Automatic Lexical Resource Acquisition for Constructing an LMF-Compatible Lexicon of Modern Standard Arabic

Mohammed Attia, Lamia Tounsi, Josef van Genabith

Outline

- Introduction
- Principles of Lexicography
- Modern Standard Arabic vs. Classical Arabic
- Review of Arabic lexicographic work
- Project Description
- Lexical Markup Framework (LMF)
- Automatic Acquisition Architecture
- Results and evaluation
- Conclusion

Introduction

- Importance of lexical resources for NLP tasks
  - The backbone of morphological analysers
  - Principal factor for coverage
  - Unknown words in parsing cause a problem and we want to minimize them as much as possible
- Advantages of automatic acquisition vs. Manual construction of dictionaries
  - Time and effort
  - Speed and efficiency
  - Consistency and quality
  - It is impractical to manually analyse large and ever-growing amounts of data
- How the resource will fit in the annotation/parsing tools
  - DCU Parser
  - DCU annotation tools

Principles of Lexicography

Definition of a dictionary

- A description of the vocabulary used by members of a speech community. A dictionary deals with:
  - conventions not idiosyncrasies
  - norms not rarities
  - probable not possible

- Lexical evidence
  - Subjective evidence
    - introspection
    - informant-testing
  - Objective evidence
    - A corpus provides typifications of the language
      - A typical lexical entry means it is both “frequent” or “recurrent” and “well-dispersed” in a corpus.
      - A typical lexical entry belongs to the stable “core” of the language.

Atkins, B. T. S. and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography.* Oxford *University Press.*

Principles of Lexicography

- **Corpora and Dictionaries**
  - Brown Corpus, 1 million words, 1960s,
    → Citations for *American Heritage Dictionary*
  - Birmingham corpus, 20 million words, 1980s
    → Cobuild English Dictionary.
  - British National Corpus (BNC), 100 million words, 1990s set the standard (balance, encoding)
  - The Oxford English Corpus, one billion words, 2000s
    → Oxford English Dictionary
  - Longman Corpus Network, 330 million word
    → Longman Dictionaries

Principles of Lexicography

- **Dictionaries before Corpora**
  - Citation banks
    - A citation is a short extract providing evidence for a word usage or meaning in authentic use.
  - Disadvantages
    - labour-intensive
    - instances of usage are authentic, but there is a big subjective element in their selection.
      - People tend to notice what is remarkable and ignore what is typical
      - bias towards the novel or idiosyncratic usages

Principles of Lexicography

- **Characteristics of a reliable corpus**
  - The corpus does not favour high class language
  - The Corpus should be large and diverse
  - The corpus should be either synchronic or diachronic
  - The corpus should be well-balanced using “stratified sampling”
  - The corpus should avoid skewing

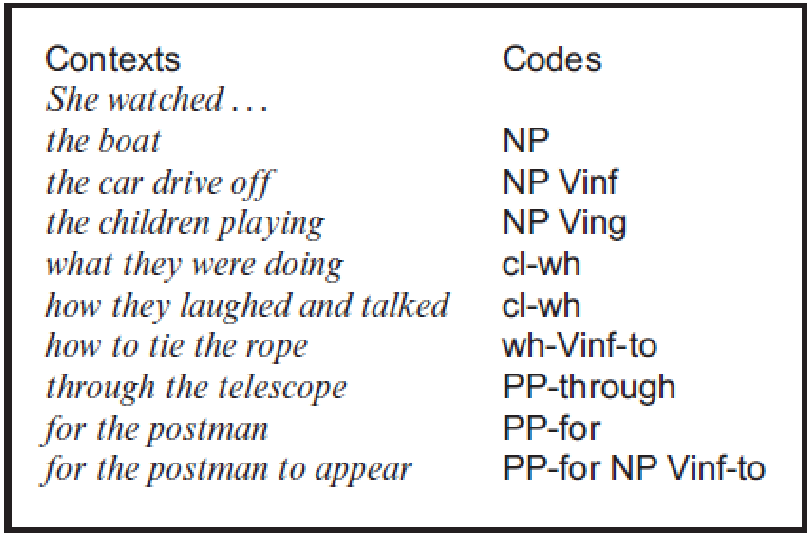
Principles of Lexicography

- **Lexical Profiling**
  - Word POS

    v, n, adj, adv, conj, det, interj, prep, pron
  - Valency Information
  - Collocations

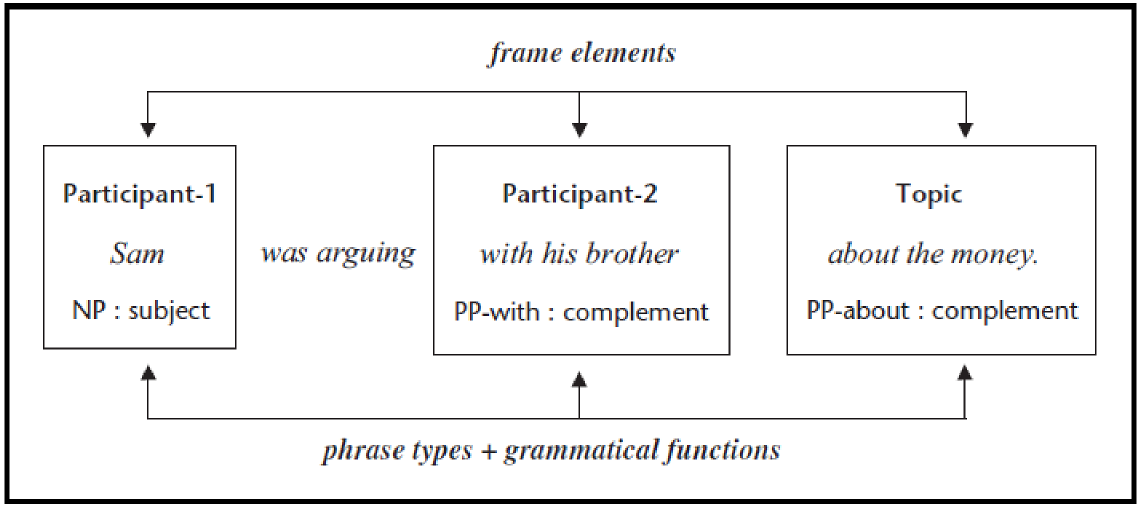
    commit a crime, sky blue, lame duck
  - Colligational preferences

    was acquitted, trials (difficult experiences)



Contexts	Codes
She watched . . .	
the boat	NP
the car drive off	NP Vinf
the children playing	NP Ving
what they were doing	cl-wh
how they laughed and talked	cl-wh
how to tie the rope	wh-Vinf-to
through the telescope	PP-through
for the postman	PP-for
for the postman to appear	PP-for NP Vinf-to

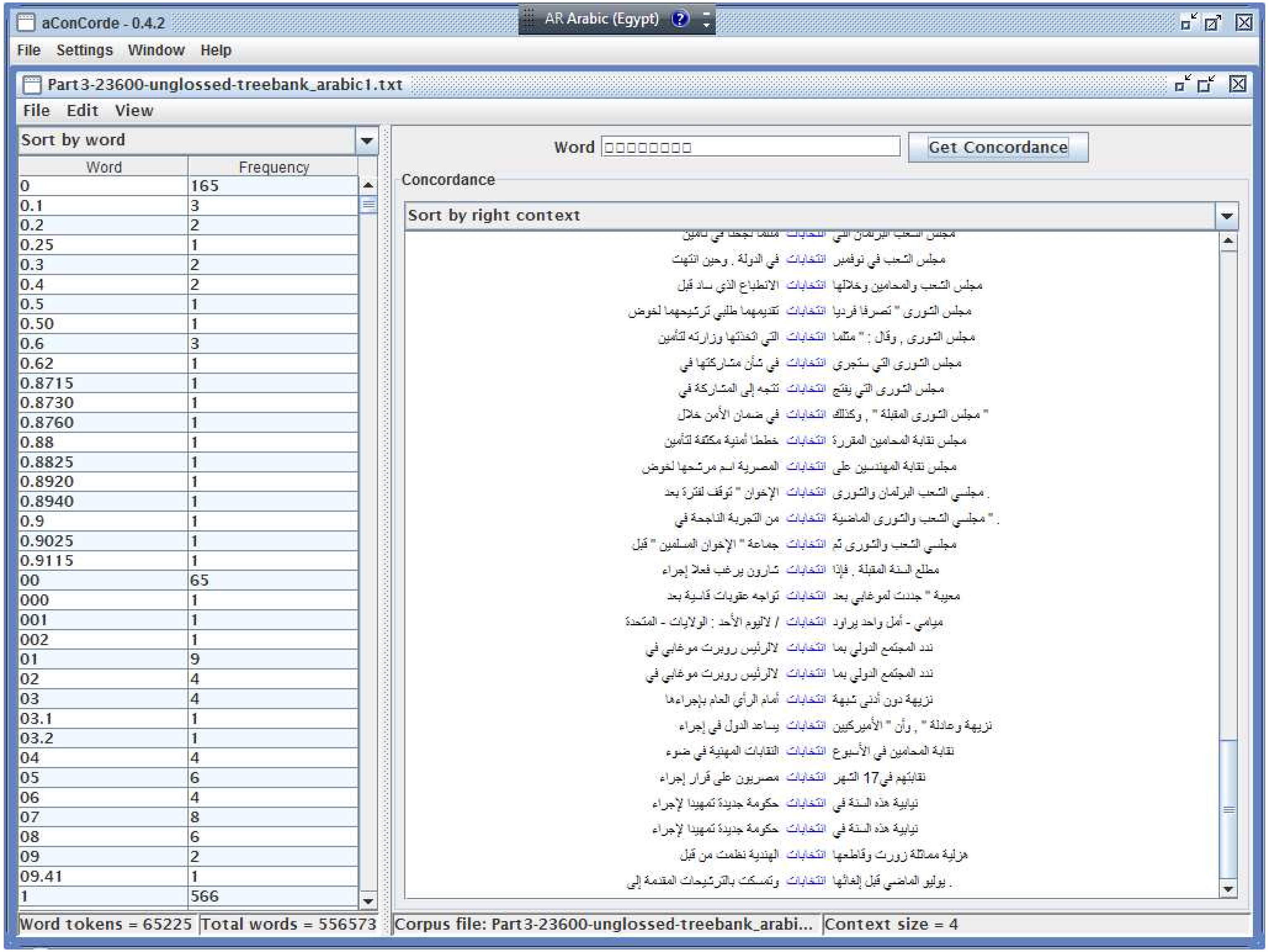
Some constructions for the verb *watch*



Principles of Lexicography

- **Lexical Profiling Software**
- Concordancers
- Word Sketch (Sketch Engine) - Adam Kilgarriff

Concordancer



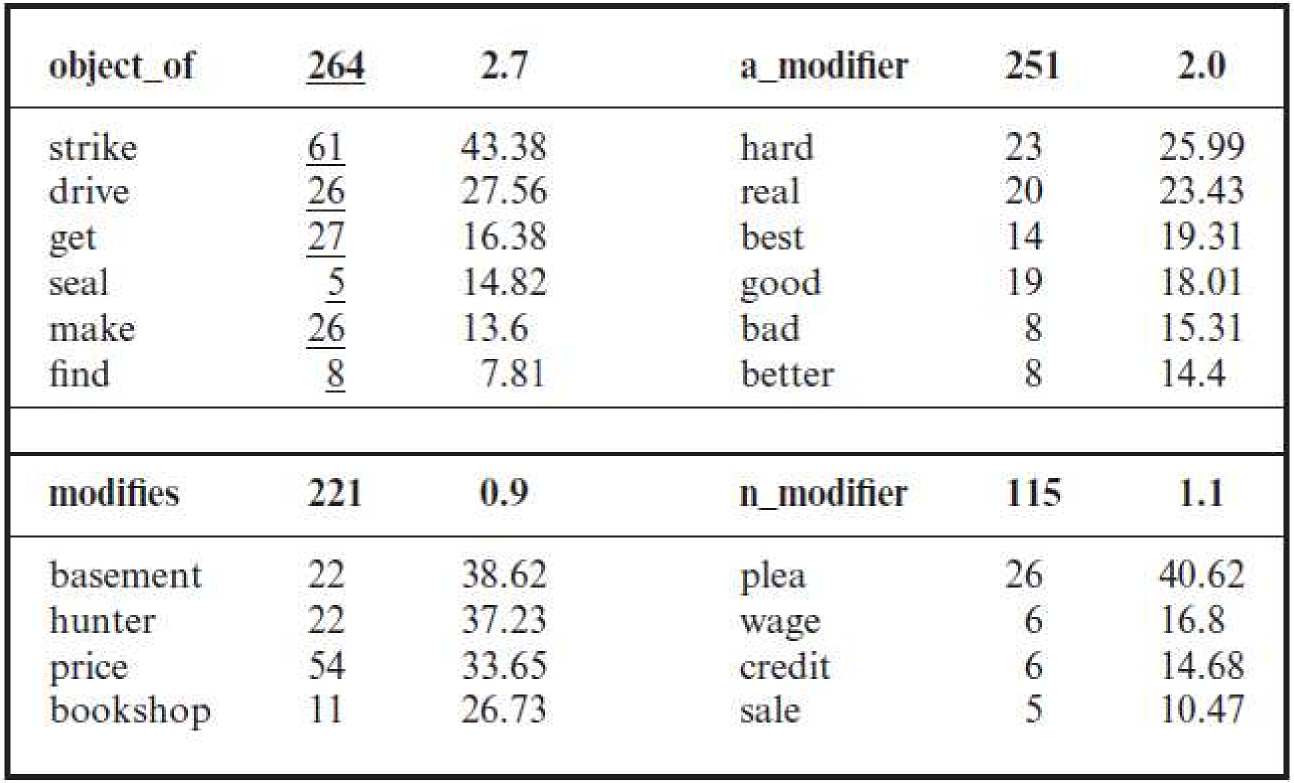
Word	Frequency
0	165
0.1	3
0.2	2
0.25	1
0.3	2
0.4	2
0.5	1
0.50	1
0.6	3
0.62	1
0.8715	1
0.8730	1
0.8760	1
0.88	1
0.8825	1
0.8920	1
0.8940	1
0.9	1
0.9025	1
0.9115	1
00	65
000	1
001	1
002	1
01	9
02	4
03	4
03.1	1
03.2	1
04	4
05	6
06	4
07	8
08	6
09	2
09.41	1
1	566

مجلس الشعب البرلمان التي انتخابات مثلما نجحنا في تأمين
مجلس الشعب في نوفمبر انتخابات في الدولة . وحين انتهت
مجلس الشعب والمحامين وخلالها انتخابات الانطباع الذي ساد قبل
مجلس الشورى " تصرفا فرديا انتخابات تقديمهما طلبي ترشيحهما لخوض
مجلس الشورى , وقال : " مثلما انتخابات التي اتخذتها وزارته لتأمين
مجلس الشورى التي ستجري انتخابات في شأن مشاركتها في
مجلس الشورى التي يفتح انتخابات تتجه إلى المشاركة في
" مجلس الشورى المقبلة " , وكذلك انتخابات في ضمان الأمن خلال
مجلس نقابة المحامين المقررة انتخابات خططا أمنية مكثفة لتأمين
مجلس نقابة المهندسين على انتخابات المصرية اسم مرشحها لخوض
. مجلسي الشعب البرلمان والشورى انتخابات الإخوان " توقف لفترة بعد
. " مجلسي الشعب والشورى الماضية انتخابات من التجربة الناجحة في
مجلسي الشعب والشورى ثم انتخابات جماعة " الإخوان المسلمين " قبل
مطلع السنة المقبلة . فإذا انتخابات شارون يرغب فعلا إجراء
معيبة " جددت لموغابي بعد انتخابات تواجه عقوبات قاسية بعد
ميامي - أمل واحد يراود انتخابات / لااليوم الأحد : الولايات - المتحدة
ندد المجتمع الدولي بما انتخابات لالرئيس روبرت موغابي في
ندد المجتمع الدولي بما انتخابات لالرئيس روبرت موغابي في
نزيهة دون أدنى شبهة انتخابات أمام الرأي العام بإجراءها
نزيهة وعادلة " , وأن " الأميركيين انتخابات يساعد الدول في إجراء
نقابة المحامين في الأسبوع انتخابات النقابات المهنية في ضوء
نقابتهم في17 الشهر انتخابات مصريون على قرار إجراء
نيابية هذه السنة في انتخابات حكومة جديدة تمهيدا لإجراء
نيابية هذه السنة في انتخابات حكومة جديدة تمهيدا لإجراء
هزلية مماثلة زورت وقاطعها انتخابات الهندية نظمت من قبل
. يوليو الماضي قبل إلغائها انتخابات وتمسكت بالترشيحات المقدمة إلى

Word tokens = 65225 | Total words = 556573

Corpus file: Part3-23600-unglossed-treebank_arabi... | Context size = 4

Sketch Engine



object_of	**264**	**2.7**	**a_modifier**	**251**	**2.0**
strike	61	43.38	hard	23	25.99
drive	26	27.56	real	20	23.43
get	27	16.38	best	14	19.31
seal	5	14.82	good	19	18.01
make	26	13.6	bad	8	15.31
find	8	7.81	better	8	14.4
modifies	**221**	**0.9**	**n_modifier**	**115**	**1.1**
basement	22	38.62	plea	26	40.62
hunter	22	37.23	wage	6	16.8
price	54	33.65	credit	6	14.68
bookshop	11	26.73	sale	5	10.47

Part of the Word Sketch for the noun *bargain*

Modern Standard Arabic vs. Classical Arabic vs. Colloquial Arabic

- Modern Standard Arabic
  - The language of modern writing, prepared speeches and the language of the news
- Classical Arabic
  - The language of Arabia before Islam and after Islam until the Medieval Times
  - Present religious teaching, poetry and scholarly literature.
- Colloquial Arabic
  - Variety of Arabic spoken regionally and which differs from one country or area to another. They are to a certain extent mutually intelligible.

Code Shifting – Diglossia – multi-layered diglossia

Modern Standard Arabic vs. Classical Arabic vs. Colloquial Arabic

- Modern Standard Arabic
  - Tendency for simplification
    - Some CA structures to die out
    - Structures marginal in CA started to have more salience
    - no strict abidance by case ending rules
  - A subset of the full range of structures, inflections and derivations available in CA
  - MSA conforms to the general rules of CA
  - How “big” or how “small” the difference (on morphological, lexical or syntactic levels) need more research and investigation

Review of Arabic lexicographic work

- *Kitab al-'Ain* by al-Khalil bin Ahmed al-Farahidi (died 789)

(refinement/expansion/organizational Improvement)

▼

- *Tahzib al-Lughah* by Abu Mansour al-Azhari (died 980)
- *al-Muheet* by al-Sahib bin 'Abbad (died 995)
- *Lisan al-'Arab* by ibn Manzour (died 1311)
- *al-Qamous al-Muheet* by al-Fairouzabadi (died 1414)
- *Taj al-Arous* by Muhammad Murtada al-Zabidi (died 1791)
- *Muheet al-Muheet* (1869) by Butrus al-Bustani
- *al-Mu'jam al-Waseet* (1960)

Review of Arabic lexicographic work

- Bilingual Dictionaries
  - Edward William Lane's *Arabic-English Lexicon* (1876) indebted to *Taj al-Arous* by al-Zabidi
  - Hans Wehr's Dictionary of Modern Written Arabic (1961)
    - Size: 45,000 entries
    - Aim: Using scientific descriptive principles to describe present-day vocabulary through wide reading in literature of every kind
    - Application
      - Selection of works by high flying poets and literary critics such as Taha Husain, Taufiq al-Hakim, Mahmoud Taimur, al-Manfalauti, Jubran Khalil Jubran
      - Use of secondary sources (dictionaries) for expansion
      - Inclusion of rarities and classisms that no longer formed a part of the living lexicon

Review of Arabic lexicographic work

- Bilingual Dictionaries
  - Landau and Brill (1959) *A Word Count of Modern Arabic Prose*
    - A word count based on 270,000 words based on the news and 60 contemporary books on:
fiction, literary criticism, history, biography, political science, religion, social studies, economics, travels and historical novels
    - 6,000 words in the news
    - 11,000 words in literature
    - 12,400 words in the combined list (does not include proper nouns)

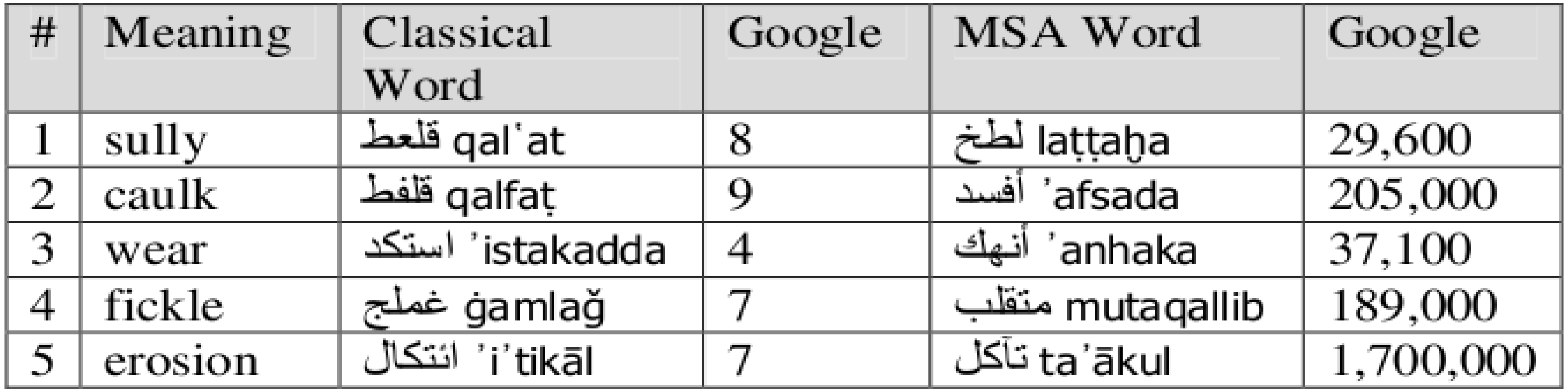
Review of Arabic lexicographic work

- Bilingual Dictionaries
  - Van Mol’s (2000) Arabic-Dutch learner’s dictionary
    - COBUILD-style, Corpus-based (3 million words)
    - Manually constructed
    - Covers the whole range of the actual vocabulary in the corpus with 17,000 entries compared to 45,000 entries in Hans Wehr
    - 5% of frequent new words not found in Hans Wehr

Review of Arabic lexicographic work

- Bilingual Dictionaries
  - Buckwalter Arabic Morphological Analyzer (2002)
    - Size: 40,222 lemmas (including 2,034 proper nouns)
    - Includes many obsolete lexical items

(But how many?)

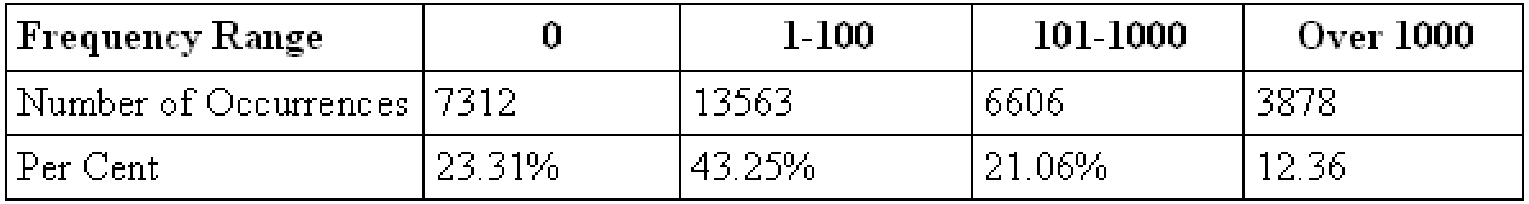


#	Meaning	Classical Word	Google	MSA Word	Google
1	sully	قلعط qalʿat	8	لطخ laṭṭaḫa	29,600
2	caulk	قلفط qalfaṭ	9	أفسد ʾafsada	205,000
3	wear	استكد ʾistakadda	4	أنهك ʾanhaka	37,100
4	fickle	غملج ġamlaǧ	7	متقلب mutaqallib	189,000
5	erosion	ائتكال ʾiʾtikāl	7	تآكل taʾākul	1,700,000

Google score for Classical vs. MSA entries

Review of Arabic lexicographic work

- Bilingual Dictionaries
  - Buckwalter Arabic Morphological Analyzer (2002)
    - Searching for Buckwalter on Aljazeera (40,205 reduced to 31,359 after removing diacritics)
    - Both false positives and false negatives are possible but the figures are still indicative



Frequency Range	**0**	**1-100**	**101-1000**	**Over 1000**
Number of Occurrences	7312	13563	6606	3878
Per Cent	23.31%	43.25%	21.06%	12.36

Project Description

- Acquisition of Arabic lexical resources
  - from corpora
  - modern language
  - varied domains
  - inducing the lexical profile for each lemma (frequency, inflections, derivations, citation, etc.)
- Production of new lexical sets
  - accumulated in a MySQL database
  - meeting Lexical Markup Framework specifications
  - exported into Lexical Markup Framework format

Project Description

- How our lexical database will be different from Buckwalter’s. We include
  - only entries attested in a corpus
  - subcategorization frames
  - +/-human semantic information for nouns
  - detailed information about derived nouns/adjectives (active or passive participle or a verbal noun, *masdar*)
  - multi-word expressions
  - classification of proper nouns: person, place, organization, etc.

Lexical Markup Framework (LMF)

ISO 24613:2008

- Aims:
  - Managing lexical resources
  - Providing a metamodel (or a super-hierarchy) to accommodate lexical information at all levels
  - Provides specifications, encoding format and naming conventions to ensure consistency
  - Enable the merger of individual electronic lexical resources
  - Allows instantiation of monolingual, bilingual or multilingual lexical resources
  - Allows work at a small scale or large scale
  - Tries to cover all natural languages (including languages with rich and complex morphology such as Arabic)

Lexical Markup Framework (LMF)

ISO 24613:2008

- History:
  - It started in 2003
  - Earlier lexicon standardization projects include GENELEX, EDR, EAGLES, MULTEXT, PAROLE, SIMPLE and ISLE
  - Project team:
    - Nicoletta Calzolari (Italy)
    - Gil Francopoulo (France)
    - Monte George (US)
    - + A panel of 60 experts
  - Published officially as an International Standard in 2008
  - Uses Unified Modeling Language (UML)
  - LMF is considered the state of the art in NLP lexicon management field

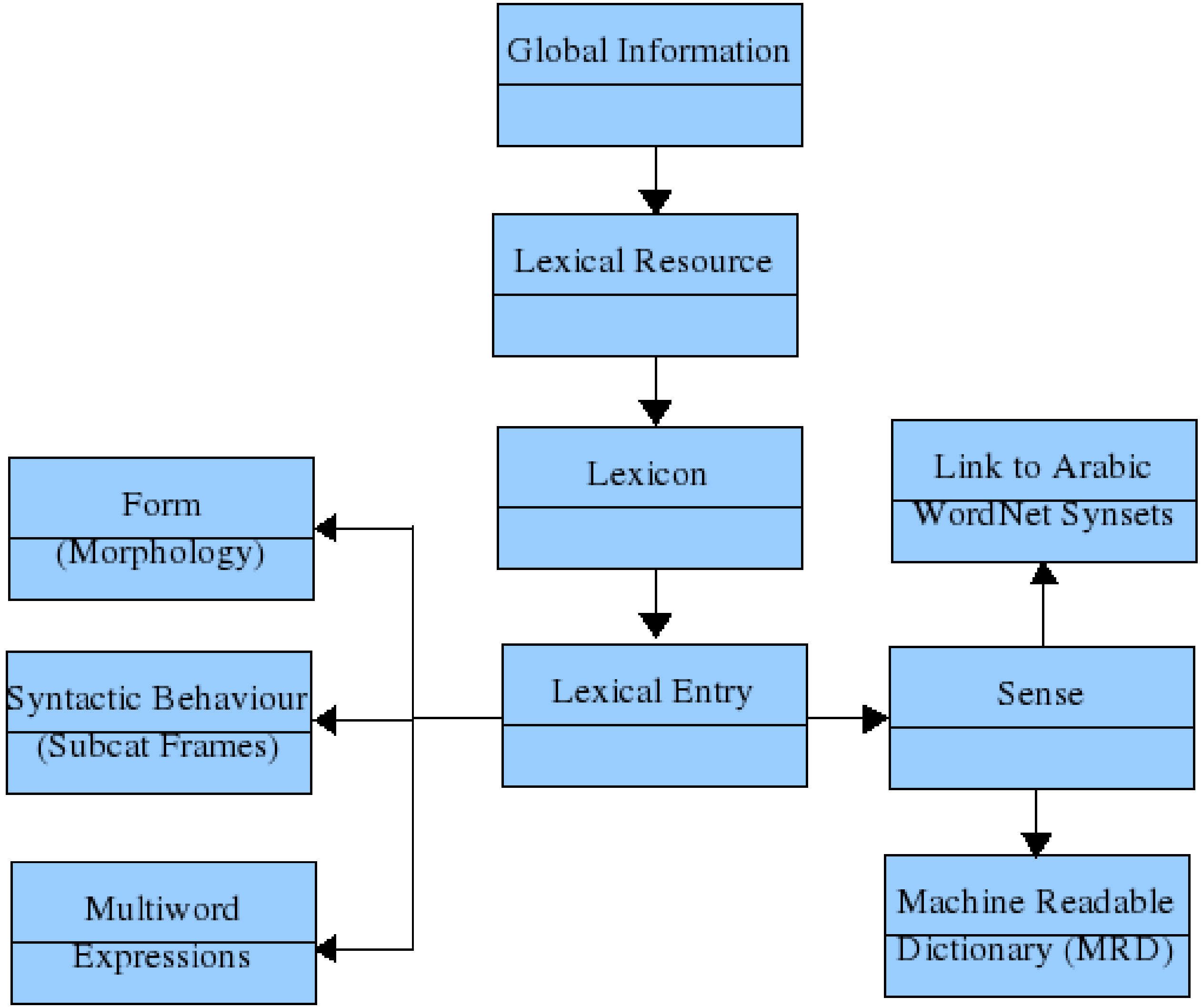
Lexical Markup Framework (LMF)

Architecture

1. Morphology extension
2. Machine Readable Dictionary extension
3. NLP syntax extension
4. NLP semantics extension
5. NLP multiword expression patterns extension

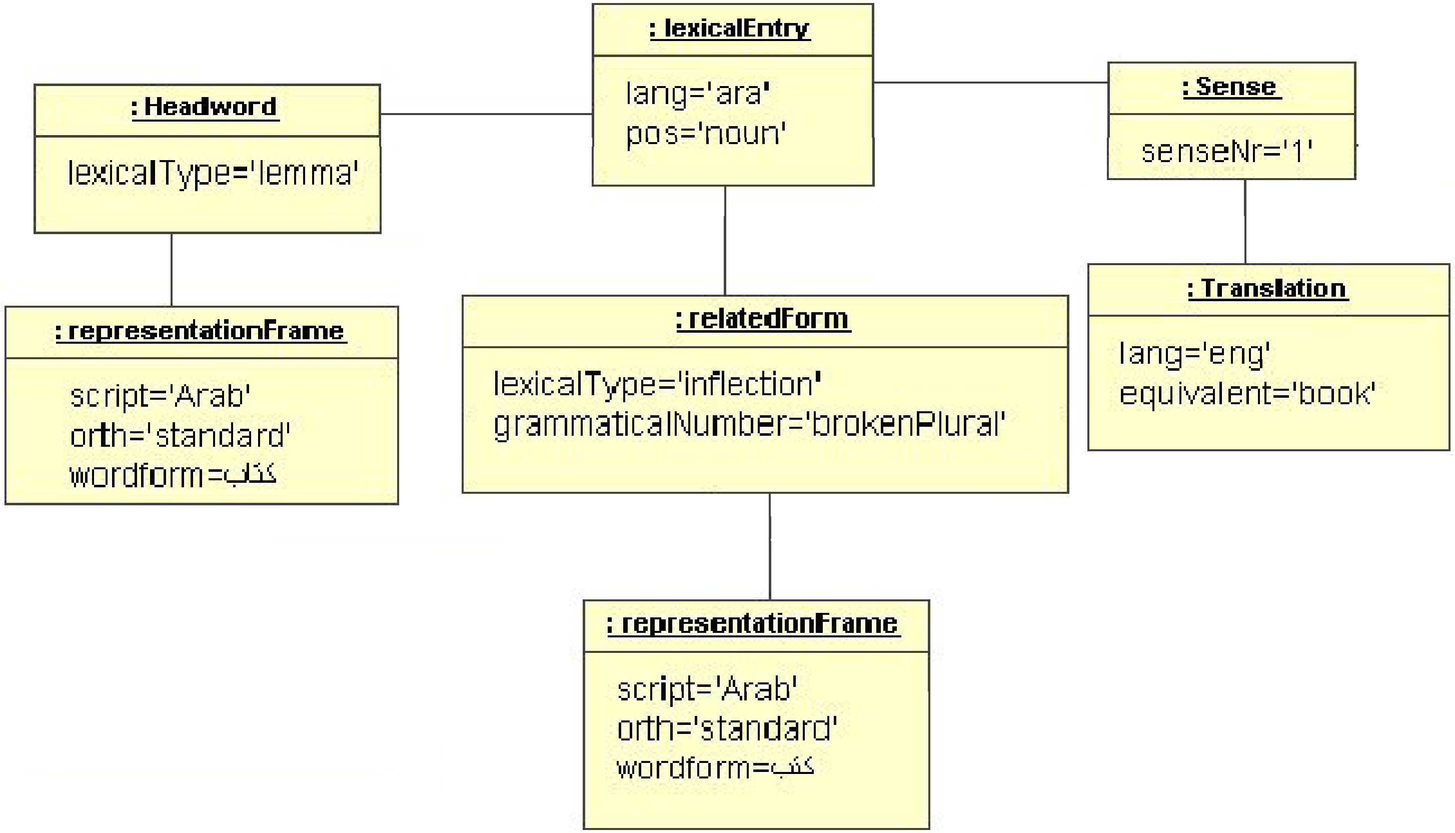
Lexical Markup Framework (LMF)

Architecture


Global Information
Lexical Resource
Lexicon
Link to Arabic WordNet Synsets
Form (Morphology)
Syntactic Behaviour (Subcat Frames)
Lexical Entry
Sense
Multiword Expressions
Machine Readable Dictionary (MRD)

Lexical Markup Framework (LMF)

Simple Example: Noun



: lexicalEntry

lang='ara'
pos='noun'

: Headword

lexicalType='lemma'

: Sense

senseNr='1'

: representationFrame

script='Arab'
orth='standard'
wordform=كتاب

: relatedForm

lexicalType='inflection'
grammaticalNumber='brokenPlural'

: Translation

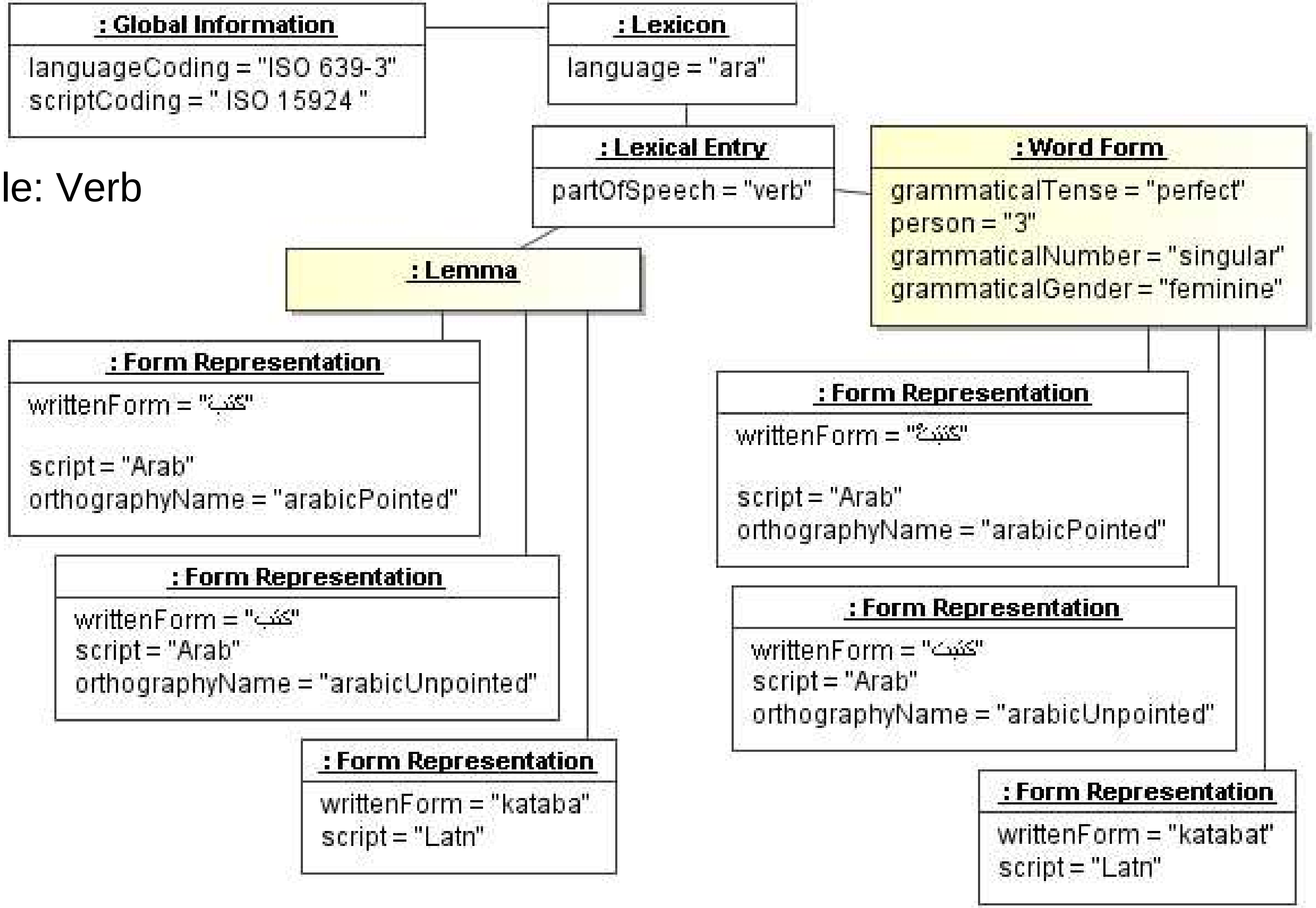
lang='eng'
equivalent='book'

: representationFrame

script='Arab'
orth='standard'
wordform=كتب

Lexical Markup Framework (LMF)

Example: Verb



: Global Information

languageCoding = "ISO 639-3"
scriptCoding = " ISO 15924 "

: Lexicon

language = "ara"

: Lexical Entry

partOfSpeech = "verb"

: Lemma

: Form Representation

writtenForm = "كَتَبَ"

script = "Arab"
orthographyName = "arabicPointed"

: Form Representation

writtenForm = "كتب"
script = "Arab"
orthographyName = "arabicUnpointed"

: Form Representation

writtenForm = "kataba"
script = "Latn"

: Word Form

grammaticalTense = "perfect"
person = "3"
grammaticalNumber = "singular"
grammaticalGender = "feminine"

: Form Representation

writtenForm = "كَتَبَتْ"

script = "Arab"
orthographyName = "arabicPointed"

: Form Representation

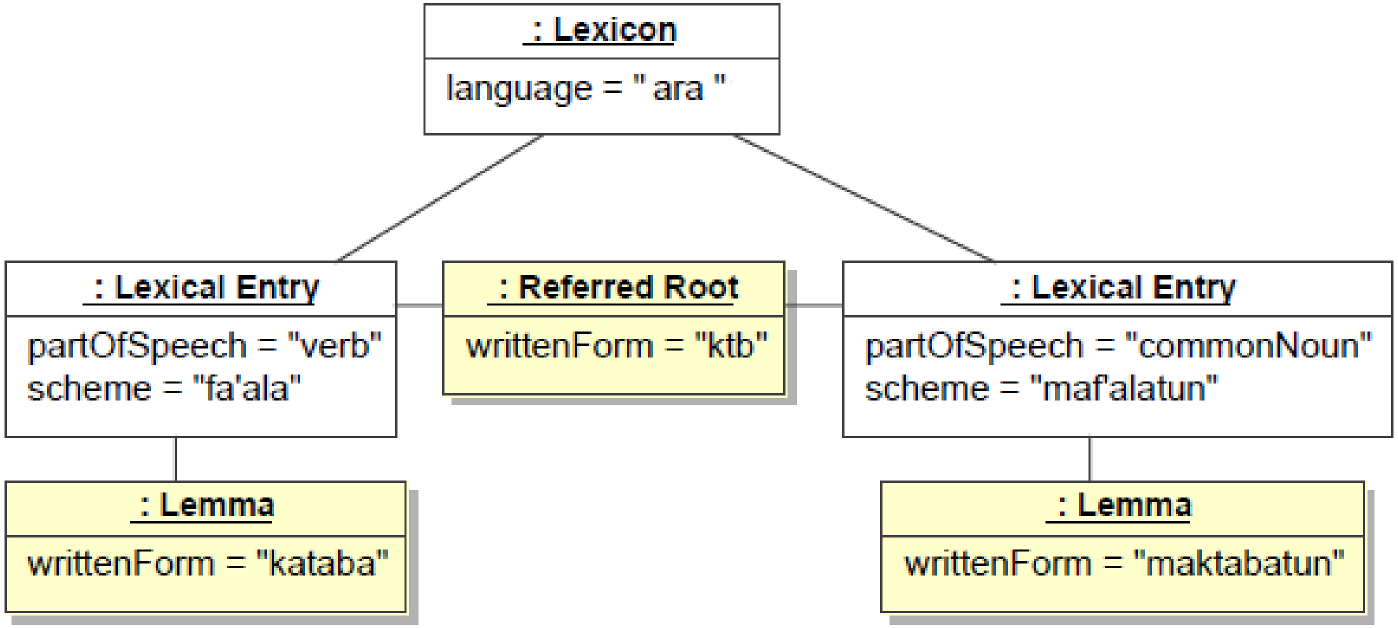
writtenForm = "كتبت"
script = "Arab"
orthographyName = "arabicUnpointed"

: Form Representation

writtenForm = "katabat"
script = "Latn"

Lexical Markup Framework (LMF)

Example: Arabic Root Management



Lexical Markup Framework (LMF)

Sources of Lexical Information

- Morphological information (ATB, Buckwalter, FST):
  - word root, lemma, form, diacritics, frequency, citations. This information will be extracted from the Arabic Treebank
- Syntactic information: Subcategorization frames (Arabic Annotation Algorithm)
- Semantic information: linking to Arabic WordNet.
- Dictionary information: translation in English (Buckwalter, Online Dictionaries, Landau’s Word Count)
- Multi-word Expression (MWE) and named entity. FST and Arabic Named Entity Lexicon extraction project

Automatic Acquisition Methodology

- Methodology
  - Starting with the annotated data to build a core lexicon
  - Moving toward un-annotated data for extension in domain and size
- Lemmatization Tools
  - Buckwalter
  - MADA-TOKAN
  - FST-Guesser

Automatic Acquisition Methodology

- Lemmatization is an essential prerequisite due to
  - derivational and inflectional nature of Arabic
  - lack of diacritics (vowel marks)
  - the employment of cliticization (affixation of function words to content words)
  - 2,552 well-formed forms for transitive verbs (شكر “shakar” (to thank))
  - 519 valid forms for regular nouns (معلم “mu’allim” (teacher))

وسيشكرونه

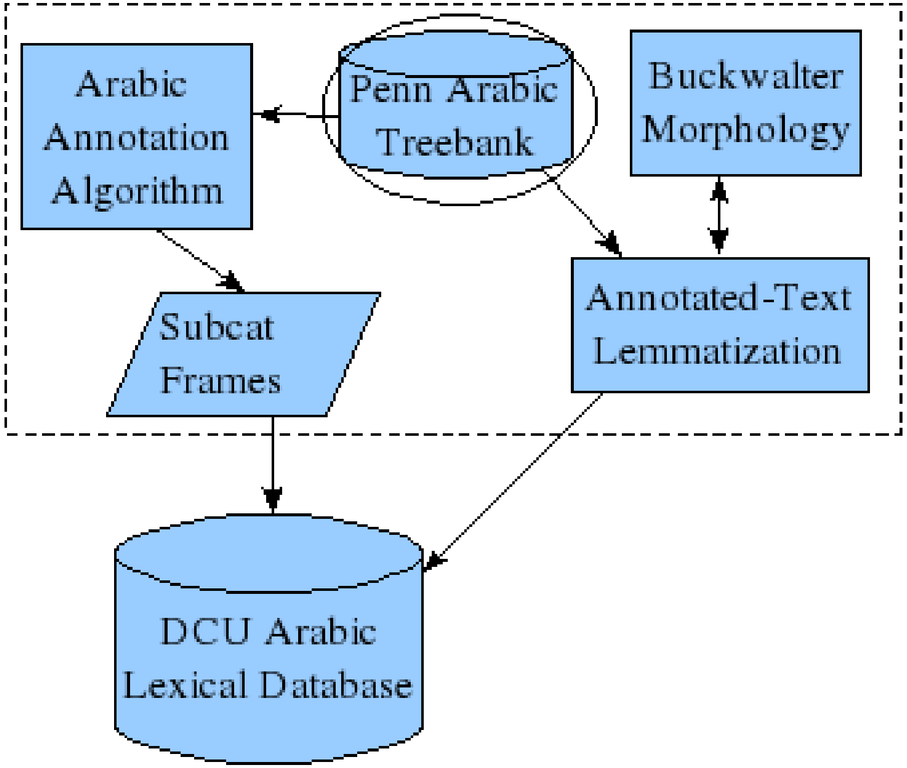
Wa-sa-ya-shkur-una-hu

And-will-thank-they-him

And they will thank him.

Automatic Acquisition

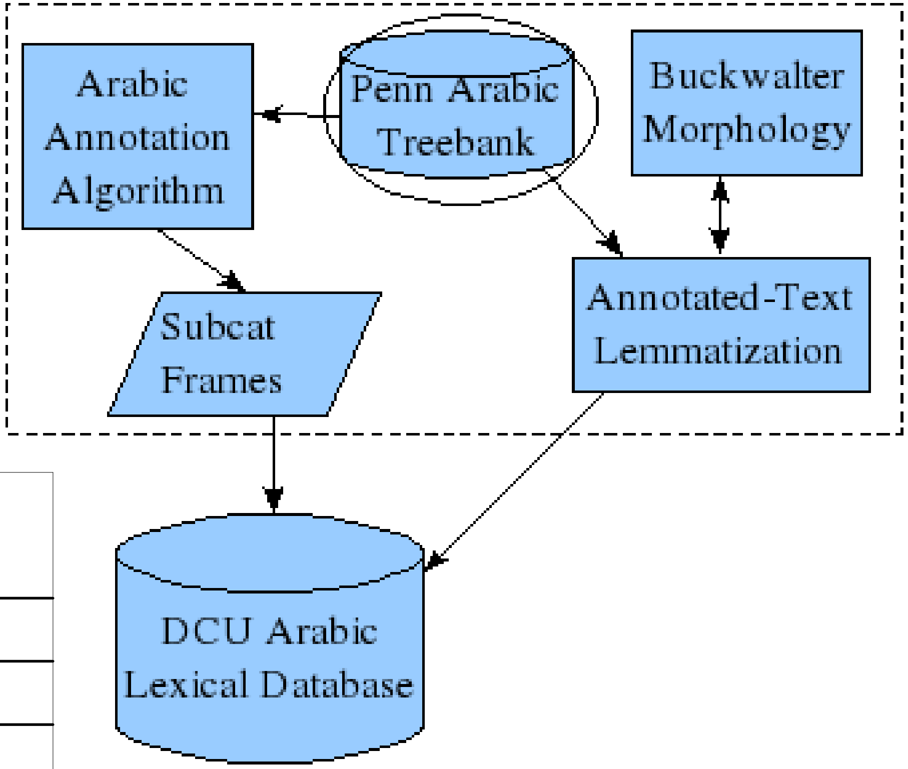
- From annotated data: Arabic Treebank (ATB)
- Advantages
  - morphologically/syntactically annotated
  - modern texts
  - tokenized and diacritized,
  - manually reviewed by human annotators
- Disadvantages
  - Not large: Only ½ million words
  - Not diverse: taken from the newswire


Arabic Annotation Algorithm
Penn Arabic Treebank
Buckwalter Morphology
Subcat Frames
Annotated-Text Lemmatization
DCU Arabic Lexical Database

Automatic Acquisition

- From annotated data: Arabic Treebank (ATB)

Results


Arabic Annotation Algorithm
Penn Arabic Treebank
Buckwalter Morphology
Subcat Frames
Annotated-Text Lemmatization
DCU Arabic Lexical Database

	Types	Unique Lemmas
Nouns	41,183	7,184
Adjectives	14,044	2,540
Verbs	17,888	2,315
Total	73,115	12,039

Testing and Evaluation

From annotated data: Arabic Treebank (ATB)

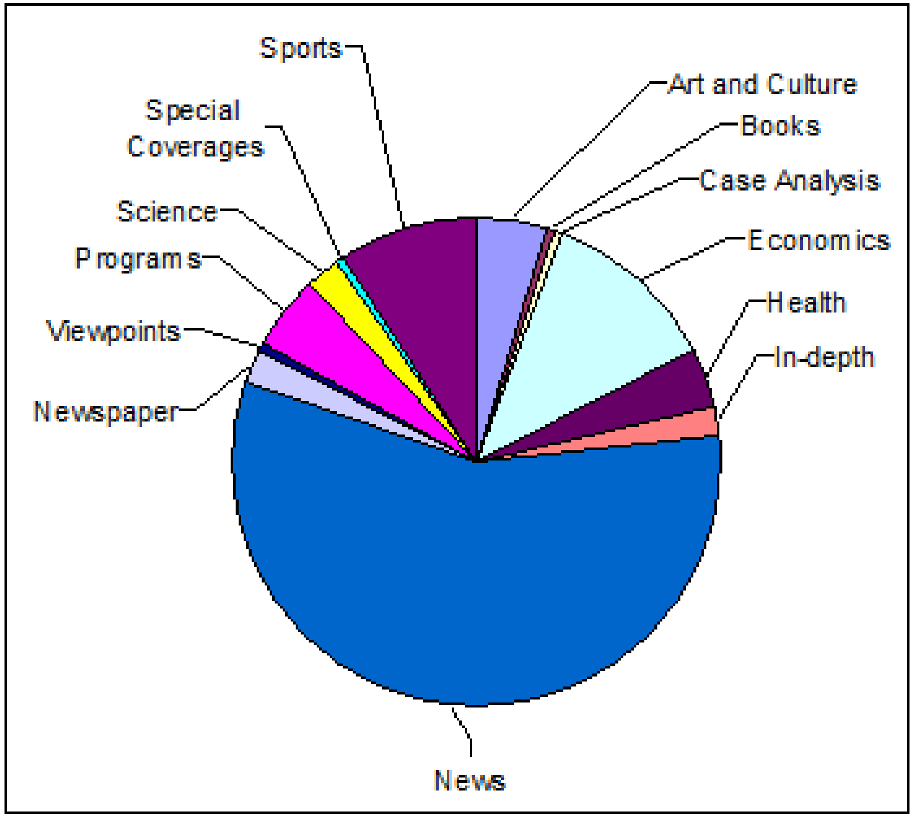
- Testing on Aljazeera Search Engine, Why?
  - The web is polluted with noisy data
  - The type of application (lexicon) has a narrow threshold for noise
  - Technically API does not allow more than 1000 searches per day

Misspellings	Google Score	CNN Score	Right Form	Google Score	CNN Score
arround	1,200,000	3	around	780,000,000	44,555
vedio	4,450,000	0	video	2,590,000,000	131,845
resaercher	6,200	0	researcher	26,500,000	19,729
possebility	31,100	0	possibility	95,100,000	38,163
bilieve	29,200	0	believe	349,000,000	44,330
perfromance	195,000	0	performance	459,000,000	17,085
mesjudge	80	0	misjudge	278,000	196
gtfrde	1,750	0			
ghgh	233,000	0			

Testing and Evaluation

From annotated data: Arabic Treebank (ATB)

- Testing on Aljazeera Search Engine, Why?
  - Aljazeera is more than just news.



Testing and Evaluation

From annotated data: Arabic Treebank (ATB)

- Testing results
- No. of Lemmas from the ATB: 12340
- After removing diacritics: 10071
- No. of Not found on Al-Jazeera 208 (2%)

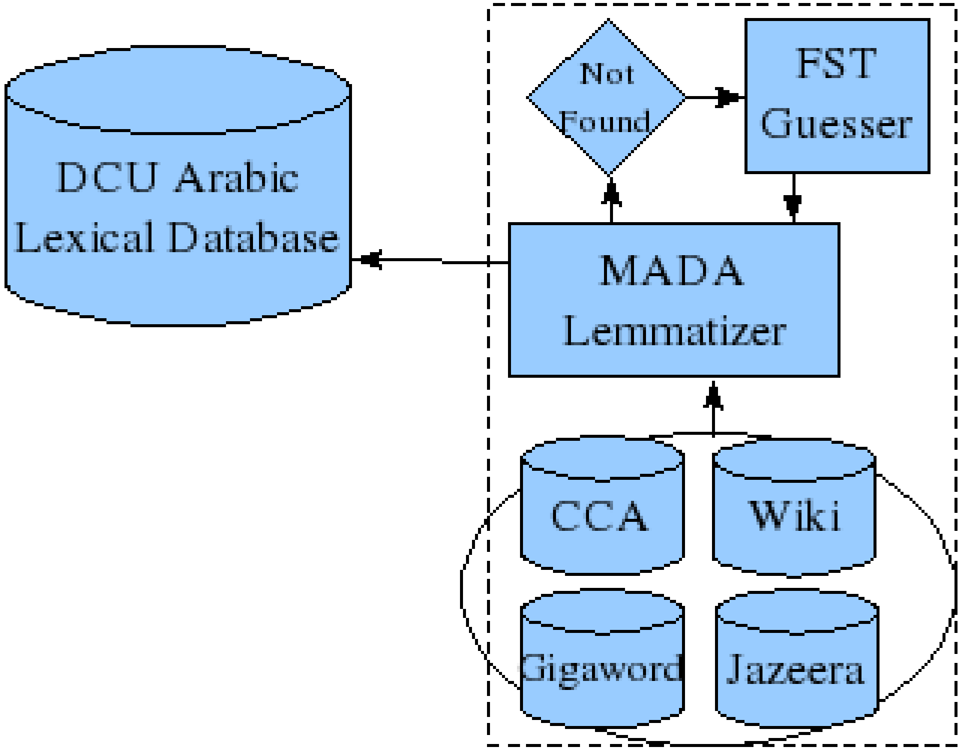
Error analysis

- Mistagging in the annotation process (baloqA' should be bi-liqA')
- Buckwalter gives the wrong lemma it should be >abora$iy~+ap not >abora$iy~
- Not found in Al-Jazeera (>adokan nor dakonA')

Automatic Acquisition Architecture

From un-annotated data

- Advantages
  - Large:
    - CCA: ½ million
    - Wiki: 40 million
    - Gigaword: 200 million
  - Diverse: taken from the various domains
- Disadvantages
  - No morphological or syntactical annotation



Automatic Acquisition Architecture

From un-annotated data

- Results for Corpus of Contemporary Arabic (CCA)

– Nominals:	240236
– Unique nominals:	12502
– Verbals:	67812
– Unique verbals:	4245
– Total:	16747
– Not intersected with ATB:	6312

Automatic Acquisition Architecture

From un-annotated data

- Testing on Aljazeera Search Engine
  - New lexical items: 6312
  - lemmas not found: 543 (9%)
  - full forms: not found: 941 (15%)
  - Neither lemma nor form: 240 (4%)

Automatic Acquisition Architecture

From un-annotated data

- Dealing with the residue (Words not analysed by MADA)
  - 6106 items from the CCA had no analysis by MADA, potentially useful
  - Error Detection (Language dependent)
    - Not including numbers or non-Arabic letters
    - Not including (".*ة.+") taa marbouta in the middle
    - Not including (".*ى.+") Alif maqsoura in the middle in the middle
    - Not including (".*اا.*") two alifs anywhere
  - Error Detection (Language independent)
  - We merge MADA with our FST Guesser to filter through these words

Automatic Acquisition Architecture

MadaOutput: *1.000000 **wAlmtswqyn**=[wAlmtswqyn_0 POS:AJ Al+ +ACC w+ +DEF MOOD:NA +PL]=NO-ANALYSIS

Guesser output:

```
والمتسوقين          +adjوالمتسوق+Guess+dual+acc+gen@                    2
والمتسوقين          +adjوالمتسوق+Guess+masc+pl+acc+gen@                 3
والمتسوقين          +adjوالمتسوقين+Guess+sg@                            1
والمتسوقين          +nounوالمتسوق+Guess+dual+acc+gen@                   0
والمتسوقين          +nounوالمتسوق+Guess+masc+pl+acc+gen@                0
والمتسوقين          +nounوالمتسوقين+Guess+sg@                           0
والمتسوقين          و+conj@ال+defArt@+adjمتسوق+Guess+dual+acc+gen@      6
والمتسوقين          و+conj@ال+defArt@+adjمتسوق+Guess+masc+pl+acc+gen@   7
والمتسوقين          و+conj@ال+defArt@+adjمتسوقين+Guess+sg@              5
والمتسوقين          و+conj@ال+defArt@+nounمتسوق+Guess+dual+acc+gen@     0
والمتسوقين          و+conj@ال+defArt@+nounمتسوق+Guess+masc+pl+acc+gen@  0
والمتسوقين          و+conj@ال+defArt@+nounمتسوقين+Guess+sg@             0
والمتسوقين          و+conj@+adjالمتسوق+Guess+dual+acc+gen@              4
والمتسوقين          و+conj@+adjالمتسوق+Guess+masc+pl+acc+gen@           5
والمتسوقين          و+conj@+adjالمتسوقين+Guess+sg@                      3
والمتسوقين          و+conj@+nounالمتسوق+Guess+dual+acc+gen@             0
والمتسوقين          و+conj@+nounالمتسوق+Guess+masc+pl+acc+gen@          0
والمتسوقين          و+conj@+nounالمتسوقين+Guess+sg@                     0
```

Automatic Acquisition Architecture

- GuessLemma: 7@متسوق@والمتسوقين@AJ
- GuessLemma: 6@متسوق@والمتسوقين@AJ
- GuessLemma: 5@متسوقين@والمتسوقين@AJ
- GuessLemma: 5@المتسوق@والمتسوقين@AJ
- GuessLemma: 4@المتسوق@والمتسوقين@AJ
- GuessLemma: 3@والمتسوق@والمتسوقين@AJ
- GuessLemma: 3@المتسوقين@والمتسوقين@AJ
- GuessLemma: 2@والمتسوق@والمتسوقين@AJ
- GuessLemma: 1@والمتسوقين@والمتسوقين@AJ

Automatic Acquisition Architecture

- Formula for giving weight to the guessing output:

Word Weight =

```
      ((# of different forms          * 2)
    +  (# of form repetitions         * 1))
       / 2
```

006@@متسوق@والمتسوقين#المتسوقين#متسوق

7@3@AJ

((3 * 2) + 7) / 2 = 6

Testing the Formula (How useful in cascading good solutions up the list)

- 57% from the top are valid for inclusion in a dictionary as is
- 6% from the bottom are valid for inclusion in a dictionary as is

Automatic Acquisition Architecture

- Arabic Wikipedia:
- First Portion 2 million words
  - MADA coverage is 96%.
  - 36,000 unique words not found by MADA
  - 22164 verbs, nouns and adjectives are collected from the first portion
  - 10712 were not found in the ATB

        7763 Nominals not found in the ATB

        2949 verbs not found in the ATB

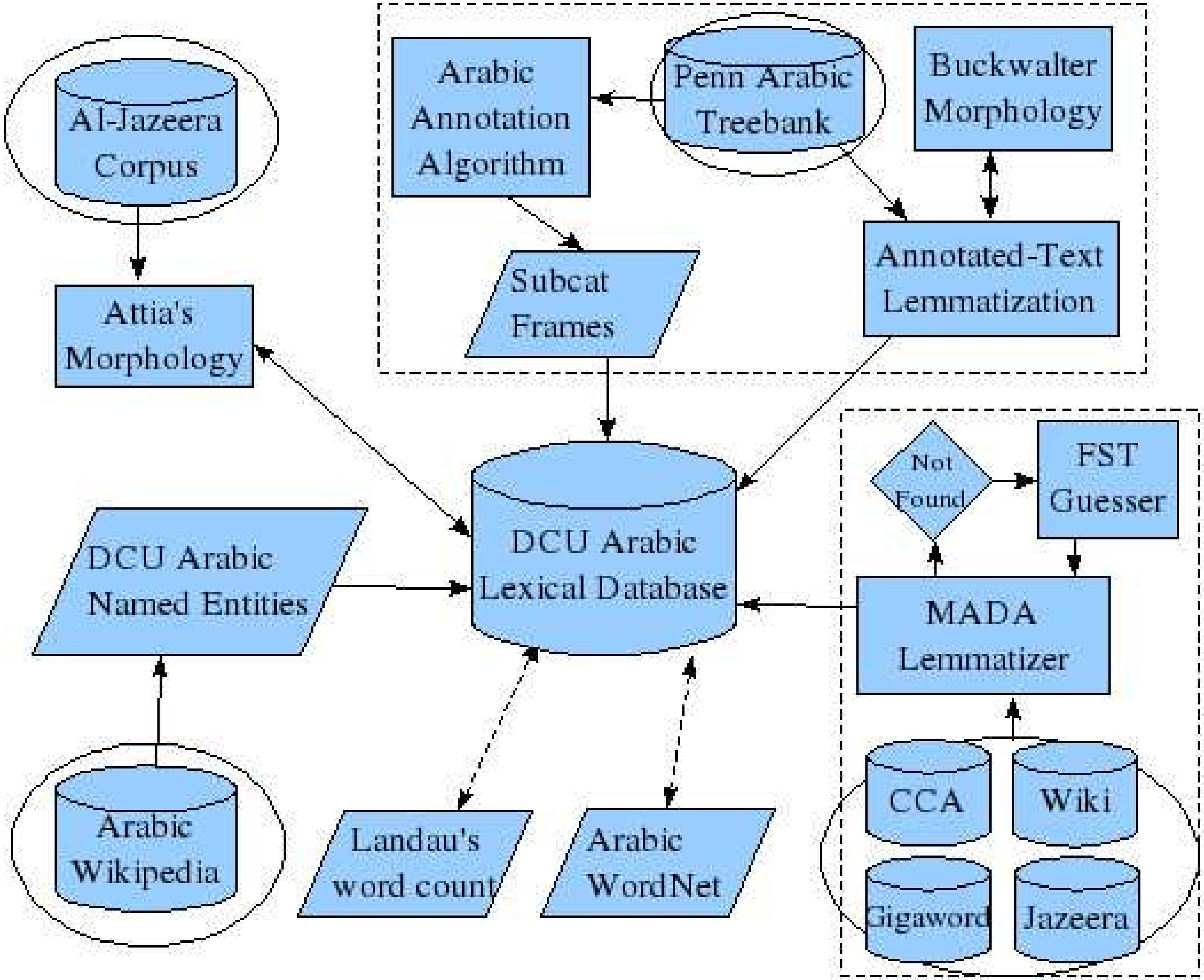
Automatic Acquisition Architecture

From un-annotated data

- Results for Arabic Wikipedia (Full Corpus)

– Nominals:	17076178
– Unique nominals:	22969
– Verbals:	2991974
– Unique verbals:	8151
– Total:	31120
– No Analysis by MADA:	1,724,200

Automatic Acquisition Architecture



Conclusion

- Size of acquired lexicon
  - 12,340 from ATB
    - 1,000 from Attia fst
    - 6,312 from CCA
    - 18,823 from Wiki
- Buckwalter contains irrelevant lexical entries (at least 1/5 is outside of MSA)
  - 23% of the lemmas are not found in Al-Jazeera
  - 20% of the lemmas are not found in Arabic Wiki (40 million words)
- Significance of frequency information to calculate word weight as a method of validation
- Using search engine as a way to flag potentially problematic entries
- Improving error detection techniques