Camel Morph MSA: A Large-Scale Open-Source Morphological Analyzer for Modern Standard Arabic

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

Computational Approaches to Modeling Language Lab, New York University Abu Dhabi, UAE
{christian.khairallah,nizar.habash}@nyu.edu

Motivations

- Morphological analyzers add value to neural models in several NLP tasks:
  - Grammatical Error Correction (Alhafni et al., 2023)
  - Morphological Disambiguation (Inoue et al., 2022)
  - Machine Translation (Oudah et al., 2019)
- Make building analysis/generation tools easier and more consistent
- Make expansion of lexicons more efficient

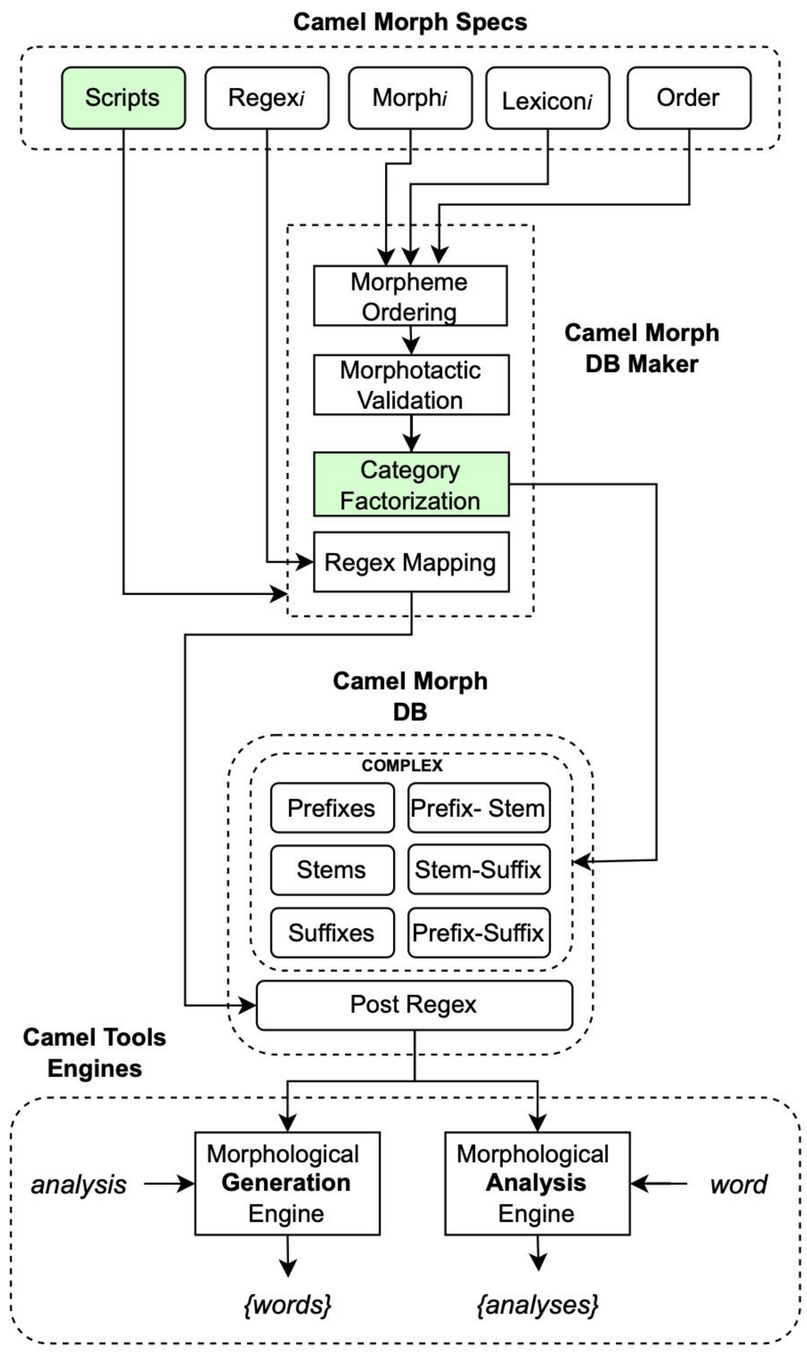
Contributions

- Released largest open-source Arabic morphological analyzer/generator
- Developing an extendable large-scale implementation using **Camel Morph**
- Benchmarking our models against a publicly available analyzer
- Our data and code are publicly available

morph.camel-lab.com

Camel Morph Framework (Habash et al., 2022)

- **Compile morphological specifications into optimized DBs**
  - ***From***: linguistic representations and rules
  - ***To***: lists of complex prefixes and suffixes, stems, and their compatibilities



Camel Morph Specifications & Database

- **Example of Morphology Specification and Validation**
  - Morph order defines the full space of all morphemes that can co-occur by their class.
  - Each form (allomorph) sets some truth conditions to be true.
  - For a word to be valid, the required truth conditions of every form (allomorph) 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	ī		m	s	i	g	MS	else
	S1b	[NomSuff.IG]	Suff.MSIG	a		m	s	i	g	MS	#dip
	S2	[NomSuff.IG]	Suff.FPIG	aAt+ī		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

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

		Count	*Verb Example*	Count	*Noun Example*	Count	*Particle Example*
Morphology Specifications	Lemma	1	ramaý 'throw' رَمَى	1	safiyr 'embassador' سَفِير	1	çalaý 'on, upon' عَلَى
	Proclitics	30	wa, fa, Âa, ... وَ، فَ، أَ، ...	19	Al, wa, fa, li, bi, ... ال، وَ، فَ، لِ، بِ، ...	9	wa, fa, Âa, ... وَ، فَ، أَ، ...
	Prefixes	13	ya, ta,... يَ، تَ، ...	N/A		N/A	
	Pre-Buffers	1	a ◌َ	N/A		N/A	
	Stems	3	رَم (ماضي)، رَم (مضارع)، رْم (أمر) ram (perfect), r.m (imperfect/command)	2	سَفِير، سُفَرَا safiyr (base), sufaraA (broken plural)	4	عَلَى، عَلِيّ، عَلَّام،... çalaý, çalay., çalaAma,...
	Post-Buffers	11	aý, ∅, ay, iy, ... ◌َي، ◌ٌ◌ي، ...	5	ŵ, ', ŷ, ∅, ... وٗ، ء، يٗ، ∅، ...	0	
	Suffixes	100	tu, ta, naA, at, ... تُ، تَ، نَا، ◌َت، ...	79	a, i, ū, aħu, aAti, ... ◌َ، ◌ِ، ◌ُ، ◌َاتِ،ةَ...	0	
	Enclitics	18	hu, hi, hum, him ... هُ، هِ، هُم، هِم، ...	19	hu, hi, hum, him ... هُ، هِ، هُم، هِم، ...	13	hu, hi, hum, haA, ... هُ، هِ، هُم، هَا، ...
	Conditions	25	#ay (defective),...	25	FP (+At suffix),...	6	enc0 (presence of object clitic)
	Order Seqs.	69	verbal orders	21	nominal orders	2	particle orders
DB	Complex Prefixes	3,129	wa, fa, Âa, waya, ... وَ، فَ، أَ، وَيَ، ...	199	wa, waAl, wali, lil, ... وَ، وَال، وَلِ، لِل، ...	14	wa, fa, >a, >awa, ... وَ، فَ، أَ، أَوَ، ...
	Complex Stems	10	ramaý, ramay., ... رَمَى، رَمِيَ، ...	4	safiyr, sufarAŵ, ... سَفِير، سُفَرَاوٗ، ...	4	çalaý, çalay., çalaAma,... عَلَى، عَلِيّ، عَلَّام،...
	Complex Suffixes	972	ta, tahu, hi, ... تُ، تَهُ، هِ،...	243	aAtu, aAtihim, aki,... ◌َاتُ، ◌َاتِهِم، ◌ِكِ، ...	14	hu, hi, hum, haA ... هُ، هِ، هُم، هَا، ...
	Compatibility Combinations	2,687	Stem-Suffix, Prefix-Stem, Prefix Suffix	394	Stem-Suffix, Prefix-Stem, Prefix Suffix	14	Stem-Suffix, Prefix-Stem, Prefix Suffix
Word Forms (unique analyses)		42,588	Word: Âaramay.tahu أَرَمَيْتَهُ `did you throw it?' Features: verb, perfective, active, 2nd person, masculine, singular, interrogative, 3rd person ms direct object	12,942	Word: walisafiyraAtihim وَلِسَفِيرَاتِهِم `and for their ambassadors' Features: noun, feminine, plural, genitive, construct, 3rd person mp possessive, conjunction, preposition	224	Word: façalay.hA فَعَلَيْهَا `so, upon it' Features: prep, 3rd person fs pron, conjunction

Evaluation

- Compare to Calima MSA (Taji et al. 2018)~SAMA (Graff et al., 2009)
- Coverage evaluation of Penn Arabic Treebank (Maamouri et al., 2004): Recall 95.9% of analyses; 90% of mismatches due to gold errors.

		Camel Morph MSA Specs		Camel Morph MSA DB		SAMA/CALIMA DB			
(a)	Lemmas (Stems)	105,102	(140,612)	105,102	(154,573)	42,218	(71,466)	Lemmas (Stems)	(c)
	Verbs	9,333	(38,156)	9,333	(47,540)	9,279	(26,343)	*Verbs*	
	Nominals	33,267	(39,837)	33,267	(44,414)	32,701	(44,742)	*Nominals*	
	Others	230	(347)	230	(347)	238	(381)	*Others*	
	Proper Nouns - Annex	62,272	(62,272)	62,272	(62,272)				
(b)	Prefix Morphs (Allom.)	60	(65)	14,726		4,640		DBPrefix Sequences	(d)
	Suffix Morphs (Allom.)	205	(406)	12,724		1,191		DBSuffix Sequences	
	Stem Buffers	111		15,044		979		Compatibility Entries	
	Unique Condition Terms	88		535,186,314	(242,824,398)	70,250,488		Unique Diacritized Forms (no Annex)	(e)
	Morph Order Sequences	122		1,447,312,125	(630,731,386)	227,471,211		Unique Analyses (no Annex)	
				8,070,764	(3,214,695)	1,514,577		Unique Analyses w/o Clitics (no Annex)	

	Camel Morph MSA		SAMA/CALIMA	
	MSA	CA	MSA	CA
Run Time (sec)	12,293	4,667	**4,231**	**1,960**
Type OOV	**67.9%**	**34.7%**	75.1%	43.2%
Token OOV	**2.3%**	**1.5%**	3.5%	2.5%
Analyses/Type	**18.9**	**21.2**	13.7	15.2
Analyses/Token	**38.6**	**45.7**	18.9	20.3