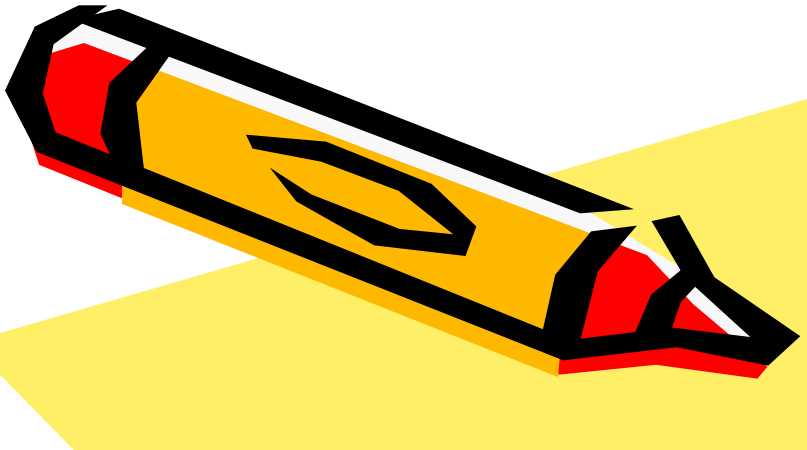


Lexicons, Corpora & Morphology

Yousef S. I. Elarian


بسم الله الرحمن الرحيم

Outline

- Lexicons
  - Lexicon
  - Lexicon Extraction
- Corpora
  - Corpus
  - Evaluation (Zipf's Law).
- Morphology
  - Arabic Morphology
    - Templative
      - Roots, Patterns, Stems
    - Concatenative
    - Concatenative *vs.* Templative
- Practical Stuff
  - Xerox
    - Buckwalters

Lexicon

- Restricted vocabulary of a(NLP) system
  - A list of all expected or allowed valid words.
- backbone of any NLP application.
- Generated:
  - Manually (many people)
  - With Computers (Today's trend)
    - Extract from corpora
    - Reduce (Stem)
    - Synthesized??
- Examples:
  - Bare
  - With description

; conjunctions

و وَ Pref-Wa and
<pos>wa/CONJ+</pos>

ف فَ Pref-Wa and;so
<pos>fa/CONJ+</pos>

One
Two
Three
Four
Five
Here
Mars
Days
Name
Go

Lexicon Extraction

- Computational-linguistic community is converging to extract the lexicon from naturally used text (newspaper, phone call).
- A large amount of representative text is gathered and processed (*Corpus*).
- Typically involves normalizing surface-words into a common basic form (e.g. roots or stems)
  - Reduce the number of entries.
  - Need Morphology!

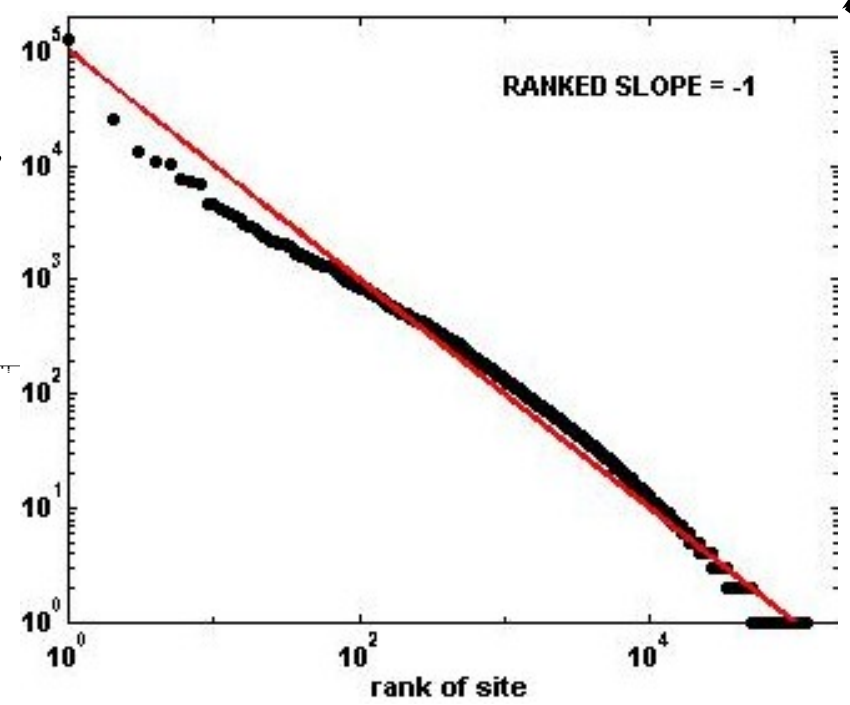
Lexicon Extraction from Corpora

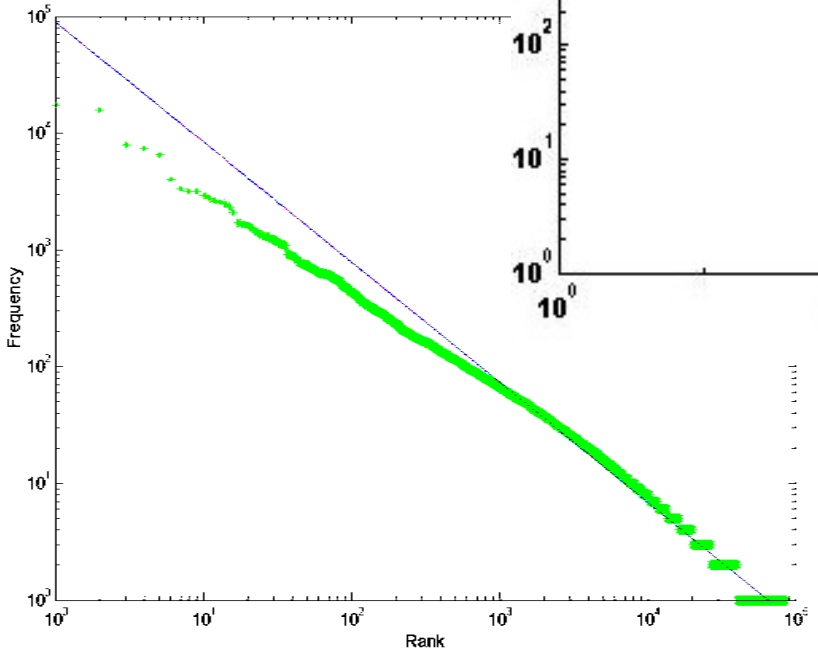
- Corpus:
  - *pl. corpuses or corpora*
  - A very large amount of NL representative text.
  - Typically (but not exclusively) from newspapers.
- Pros:
  - Capture the frequencies of NL. (Utterances.)
- Cons:
  - Never complete.
  - Typos.
- Example
  - CCA

Evaluating Corpora
Zipf's law

- Empirical law
  - Measures corpus quality
  - Theory

- $f \times r = k$.





- log-log plot

Morphology

Morphology

علم الصرف

- The (grammatical) study of the (internal) structure of words.
- A **morpheme** is defined as the minimal meaningful unit of a language.
- Types (by August Schleicher)
  - Analytic (Isolating)
  - Concatenative (Agglutinative)
    - Prefix — informal — سيذهب
    - Suffix — formalize — ذهبوا
    - Circumfix — informalize — يذهبان
  - Templatic (Fusional)
    - Root — mouse — ذهب
    - Pattern (infix). mice — ذاهب

Chinese

English

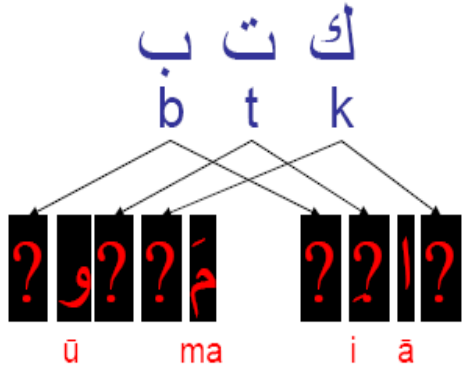
Arabic

Turkish

Templatic Morphology

- Starts from Roots & Patterns
- Examples:

- Root



- Pattern
- Lexeme

كاتب
kātib

مكتوب
maktūb

	Pattern	Pattern Meaning	Example	Gloss
I	1a2a3	Basic sense of root	ktb → katab	write
II	1a22a3	Intensification, causation	ktb → kattab	dictate
III	1aA2a3	Interaction with others	ktb → kaAtab	correspond with
IV	Aa12a3	Causation	jls → Ajlas	seat
V	ta1a22a3	Reflexive of Pattern II	Elm → taEal~am	learn
VI	ta1aA2a3	Reflexive of Pattern III	ktb → takaAtab	correspond
VII	Ain1a2a3	Passive of Pattern I	ktb → Ainkatab	subscribe/enroll
VIII	Ai1ta2a3	Acquiescence, exaggeration	ktb → Aiktatab	register
IX	Ai12a33	Transformation	Hmr → AiHmarr	Turn red/blush
X	Aista12a3	Requirement	ktb → Aistaktab	ask/make_write

Templatic Morphology

Roots

- Primary lexical unit of a word
  - Carries semantic content.
  - Cannot be reduced.
  - Left when all, including internal, morphologically added structure has been wrung out.
- In Arabic:
  - An ordered sequence of 3, 4, or 5 letters.
  - bare verb.

Templatic Morphology

Patterns

- AKA measures or forms.
- Inflectional morphemes
  - Non(purely)-concatenative.
- General moulds.
- A sequence of constant and variable characters.
  - Variable characters: (ف، ع، ل) = (1, 2, 3).
    - To be substituted by the letters of the Arabic root.

كتب كاتب

Concatenative Morphology

Concatenative Morphology

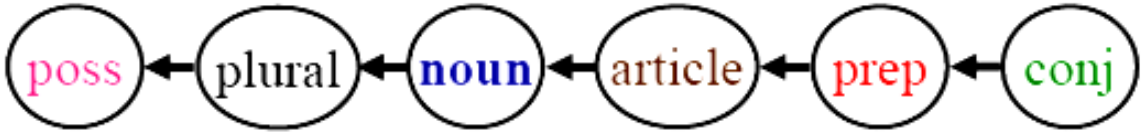
- Starts from stems.
- Minimal surface-form
  - Nouns, verbs, & Particles.
  - But not all surface-words are stems.
  - Roots + Patterns

`Root.GeneralMeaning + Patten.specificMeaning`

- Only further *Circum*fixation allowed
  - No further *in*fixation.

Concatenative Morphology

- Noun Examples



وكبيوتنا	وللمكتبات
/wakabiyūtinā/	/walilmaktabāt/
و+ ك + بيوت + نا	و+ل+ال+مكتبة+ات
wa+ka+biyūt+nā	wa+li+al+maktaba+āt
and+like+houses+our	and+for+the+library+plural
And like our houses	*And for the libraries*

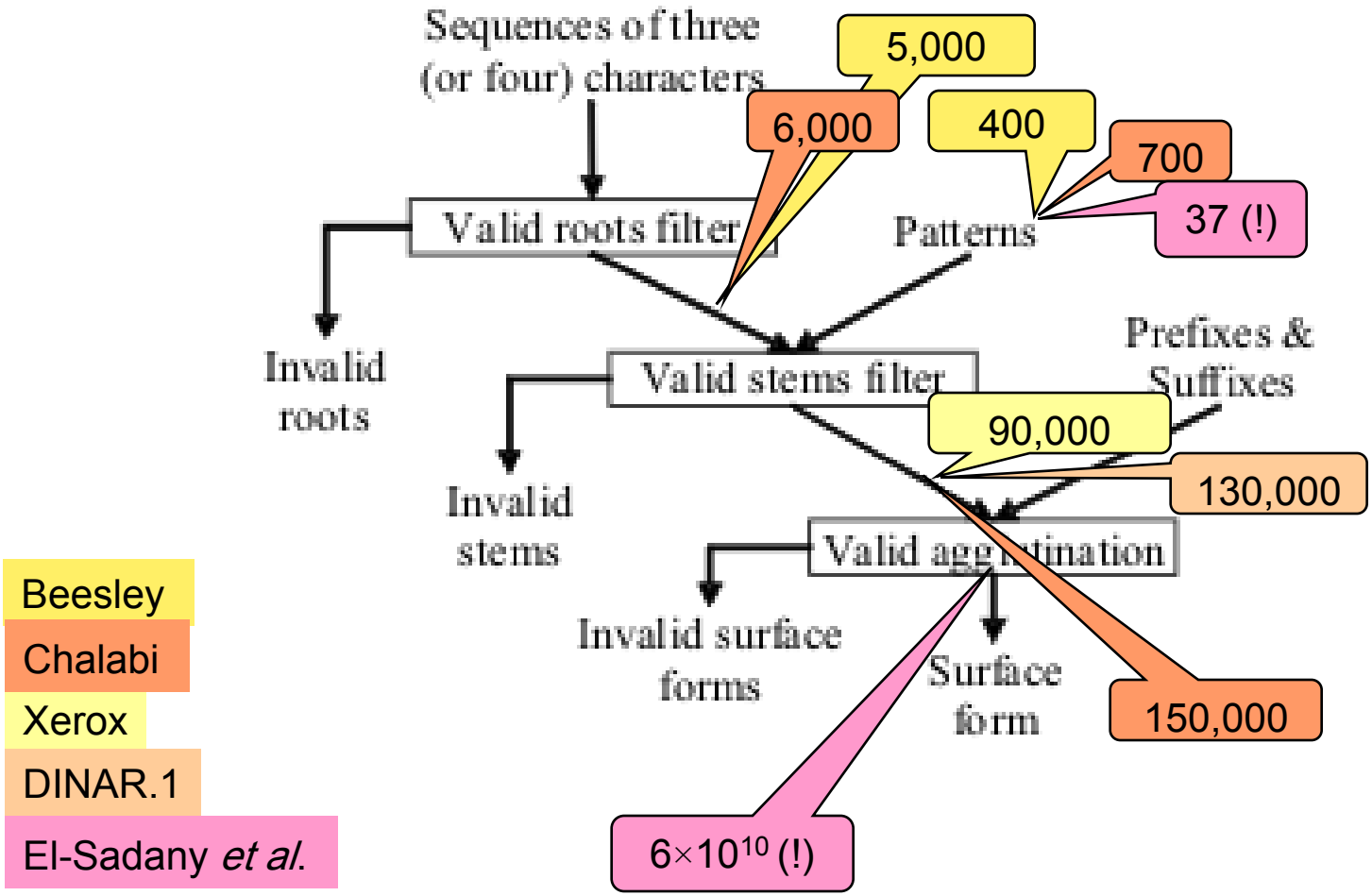
Concatenative Morphology (Cont.)

- Verb Examples

وسنقولها	فقلناها
/wasanaqūluhā/	/faqulnāhā/
و+ س+ ن+ قول + ها	ف+ قال+ نا+ ها
wa+sa+na+qūl+u+hā	fa+qul+na+hā
and+will+we+say+it	so+said+we+it
And we will say it	*So we said it.*

Arabic Morphology

- Statistics



Templatic vs. Concatenative

Concatenative *vs.* Templatic

Unit	stem	Root & Pattern
# entries in lexicon	# Roots * # templates (>80K)	# Roots (5 – 6 K)
Cons.	Size	Computations
Abstract-ness	Standalone Word	Pure Semantic field
Use (Xerox)	Buckwalter ++	Beesley --

Practical Stuff

Xerox

- Worked on both Morph. Anal:
  - Internal (Root-pattern)
    - Beesley
    - ((Roots + patterns) + circumfixes)
    - 5,000 Root ∞ 400 pattern
  - External (Stems)
    - Buckwalter
    - (Stems + circumfixes)
    - Starts from over 80,000 stems
- Has a very popular *transliteration* system.
  - Named after Buckwalter.
  - A semi-standard now.

ء '	ذ *	ل l
آ ا	ر r	م m
أ >	ز z	ن n
ؤ &	س s	ه h
إ <	ش S	و w
ئ }	ص S	ى Y
ا A	ض D	ي y
ب b	ط T	ً F
ة p	ظ Z	ٌ N
ت t	ع E	ٍ K
ث v	غ g	َ a
ج j	ـ _	ُ u
ح H	ف f	ِ i
خ x	ق q	ّ ~
د d	ك k	ْ o

Buckwalter's AraMorph as an example

Goals

- Morphotactics & morphophonemic rules built in the lexicon
  - A single lexicon of prefixes/suffixes including all valid concatenations
  - Orthographic variations = additional dictionary entries.
- Lexical tagging
  - Stems rather than root and patterns.

Buckwalter's AraMorph

- Availabe for free
  - Original Perl version
  - Java version (Pierrick Brihaye)
- morphotactics and orthographic rules built-in (in lexicons).
  - E.g. contains: ل، ال، لل
- 3 Morpheme Lexicons
  - Stems, prefixes, suffixes.
- 3 Compatibility tables:
  - specify allowed concatenations
    - Prefix-Stem
    - Stem-Suffix
    - Prefix-Suffix

Buckwalter's Files

- Abstract

File kind	*File name*	*Entries*	*Forms*	*Size KB*
Lexicons	dictPrefixes	78	299	26.7
	dictStems	38,600	82,158	3,727.36
	dictSuffixes	206	618	89.5
Compatibil-ity tables	tableAB	1,648	1,648	54.9
	tableBC	1,285	1,285	15.3
	tableAC	598	598	27.7

Buckwalter's Files

- Sample from "dictPrefixes"

w	wa	Pref-Wa	and <pos>wa/CONJ+</pos>
f	fa	Pref-Wa	and;so <pos>fa/CONJ+</pos>
b	bi	NPref-Bi	by;with <pos>bi/PREP+</pos>
k	ka	NPref-Bi	like;such as <pos>ka/PREP+</pos>
wb	wabi	NPref-Bi	and + by/with <pos>wa/CONJ+bi/PREP+</pos>
fb	fabi	NPref-Bi	and + by/with <pos>fa/CONJ+bi/PREP+</pos>
wk	waka	NPref-Bi	and + like/such as <pos>wa/CONJ+ka/PREP+</pos>
fk	faka	NPref-Bi	and + like/such as <pos>fa/CONJ+ka/PREP+</pos>
Al	Al	NPref-Al	the <pos>Al/DET+</pos>

Buckwalter's Files

- Sample from "dictStems"

```
;--- ktb
;; katab-u_1
ktb         katab      PV          write
ktb         kotub      IV          write
ktb         kutib      PV_Pass     be written;be fated;be destined
ktb         kotab      IV_Pass_yu          be written;be fated;be destined
;; kAtab_1
kAtb        kAtab      PV          correspond with
kAtb        kAtib      IV_yu       correspond with
;; >akotab_1
>ktb        >akotab    PV          dictate;make write
Aktb        >akotab    PV          dictate;make write
ktb         kotib      IV_yu       dictate;make write
ktb         kotab      IV_Pass_yu          be dictated
;; kitAboxAnap_1ktAbxAn
kitAboxAn           NapAt      library;bookstore
```

Buckwalter's Files

- Sample from "dictSuffixes"

```
p          ap          NSuff-ap   [fem.sg.]      <pos>+ap/NSUFF_FEM_SG</pos>
ty         atayo       NSuff-tay  two
<pos>+atayo/NSUFF_FEM_DU_ACCGEN_POSS</pos>
tyh        atayohi     NSuff-tay  his/its two
<pos>+atayo/NSUFF_FEM_DU_ACCGEN_POSS+hu/POSS_PRON_3MS</pos>
tyhmA      atayohimA   NSuff-tay  their two
<pos>+atayo/NSUFF_FEM_DU_ACCGEN_POSS+humA/POSS_PRON_3D</pos>
tyhm       atayohim    NSuff-tay  their two
<pos>+atayo/NSUFF_FEM_DU_ACCGEN_POSS+hum/POSS_PRON_3MP</pos>
tyhA       atayohA     NSuff-tay  its/their/her two
<pos>+atayo/NSUFF_FEM_DU_ACCGEN_POSS+hA/POSS_PRON_3FS</pos>
tyhn       atayohin~a  NSuff-tay  their two
<pos>+atayo/NSUFF_FEM_DU_ACCGEN_POSS+hun~a/POSS_PRON_3FP</pos>
```

Buckwalter's Files

- Sample from:

"TableAB" "TableAC" and "TableBC"

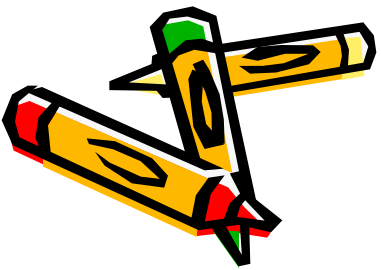
NPref-Al N	NPref-Al Suff-0	PV PVSuff-a
NPref-Al N-ap	NPref-Al NSuff-u	PV PVSuff-ah
NPref-Al N-ap_L	NPref-Al NSuff-a	PV PVSuff-A
NPref-Al N/At	NPref-Al NSuff-i	PV PVSuff-Ah
NPref-Al N/At_L	NPref-Al NSuff-An	PV PVSuff-at
NPref-Al N/ap	NPref-Al NSuff-ayn	PV PVSuff-ath

1st Step: existance

- Arabic dictionary look-up consists of asking, for each segmentation:
  - does the prefix exist in the lexicon of prefixes?
    - if so, does the stem exist in the lexicon of stem?
      - if so, does the suffix exist in the lexicon of suffixes

2nd Step: Compatibility

- If all three word elements (prefix, stem, suffix) are found, ask:
  - is the morphological category of the prefix compatible with the morphological category of the stem?
    - if so, is the morphological category of the prefix compatible with the morphological category of the suffix?
      - if so, is the morphological category of the stem compatible with the morphological category of the suffix?



Links:

- http://www.qamus.org/morphology.htm
- http://students.cs.byu.edu/~jonsafar/cgi-bin/aramorph_fast.cgi
- http://www.AraMorph.nongnu.org

Finally

- A Program run.
- Questions?
- Main References
  - Elarian YS. Lexicon Generation for Arabic Optical Text Recognition [dissertation]. Jordanian University of Science and Technology; 2006, August.
  - Habash N. Introduction to Arabic natural language processing. ACL'05 Tutorial; 2005 June 25; Ann Arbor, USA.
  - Wikipedia, the free encyclopedia. [Online] [Accessed 2006 December]. Available from URL: http://en.wikipedia.org/wiki/.

- Thanks.