جامعة نيويورك أبوظبي

NYU | ABU DHABI

مختبر كامل

CAMeL Lab

Computational Morphology and Lexicography Modeling of Modern Standard Arabic Nominals

Christian Khairallah, Reham Marzouk, Salam Khalifa
Mayar Nassar and Nizar Habash

جامعة نيويورك أبوظبي

NYU ABU DHABI

مختبر كامل

CAMeL Lab

- **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

جامعة نيويورك أبوظبي

NYU ABU DHABI

مختبر كامل

CAMeL Lab

- **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)

جامعة نيويورك أبوظبي

NYU ABU DHABI

مختبر كامل

CAMeL Lab

- **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**

Arabic Nominal Modeling Challenges

- **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

Arabic Nominal Modeling Challenges

- Gender-Number-Case-State Discrepancies

Function → Form

	ni	nd	nc	gi	gd	gc	
ms	ٌ ũ	ُ u		ٍ ĩ	ِ i		+MS
fs	ةٌ aħũ	ةُ aħu		ةٍ aħĩ	ةِ aħi		+FS
mp	ُونَ uwna		ُو uw	ِينَ iyna		ِي iy	+MP
fp	اتٌ aAtũ	اتُ aAtu		اتٍ aAtĩ	اتِ aAti		+FP

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

Arabic Nominal Modeling Challenges

- Gender-Number-Case-State Discrepancies

Function → Form

	ni	nd	nc	gi	gd	gc	
ms	ٌ ũ	ُ u		ٍ ĩ	ِ i		+MS
fs	ةٌ aħũ	ةُ aħu		ةٍ aħĩ	ةِ aħi		+FS
mp	ونَ uwna	وُ uw		ينَ iyna	يِ iy		+MP
fp	اتٌ aAtũ	اتُ aAtu		اتٍ aAtĩ	اتِ aAti		+FP

Lemma + Function → Stem + Form

			Features			
Lemma	Gloss	Stem	ms	mp	fs	fp
muwaĎ~af	employee	*muwaĎ~af*	+MS	+MP	+FS	+FP
safiyr	ambassador	*safiyr*	+MS		+FS	+FP
		sufaraA'		+MS		
nAr	fire	*nAr*			+MS	
		niyrAn				+MS
xaliyfaħ	caliph	*xaliyf*	+FS			
		xulafA'		+MS		

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

Arabic Nominal Modeling Challenges

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

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

Word	(a) وَلِسَفِيرَاتِهِم walisafiyraAtihim *`and for their ambassadors [f]'*					
Surface Segmentation	**Proclitics**			**Baseword**		**Enclitic**
				Stem	**Suffixes**	
	wa+	li+		safiyr	+aAt	+i	+him

Word	(b) وَلِسُفَرَائِهِم walisufaraAŷihim *`and for their ambassadors [m]'*					
Surface Segmentation	**Proclitics**			**Baseword**		**Enclitic**
				Stem	**Suffixes**	
	wa+	li+		sufaraAŷ	+i	+him

Arabic Nominal Modeling Challenges

- Sound Plural
  vs Broken Plural

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

Word	(a) وَلِسَفِيرَاتِهِم walisafiyraAtihim *`and for their ambassadors [f]'*										
Surface Segmentation	**Proclitics**			**Baseword**							**Enclitic**
				Stem			**Suffixes**				
	wa+	li+		safiyr			+aAt		+i		+him
Morpheme & Features	**prc2**	**prc1**	**prc0**	**lex**	**root**	**pattern**	**gen**	**num**	**cas**	**stt**	**enc0**
	wa+	li+	∅	safiyr	s.f.r	1a2iy3	f	p	g	c	+hum

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

Arabic Nominal Modeling Challenges

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

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

sufaraA	*'*	*a*
sufaraA	*'*	*a +hum*
sufaraA	*'*	*u*
sufaraA	*ŵ*	*u +hum*
sufaraA	*'*	*i*
sufaraA	*ŷ*	*i +him*

Word	(a) وَلِسَفِيرَاتِهِم walisafiyraAtihim *`and for their ambassadors [f]'*										
Surface Segmentation	**Proclitics**			**Baseword**							**Enclitic**
				Stem			**Suffixes**				
	wa+	li+		safiyr			+aAt		+i		+him
Morpheme & Features	**prc2**	**prc1**	**prc0**	**lex**	**root**	**pattern**	**gen**	**num**	**cas**	**stt**	**enc0**
	wa+	li+	∅	safiyr	s.f.r	1a2iy3	f	p	g	c	+hum
Buckwalter Database	**DBPrefix**			**DBStem**			**DBSuffix**				
	wali+			safiyr			+aAtihim				
Camel Morph Specs	**[Conj]**	**[Prep]**	**[Art]**	**[Stem]**		**[Buffer]**	**[Suff]**				**[Pron]**
	wa+	li+	∅	safiyr		∅	+aAt		+i		+him

Word	(b) وَلِسُفَرَائِهِم walisufaraAŷihim *`and for their ambassadors [m]'*										
Surface Segmentation	**Proclitics**			**Baseword**							**Enclitic**
				Stem					**Suffixes**		
	wa+	li+		sufaraAŷ					+i		+him
Morpheme & Features	**prc2**	**prc1**	**prc0**	**lex**	**root**	**pattern**	**gen**	**num**	**cas**	**stt**	**enc0**
	wa+	li+	∅	safiyr	s.f.r	1u2a3aA'	m	p	g	c	+hum
Buckwalter Database	**DBPrefix**			**DBStem**					**DBSuffix**		
	wali+			sufaraAŷ					+ihim		
Camel Morph Specs	**[Conj]**	**[Prep]**	**[Art]**	**[Stem]**		**[Buffer]**	**[Suff]**				**[Pron]**
	wa	li	∅	sufaraA		ŷ	∅		+i		+him

Arabic Nominal Modeling Challenges

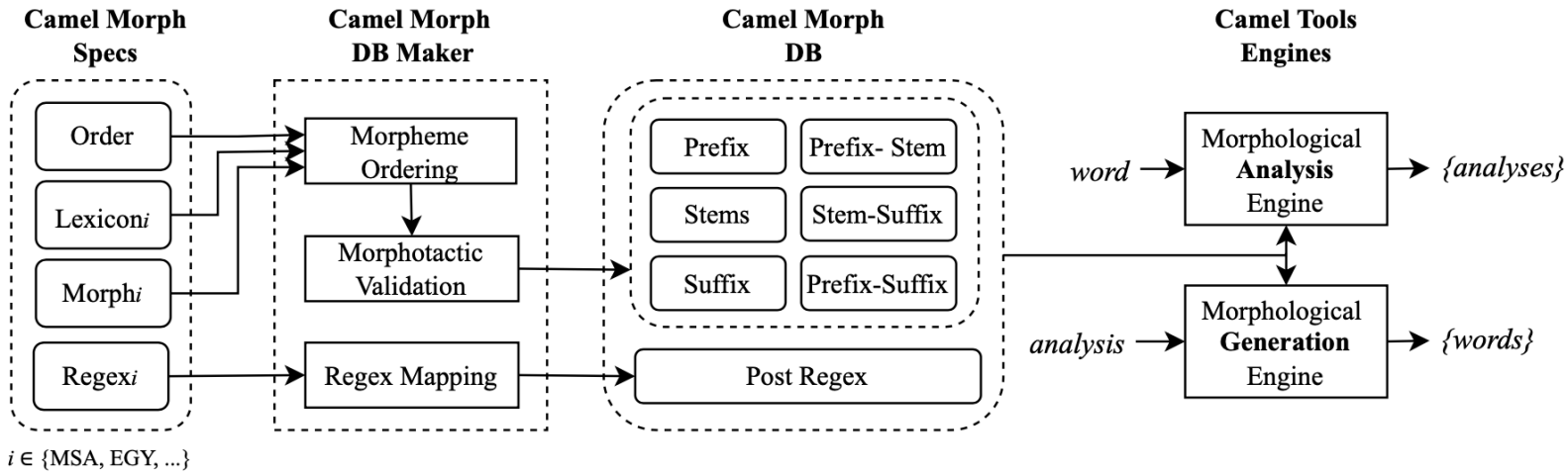
- Sound Plural vs Broken Plural
- Allomorphs
  - Stems
  - Buffers
  - Enclitics
- Buckwalter DBPrefix DBStem DBSuffix
- Camel Morph Specifications

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

Word	(a) وَلِسَفِيرَاتِهِم walisafiyraAtihim `*and for their ambassadors [f]*'										
Surface Segmentation	**Proclitics**			**Baseword**							**Enclitic**
				Stem			**Suffixes**				
	wa+	li+		safiyr			+aAt		+i		+him
Morpheme & Features	**prc2**	**prc1**	**prc0**	**lex**	**root**	**pattern**	**gen**	**num**	**cas**	**stt**	**enc0**
	wa+	li+	∅	safiyr	s.f.r	1a2iy3	f	p	g	c	+hum
Buckwalter Database	**DBPrefix**			**DBStem**			**DBSuffix**				
	wali+			safiyr			+aAtihim				
Camel Morph Specs	**[Conj]**	**[Prep]**	**[Art]**	**[Stem]**		**[Buffer]**	**[Suff]**				**[Pron]**
	wa+	li+	∅	safiyr		∅	+aAt		+i		+him

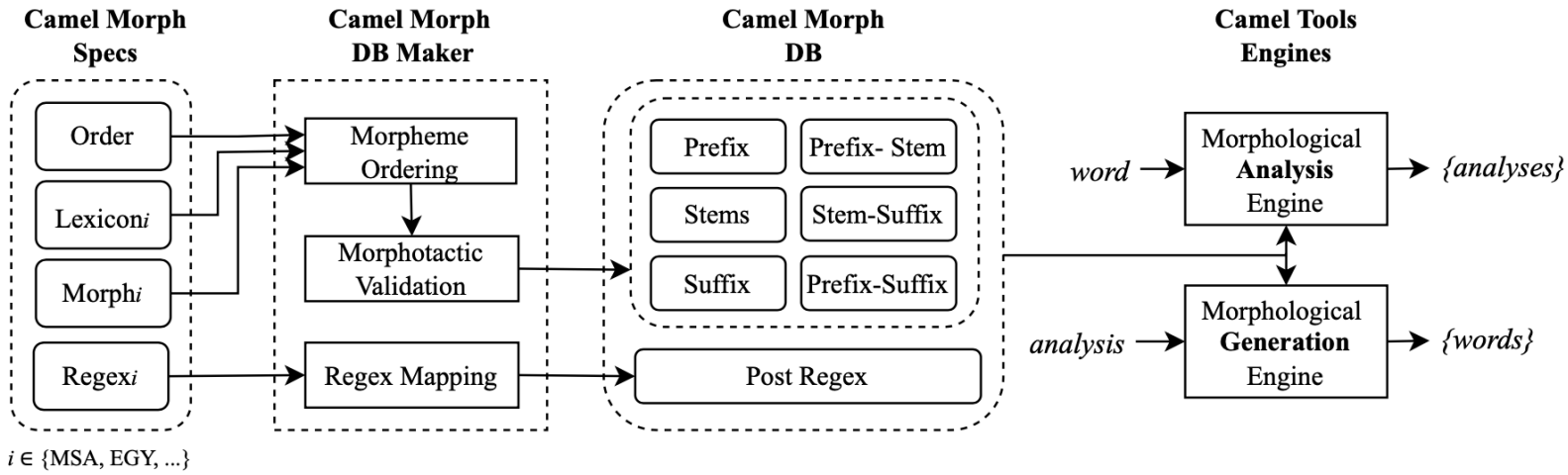
Word	(b) وَلِسُفَرَائِهِم walisufaraAŷihim `*and for their ambassadors [m]*'										
Surface Segmentation	**Proclitics**			**Baseword**							**Enclitic**
				Stem					**Suffixes**		
	wa+	li+		sufaraAŷ					+i		+him
Morpheme & Features	**prc2**	**prc1**	**prc0**	**lex**	**root**	**pattern**	**gen**	**num**	**cas**	**stt**	**enc0**
	wa+	li+	∅	safiyr	s.f.r	1u2a3aA'	m	p	g	c	+hum
Buckwalter Database	**DBPrefix**			**DBStem**					**DBSuffix**		
	wali+			sufaraAŷ					+ihim		
Camel Morph Specs	**[Conj]**	**[Prep]**	**[Art]**	**[Stem]**		**[Buffer]**	**[Suff]**				**[Pron]**
	wa	li	∅	sufaraA		ŷ	∅		+i		+him

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		
	DBPrefix	**DBStem**	**DBSuffix**
O1	[Conj] [Prep]	[NomStem] [NomBuff]	[NomSuff.IG]
O2	[Conj] [Prep]	[NomStem] [NomBuff]	[NomSuff.CG] [Pronoun]
O3	[Conj] [Prep] [Determiner]	[NomStem] [NomBuff]	[NomSuff.DG]

		Class	Lemma/ Morpheme	Form	Gloss	gen	num	stt	cas	Set Conds	Required Conds				
Lexicon	L1a	[NomStem]	*safiyr*	*safiyr*	ambassador	-	-	-	-		MS|FS|FP		✓		✓
	L1b	[NomStem]	*safiyr*	*sufaraA*	ambassador	m	p	-	-	#A' #dip	MS	✓		✓	
Prc	P1	[Determiner]													
	P2	[Determiner]	Prc.Al	Al	the										✓
Buffers	B1	[NomBuff]									else				
	B2a	[NomBuff]		'							#A'	✓			
	B2b	[NomBuff]		ŷ							#A' obj suff-i			✓	
	B2c	[NomBuff]		ŵ							#A' obj suff-u				
Suffixes	S1a	[NomSuff.IG]	Suff.MSIG	ĩ		m	s	i	g	MS	else				
	S1b	[NomSuff.IG]	Suff.MSIG	a		m	s	i	g	MS	#dip	✓			
	S2	[NomSuff.IG]	Suff.FPIG	aAt+ĩ		f	p	i	g	FP					
	S3	[NomSuff.CG]	Suff.MSCG	i		m	s	c	g	MS suff-i				✓	
	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					✓
Enclitics	C1	[Pronoun]													
	C2a	[Pronoun]	Pron.3MP	hum	their					obj	else				
	C2c	[Pronoun]	Pron.3MP	him	their					obj	suff-i		✓	✓	

sufaraA+'+a	O1
safiyr+aAt+i+him	O2
sufaraA+ŷ+i+him	O2
Al+safiyr+aAt+i	O3

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

	Morph Order		
	DBPrefix	**DBStem**	**DBSuffix**
O1	[Conj] [Prep]	[NomStem] [NomBuff]	[NomSuff.IG]
O2	[Conj] [Prep]	[NomStem] [NomBuff]	[NomSuff.CG] [Pronoun]
O3	[Conj] [Prep] [Determiner]	[NomStem] [NomBuff]	[NomSuff.DG]

		Class	Lemma/ Morpheme	Form	Gloss	gen	num	stt	cas	Set Conds	Required Conds	sufaraA+'+a O1	safiyr+aAt+i+him O2	sufaraA+ŷ+i+him O2	Al+safiyr+aAt+i O3
Lexicon	L1a	[NomStem]	*safiyr*	*safiyr*	ambassador	-	-	-	-		MS|FS|FP		✓		✓
	L1b	[NomStem]	*safiyr*	*sufaraA*	ambassador	m	p	-	-	#A' #dip	MS	✓		✓	
Prc	P1	[Determiner]													
	P2	[Determiner]	Prc.Al	Al	the										✓
Buffers	B1	[NomBuff]									else				
	B2a	[NomBuff]		'							#A'	✓			
	B2b	[NomBuff]		ŷ							#A' obj suff-i			✓	
	B2c	[NomBuff]		ŵ							#A' obj suff-u				
Suffixes	S1a	[NomSuff.IG]	Suff.MSIG	ĩ		m	s	i	g	MS	else				
	S1b	[NomSuff.IG]	Suff.MSIG	a		m	s	i	g	MS	#dip	✓			
	S2	[NomSuff.IG]	Suff.FPIG	aAt+ĩ		f	p	i	g	FP					
	S3	[NomSuff.CG]	Suff.MSCG	i		m	s	c	g	MS suff-i				✓	
	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					✓
Enclitics	C1	[Pronoun]													
	C2a	[Pronoun]	Pron.3MP	hum	their					obj	else				
	C2c	[Pronoun]	Pron.3MP	him	their					obj	suff-i		✓	✓	

- 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		
	DBPrefix	**DBStem**	**DBSuffix**
O1	[Conj] [Prep]	[NomStem] [NomBuff]	[NomSuff.IG]
O2	[Conj] [Prep]	[NomStem] [NomBuff]	[NomSuff.CG] [Pronoun]
O3	[Conj] [Prep] [Determiner]	[NomStem] [NomBuff]	[NomSuff.DG]

		Class	Lemma/ Morpheme	Form	Gloss	gen	num	stt	cas	Set Conds	Required Conds
Lexicon	L1a	[NomStem]	*safiyr*	*safiyr*	ambassador	-	-	-	-		MS|FS|FP
	L1b	[NomStem]	*safiyr*	*sufaraA*	ambassador	m	p	-	-	#A' #dip	MS
Prc	P1	[Determiner]									
	P2	[Determiner]	Prc.Al	Al	the						
Buffers	B1	[NomBuff]									else
	B2a	[NomBuff]		'							#A'
	B2b	[NomBuff]		ŷ							#A' obj suff-i
	B2c	[NomBuff]		ŵ							#A' obj suff-u
Suffixes	S1a	[NomSuff.IG]	Suff.MSIG	ĩ		m	s	i	g	MS	else
	S1b	[NomSuff.IG]	Suff.MSIG	a		m	s	i	g	MS	#dip
	S2	[NomSuff.IG]	Suff.FPIG	aAt+ĩ		f	p	i	g	FP	
	S3	[NomSuff.CG]	Suff.MSCG	i		m	s	c	g	MS suff-i	
	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	
Enclitics	C1	[Pronoun]									
	C2a	[Pronoun]	Pron.3MP	hum	their					obj	else
	C2c	[Pronoun]	Pron.3MP	him	their					obj	suff-i

sufaraA+'+a	O1
safiyr+aAt+i+him	O2
sufaraA+ŷ+i+him	O2
Al+safiyr+aAt+i	O3

	✓		✓	
✓		✓		
				✓
✓				
		✓		
✓				
		✓		
	✓			
				✓
	✓	✓		

- 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 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 Specs		Our DB		Calima 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		83,649,166		28,359,701		**Unique Diacritized Forms**	(d)
	Morph Order Lines	42		246,880,683		126,176,265		**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 Specs		Our DB		Calima 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		83,649,166		28,359,701		**Unique Diacritized Forms**	(d)
	Morph Order Lines	42		246,880,683		126,176,265		**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

جامعة نيويورك أبوظبي

NYU | ABU DHABI

مختبر كامل

CAMeL Lab

Thank you!

morph.camel-lab.com

nizar.habash@nyu.edu