A Lexical Database for Modern Standard Arabic Interoperable with a Finite State Morphological Transducer

Mohammed Attia, Pavel Pecina, Antonio Toral, Lamia Tounsi, Josef van Genabith

National Centre for Language Technology (NCLT),
School of Computing, Dublin City University

Funded by:
Enterprise Ireland, the Irish Research Council for Science Engineering and Technology (IRCSET), and
the EU projects PANACEA and META-NET

Contribution

- We develop a finite state morphological transducer for Modern Standard Arabic
  1. Open source, distributed under the GPLv3 license
  2. Large scale, more than 30,000 lemmas
  3. Corpus based, truly representative of Modern Standard Arabic and not Classical Arabic.
  4. Compatible with Foma, an open-source fst compiler

Short Tutorial

(1) Download Foma

http://foma.sourceforge.net

(2) Download AraComLex

http://aracomlex.sourceforge.net

(3) Build the transducer: README

The transducer online

- You can test the transducer online:

  http://www.cngl.ie/aracomlex

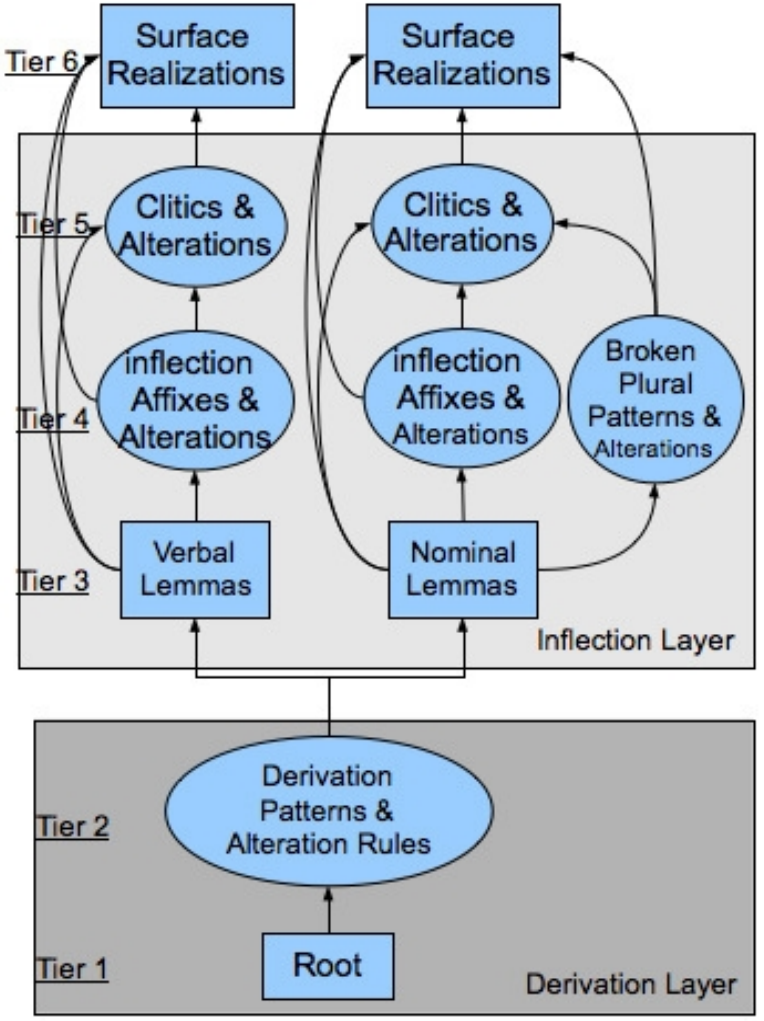
Introduction

- Modern Standard Arabic vs. Classical Arabic
- Current State of Arabic Lexicography
  - Lexicons are not corpus-based
  - Buckwalter Arabic Morphological Analyser
- Importance of Lexical Resources

Introduction

- Arabic Morphotactics

Root	درس drs			
Pattern	$R_1aR_2aR_3a$	$R_1aR_2R_2aR_3a$	$R_1\bar{a}R_2iR_3$	$muR_1aR_2R_2iR_3$
POS	V	V	N	N
lemma	d a r a s a 'study'	d a r r a s a 'teach'	d ā r i s 'student'	m u d a r r i s 'teacher'



Aim

- Constructing a lexical database for Modern Standard Arabic
- Building a finite-state morphological transducer

Methodology

- Using a medium-scale manually created lexicon of 10,799 lemmas, with detailed info for:
  - Nouns (human/nonhuman, POS, Continuation Classes)
  - Verbs (transitive/intransitive, allow passive, allow imperative)
- Using statistics from a 1 billion word corpus
  - 90% from the LDC's Arabic Gigaword
  - 10% collected from the Al-Jazeera website
- Using a pre-annotation tool: MADA+TOKAN

Methodology

- Using Finite State Technology (XFST)
  - Bidrectional: Suitable for analysis and generation
  - handles concatenative and non-concatenative morphotactics
  - Speed and efficiency in dealing with millions of paths
  - Handles separated dependencies.
  - Handles phonological and orthographic changes through alteration rules.

Methodology

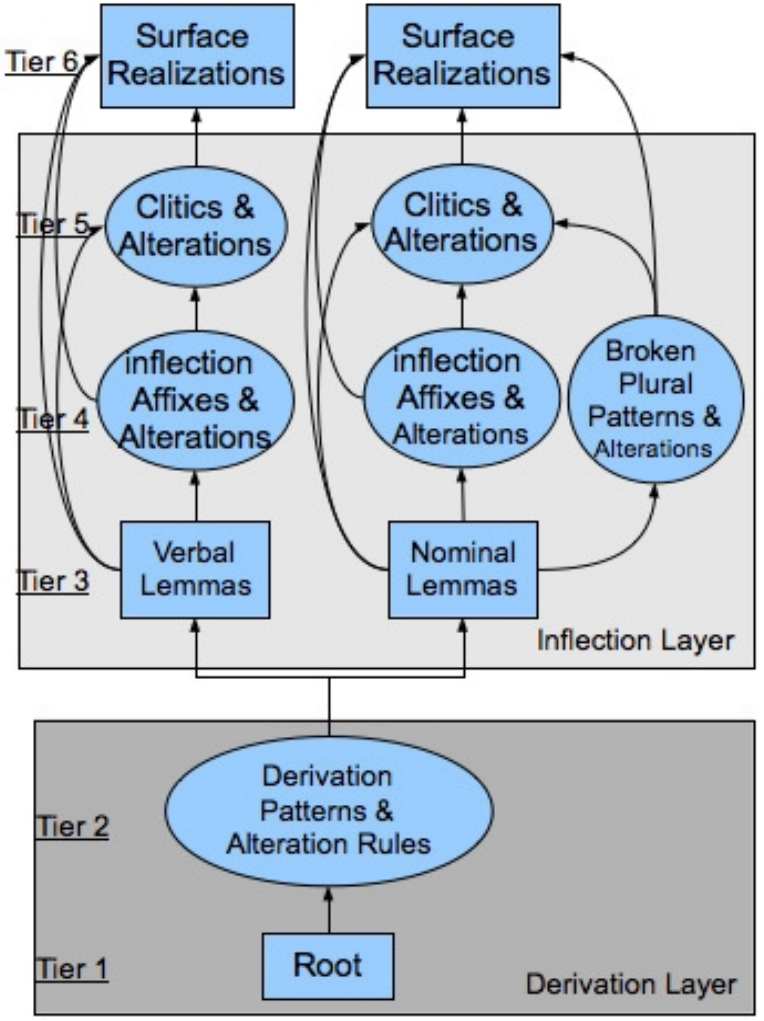
- Design Approach: Three approaches
  - Root-based Morphology

    Xerox Arabic FTM
  - Stem-based morphology

    Buckwalter

$kr	$akar	PV	thank;give thanks
$kr	$okur	IV	thank;give thanks

  - Lemma-based morphology



Methodology

Our Approach:

Lemma-based morphology

```
(1) LEXICON Verbs
^ss^شَكَرَ['thank']^se^              Transitive;
^ss^فَرِحَ['be-happy']^se^@D.V.P@    Intransitive;
^ss^أَمَرَ['order']^se^@D.M.I@       Transitive;
^ss^قَالَ['say']^se^                 Intransitive;

(2) LEXICON Nouns
+masc+human^ss^مُعَلِّم['teacher']^se^      FemMascduFemduFemplMascpl;
+masc+human^ss^طَالِب['student']^se^        FemMascduFemduFempl;
+masc+nonhuman^ss^كِتَاب['book']^se^        Mascdu;
+fem+nonhuman^ss^كُرَّاسَة['notebook']^se^  DualFempl;
```

	Masculine Singular	Feminine Singular	Masculine Dual	Feminine Dual	Masculine Plural	Feminine Plural	Continuation Class
1	مُعَلِّم *muʿallim* 'teacher'	مُعَلِّمَة *muʿallimat*	مُعَلِّمَان *muʿallimān*	مُعَلِّمَتَان *muʿallimatān*	مُعَلِّمُون *muʿallimuwn*	مُعَلِّمَات *muʿallimāt*	Fem-Mascdu-Femdu-Mascpl-Fempl
2	طَالِب *ṭālib* 'student'	طَالِبَة *ṭālibat*	طَالِبَان *ṭālibān*	طَالِبَتَان *ṭālibatān*	-	طَالِبَات *ṭālibāt*	Fem-Mascdu-Femdu-Fempl
3	تَحْضِيرِيّ *taḥḍiyriyy* 'preparatory'	تَحْضِيرِيَّة *taḥḍiyriyyat*	تَحْضِيرِيَّان *taḥḍiyriyyān*	تَحْضِيرِيَّتَان *taḥḍiyriyyatān*	-	-	Fem-Mascdu-Femdu
4	-	بَقَرَة *baqarat* 'cow'	-	بَقَرَتَان *baqaratān*	-	بَقَرَات *baqarāt*	Femdu-Fempl
5	تَنَازُل *tanāzul* 'concession'	-	-	-	-	تَنَازُلَات *tanāzulāt*	Fempl
6	-	ضَحِيَّة *ḍaḥiyyat* 'victim'	-	ضَحِيَّتَان *ḍaḥiyyatān*	-	-	Femdu
7	مَحْض *maḥḍ* 'mere'	مَحْضَة *maḥḍat*	-	-	-	-	Fem
8	اِمْتِحَان *imtiḥān* 'exam'	-	اِمْتِحَانَان *imtiḥānān*	-	-	اِمْتِحَانَات *imtiḥānāt*	Mascdu-Femdu
9	طَيَّار *ṭayyār* 'pilot'	-	-	-	.tayyAruwn	-	Mascdu-Mascpl
10	كِتَاب *kitāb* 'book'	-	كِتَابَان *kitābān*	-	-	-	Mascdu
11	دِيمُقْرَاطِيّ *diymuqrāṭiyy* 'democrat'	-	-	-	دِيمُقْرَاطِيُّون *diymuqrāṭiyyuwn*	-	Mascpl
12	خُرُوج *ḫuruwǧ* 'exiting'	-	-	-	-	-	NoNum
13	مَبَاحِث *mabāḥiṯ* 'investigators'	-	-	-	-	-	Irreg_pl

Methodology

Alteration Rules:

Alteration Rules are used for handling discrepancies between surface forms and underlying representation or lemmas. We have 130 replace rules.

a -> b || L _ R

Results to Date

- Start-off with a seed lexicon
  - Four Lexical Databases, manually constructed
    - 5,925 nominal lemmas
    - 1,529 verb lemmas
    - 490 patterns (456 for nominals and 34 for verbs)
    - lemma-root look up database

Results to Date

- Automatically Extending the Lexical Database: Lexical Enrichment
  - Data-driven filtering technique
    - 40,648 lemmas (in Buckwalter or SAMA 3.1)
    - Statistics from three web search engines
    - Statistics from the corpus annotated by MADA
    - 29,627 lemmas (left after filtering)

Results to Date

Automatically Extending the Lexical Database: Feature Enrichment

- Machine Learning
- Multilayer Peceptron classification algorithm
- Training Data: 4,816 nominals and 1,448 verbs
- Classes for nominals: continuation classes (or inflection paths), the semantico-grammatical feature of humanness, and POS (noun or adjective)
- Classes for verbs: transitivity, allowing the passive voice, and allowing the imperative mood
- We feed these datasets with frequency statistics from the corpus and build a vector grid.

Results to Date

- Extending the Lexical Database
  - Feature enrichment using Machine Learning

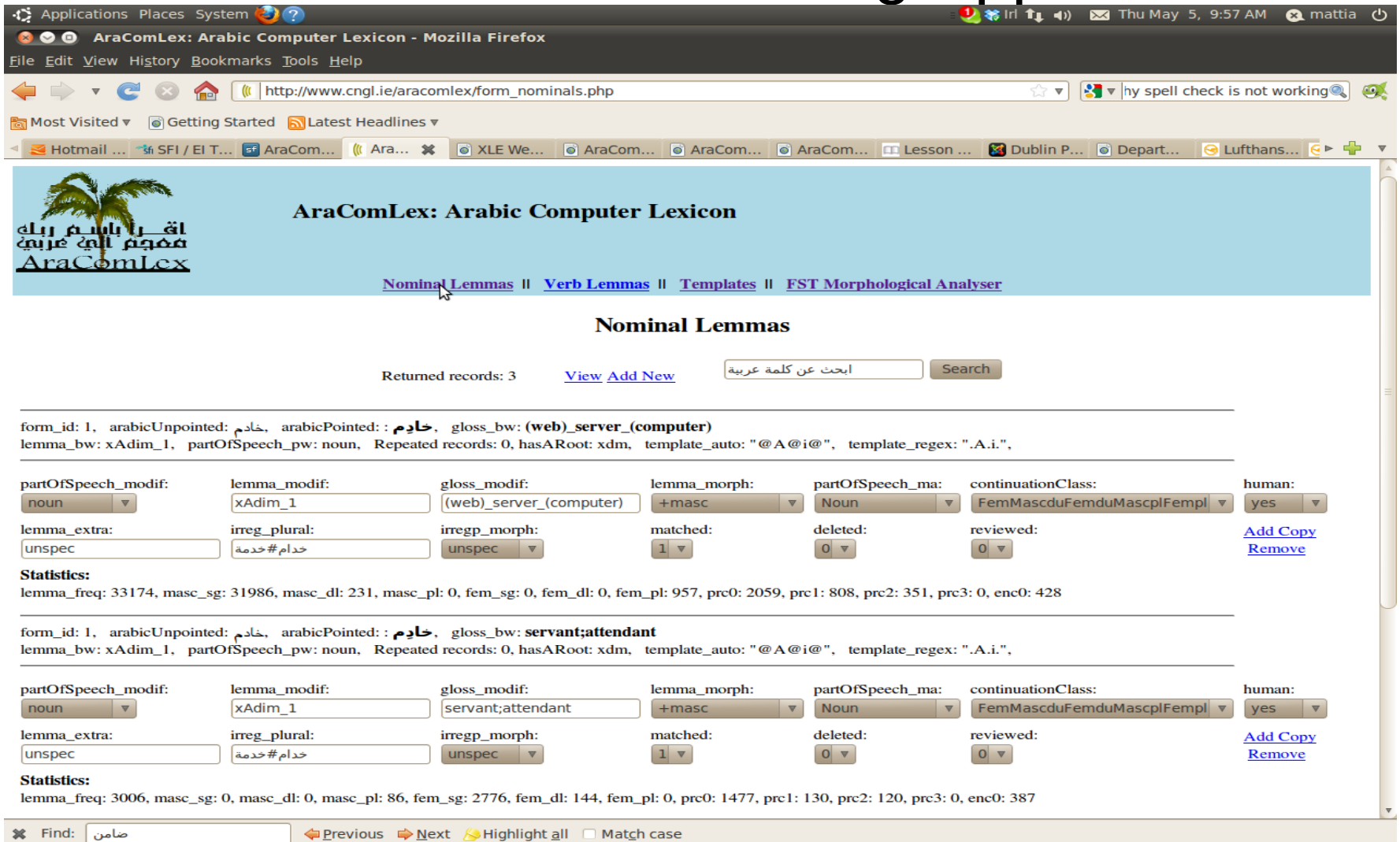
Exp.	Classes	Features	P	R	F
	Nominals				
1	Continuation Classes: 13 classes	number, gender, case, clitics	0.62	0.65	0.63
2	Human: yes, no, unspecified		0.86	0.87	0.86
3	POS: noun, adjective		0.85	0.86	0.85
	Verbs				
4	Transitivity: transitive, intransitive	number, gender, person, aspect, mood, voice, clitics	0.85	0.85	0.84
5	Allow passive: yes, no		0.72	0.72	0.72
6	Allow imperative: yes, no		0.63	0.65	0.64

Results to Date

- Extending the Lexical Database
  - With Machine Learning we add:
    - 18,000 new lemmas:
      - 12,974 nominals
      - 5,034 verbs

Results to Date

- AraComLex Lexicon Writing Application



Results to Date

- FST Morphology Coverage and RPW Results
  - a test corpus of 800,000 words, divided as
    - 400,000 for Semi-Literary text
    - 400,000 for General News texts.

Morphology	No. of Lemmas	General News		Semi-Literary	
		Coverage	Rate per word	Coverage	Rate per word
AraComLex 1.0	10,799	79.68%	1.67	69.37%	1.62
AraComLex 2.0	28,807	86.89%	2.10	85.14%	2.09
AraComLex 2.1	30,587	87.13%	2.09	85.73%	2.08
SAMA	40,648	88.13%	5.32	86.95%	5.30

Future Work

- Going beyond SAMA
- Including Named Entities and MWEs
- Building a spell checker

Conclusion

- Open-source finite state transducer for Modern Standard Arabic (AraComLex) distributed under the GPLv3 license.
- We successfully use machine learning to predict morpho-syntactic features for newly acquired words.
- Comparing our morphological transducer to SAMA, we find that we achieve comparable coverage and lower rate of analyses per word.