSuff.DG]							
				1							1						
		Class	Lemma/ Morpheme	Form	Gloss	gen	num	stt	cas	Set Conds	Required Conds						
con	L1a	[NomStem]	safiyr	safiyr	ambassador	-	-	-	-		MS FS FP		1		1		
Lexi	L1b	[NomStem]	safiyr	sufaraA	ambassador	m	p	-	-	#A' #dip	MS	√		1			
2	P1	[Determiner]															
Ā	P2	[Determiner]	Prc.Al	Al	the								Ι	Ι	✓		
	B1	[NomBuff]									else						
fers	B2a	[NomBuff]		'			[				#A'	<ul> <li>✓</li> </ul>	Ι		[		
Buf	B2b	[NomBuff]		ŷ							#A' obj suff-i			1			
	B2c	[NomBuff]		ŵ							#A' obj suff-u						
	S1a	[NomSuff.IG]	Suff.MSIG	ĩ		m	s	i	g	MS	else						
	S1b	[NomSuff.IG]	Suff.MSIG	a		m	s	i	g	MS	#dip	√	ļ	ļ	ļ		
les	S2	[NomSuff.IG]	Suff.FPIG	aAt+ĩ		f	p	i	g	FP			ļ				
lfi	S3	[NomSuff.CG]	Suff.MSCG	i		m	s	c	g	MS suff-i			ļ	1	ļ		
s.	S4	[NomSuff.CG]	Suff.FPCG	aAt+i		f	p	c	g	FP suff-i			√	<b>_</b>	ļ		
	S5	[NomSuff.DG]	Suff.MSDG	i		m	s	d	g	MS			ļ	ļ	ļ		
	S6	[NomSuff.DG]	Suff.FPDG	aAt+i		f	p	d	g	FP					✓		
ics	C1	[Pronoun]					 	L					ļ	ļ	ļ		
nclit	C2a	[Pronoun]	Pron.3MP	hum	their					obj	else						
E	C2c	[Pronoun]	Pron.3MP	him	their					obj	suff-i		✓	$\checkmark$			

- Each form (allomorph) sets some truth conditions to be true.
- For a word to be valid, the required truth conditions of every form (allomorphs) in it must be already set by some other allomorph.

			Morph Order s															01
		DBPre	efix		DBS	tem				DE	Suffix		sat	iyr+	aA	t+i+h	im	02
	01	[Conj] []	Prep]		[NomStem]	[Non	nBuff]			[Non	nSuff.IG]			su	ıfar	∵aA+ŷ	+i+him	02
	O2	[Conj] []	Prep]		[NomStem]	[Non	nBuff]			[NomSuff.	CG] [Pronoun]				-	Al+s	afiyr+aAt+i	03
	O3	[Conj] [Prep]	Determiner]		[NomStem] [NomBuff]					[NomSuff.DG]								
		Class	Lemma/ Morpheme	Form	Gloss	gen	num	stt	cas	Set Conds	Required Conds							
con	Lla	[NomStem]	safiyr	safiyr	ambassador	-	-	-	-		MS FS FP		√			1		
Lexi	L1b	[NomStem]	safiyr	sufaraA	ambassador	m	p	-	-	#A' #dip	MS	<ul> <li>✓</li> </ul>			1			
rc	P1	[Determiner]																
P	P2	[Determiner]	Prc.Al	Al	the											√		
	B1	[NomBuff]									else							
ffers	B2a	[NomBuff]		'							#A'	1						
Buf	B2b	[NomBuff]		ŷ							#A' obj suff-i			,	1			
	B2c	[NomBuff]		ŵ							#A' obj suff-u							
	S1a	[NomSuff.IG]	Suff.MSIG	ĩ		m	s	i	g	MS	else							
	S1b	[NomSuff.IG]	Suff.MSIG	a		m	s	i	g	MS	#dip	√	·					
Kes	S2	[NomSuff.IG]	Suff.FPIG	aAt+ĩ		f	p	i	g	FP								
nffiy	S3	[NomSuff.CG]	Suff.MSCG	i		m	s	c	g	MS suff-i					1			
S	S4	[NomSuff.CG]	Suff.FPCG	aAt+i		f	p	c	g	FP suff-i			_ √					
	S5	[NomSuff.DG]	Suff.MSDG	i		m	s	d	g	MS								
	S6	[NomSuff.DG]	Suff.FPDG	aAt+i		f	p	d	g	FP						√		
tics	C1	[Pronoun]					 	ļ	 		 							
nclit	C2a	[Pronoun]	Pron.3MP	hum	their					obj	else							
E	C2c	[Pronoun]	Pron.3MP	him	their					obj	suff-i		_ √		1			

- Each form (allomorph) sets some truth conditions to be true.
- For a word to be valid, the required truth conditions of every form (allomorphs) in it must be already set by some other allomorph.

# Results

		Our S	specs
(a)	Lemmas (Stems)	27,023	(33,497)
	Noun	19,858	(25,293)
	Adjective	6,922	(7,921)
	Comparative Adjective	243	(283)
(b)	DBPrefix Morphemes (Allom.)	18	(20)
	DBSuffix Morphs (Allom.)	99	(197)
	Stem Buffers	22	
	Unique Condition Terms	51	
	Morph Order Lines	42	

- Intensive semi-automatic process for creating all the entries and quality checking them.
- Multiple annotators involved in Morphological and Lexicography design.

# Results

		Our S	Specs	Ou	r DB	Calim	a MSA		
(a)	Lemmas (Stems)	27,023	(33,497)	27,023	(37,910)	26,990	(38,323)	Lemmas (Stems)	(a)
	Noun	19,858	(25,293)	19,858	(28,302)	19,970	(29,370)	Noun	
	Adjective	6,922	(7,921)	6,922	(9,184)	6,808	(8,703)	Adjective	
	Comparative Adjective	243	(283)	243	(424)	212	(250)	Comparative Adjective	
(b)	DBPrefix Morphemes (Allom.)	18	(20)		213		77	DBPrefix Sequences	(c)
	DBSuffix Morphs (Allom.)	99	(197)		614		391	DBSuffix Sequences	
	Stem Buffers	22			3,442		1,423	Compatibility Tables	
	Unique Condition Terms	51		8	3,649,166	2	8,359,701	Unique Diacritized Forms	(d)
	Morph Order Lines	42		24	6,880,683	12	6,176,265	Unique Analyses	
					1,300,068		1,041,949	Unique Analyses (no Clitics)	

- We compare our compiled DB with Calima MSA (Taji et al. 2018), based on SAMA (Graff et al., 2009)/BAMA (Buckwalter, 2004)
- Lemmas: Comparable
- Stems: Our Specs < Our DB ≈ Calima MSA
- Forms + Analyses : Our DB >> Calima MSA
  - Consistent and extended modeling of features and affixes

# Results

		Our S	specs	Our	· DB	Calima	a MSA		
(a)	Lemmas (Stems)	27,023	(33,497)	27,023	(37,910)	26,990	(38,323)	Lemmas (Stems)	(a)
	Noun	19,858	(25,293)	19,858	(28,302)	19,970	(29,370)	Noun	
	Adjective	6,922	(7,921)	6,922	(9,184)	6,808	(8,703)	Adjective	
	Comparative Adjective	243	(283)	243	(424)	212	(250)	Comparative Adjective	
(b)	<b>DBPrefix Morphemes (Allom.)</b>	18	(20)		213		77	DBPrefix Sequences	(c)
	DBSuffix Morphs (Allom.)	99	(197)		614		391	DBSuffix Sequences	
	Stem Buffers	22			3,442		1,423	Compatibility Tables	
	<b>Unique Condition Terms</b>	51		83	3,649,166	28	8,359,701	Unique Diacritized Forms	(d)
	Morph Order Lines	42		240	6,880,683	120	6,176,265	Unique Analyses	
					1,300,068		1,041,949	Unique Analyses (no Clitics)	

- Coverage evaluation of Penn Arabic Treebank (Maamouri et al., 2004)
  - Recall 95.3% of analyses
  - 86% of mismatches due to gold errors.

# **Conclusions & Future Work**

- Presented a review of challenges in modeling MSA nominals
- Developed and benchmarked a large-scale implementation using Camel Morph
- All models and code are publicly available
- We plan to work on other POS and other Arabic dialects
- We want to tackle challenges such as noisy spelling, dialect-MSA intra-word code switching, template-based backoff modeling, and automatic learning of lexicon entries
- We plan to evaluate our models on downstream applications





# Thank you!

<u>morph.camel-lab.com</u> <u>nizar.habash@nyu.edu</u>