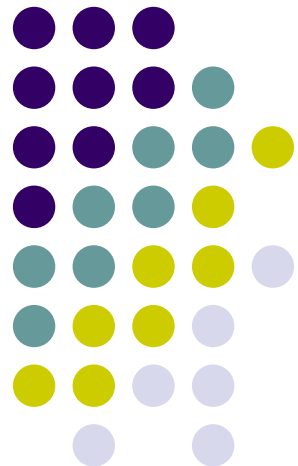
Exploring Syllables, Romanization, and Analogy in Names

Deryle Lonsdale

BYU Linguistics

lonz@byu.edu



Proper nouns and analogy

- Proper nouns are interesting linguistically
  - Phonology: sound sequences, syllable structure
  - Orthography: how writing systems do(n't) reflect sounds
  - Semantics: meaning, denotation
  - Pragmatics: culture, religion, history
  - Translation: crosslinguistic issues
- Analogy, a general cognitive strategy, can help in explaining many of these phenomena

Arabic script

- Arabic is a Semitic language
- Arabic script is also used for other languages, including non-Semitic ones
  - Urdu: Pakistan (Indo-Aryan)
  - Persian/Farsi: Iran (Indo-Iranian)
  - Pashto: Afghanistan (Indo-Iranian)
- It's an (impure) abjad
  - Abjad: alphabet but (some) symbols missing
  - No short vowels, though long ones are usually represented

Names in Arabic script

- Written right-to-left
- No capital letters
- Vocalization: add missing short vowels
- Romanization: converting words to Roman script languages (e.g. English)

أبومصعب الزرقاوي

Abu M(u)sab al-Z(a)rqawi

محمود احمدی‌نژاد

M(a)hmoud Ahm(a)din(e)jad

Common techniques used

- Lexicographic: dictionary lookup
- Bitext mining: previous translations
- Text-to-speech phonemicization
  - Usually transduction via finite-state methods
- Machine learning
  - Statistical/stochastic approaches (e.g. n-grams)
  - Entropy/noisy channel approaches
  - Rule-based transformational approaches
  - Exemplar-based approaches

Analogical modeling

- Exemplar-based machine learning approach
- Analogy is the basic operation
- Useful for modeling natural language phenomena
  - Particularly low-level issues: phonology, orthography, morphology
- No explicit rules, just store of vectorized exemplar data
- Flexible input, output, reporting, metrics

The task(s)

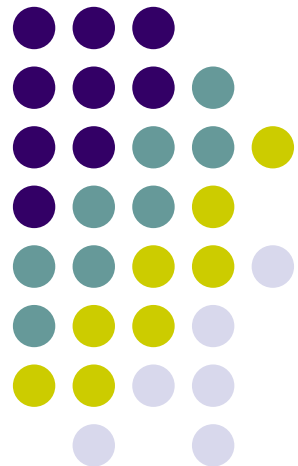
- Process Farsi names (Arabic script):
  1) Arabic script → vocalized Arabic script
  2) Arabic script → vocalized romanization
- 23,000 items with three types of proper noun information (given name(s), last name(s), location)
  - Arabic script and one romanization

Sample data

- هپکو | غلامعباس | ابراهیمی عراق نژاد
  hepko | Ghulam Abaas | Ebrahimi Iraq Nezhad
- خانه سازی قنات | ناصر | ابراهیمی عراقی
  Khanah Saazi Qnaat | Naser | Ebrahimi Iraqi
- شهید شیرودی | غلامرضا | ابراهیمی عراقی نژاد
  Shaheed sherodi | Ghulam Reza | Ebrahimi Iraqi Nezhad
- شهر صنعتی | عبدالعباس | آل بوسویلم
  Shaher Sunhati | Abdul Abaas | Aal Busuylam
- جهانگیری | محمد | آلبوغبیش
  Jahangeeri | Mohammad | Aalbughabish
- شهید رجائی | مسعود | آل بیگی غلامی
  Shaheed Rijahee | Masood | Aal Baigi Ghulami

Task 1

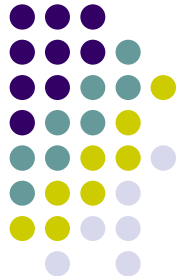
Provide Arabic-script vocalization



Issues in vocalization

- Variable placement: metathesis-like
  - Ahm(a)di / Ah(a)mdi
- Diphthongs and glides are problematic
  - Baizaa hee / Baizayee
  - Ahsaanian / Ahsaaneean
- Nasalization
- Vowels (short & long) are notoriously variable in English (ghoti, ghoughpteighbteau)
  - Imami / Imaami

Step 1: Transliterate



kukb+slTAn	Kowkab+Sultan
zhrA	Zahra
jmilh	Jamila
}biH+Alh	Zabeeulah
}biH+A...	Zabee+A&
Sdiqh	Sideeqa
Dmir	Zameer
ESmt	Esmat
ElirDA	Ali+Reza
GlAmEli	Ghulam+Ali
mHmd+Hsin	Mohmmad+Hussian
mHmd+Eli	Mohmmad+Ali

Step 2: Capture pairings

- Wrote finite-state automaton to capture correspondences between Arabic / romanization
- Sliding window across names, 1 character at a time
  - Prefer 1-1 mappings, but allow for others
- Result: training vectors with 31 orthographic features
  - Outcomes are 0-3 character realizations

Sample vectors

```
H  , = = = = = = = = = = = = = = = H A j + m H m d + x A n i = = =
A  , = = = = = = = = = = = = = = H A j + m H m d + x A n i = = = =
j  , = = = = = = = = = = = = = H A j + m H m d + x A n i = = = = =
+  , = = = = = = = = = = = = H A j + m H m d + x A n i = = = = = =
m  , = = = = = = = = = = = H A j + m H m d + x A n i = = = = = = =
oH , = = = = = = = = = = H A j + m H m d + x A n i = = = = = = = =
am , = = = = = = = = = H A j + m H m d + x A n i = = = = = = = = =
ad , = = = = = = = = H A j + m H m d + x A n i = = = = = = = = = =
+  , = = = = = = = H A j + m H m d + x A n i = = = = = = = = = = =
x  , = = = = = = H A j + m H m d + x A n i = = = = = = = = = = = =
A  , = = = = = H A j + m H m d + x A n i = = = = = = = = = = = = =
n  , = = = = H A j + m H m d + x A n i = = = = = = = = = = = = = =
i  , = = = H A j + m H m d + x A n i = = = = = = = = = = = = = = =
```

Sample generated outputs

خرمن+بیز

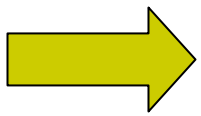
78.55 xorami+biz

77.72 xrami+biz

76.69 xarami+biz

76.52 xoraman+biz

75.69 xrman+biz



78.55 خُرَمی+بیز

77.72 خرَمی+بیز

76.69 خَرَمی+بیز

76.52 خُرَمَن+بیز

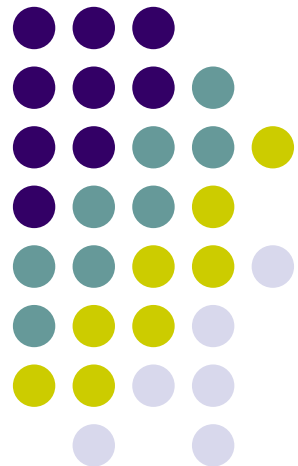
75.69 خرمَن+بیز

Sample vocalized output

صغرى
75.00 صَغَرى
71.43 صَغَرى
64.29 صوغرى
64.29 صَغَرَى
60.71 صوغَرى
60.71 صَغَرَى
53.57 صوغرَى
50.00 صوغَرَى

Task 2

Provide vocalized romanization



Issues in romanization

- Arabic sounds do not always map to English symbols
- Not just one-to-one correspondence
- Divine name often elided
  - آیت ا...غفاری Ayatullah Ghafari
- Syllable boundaries are unclear
  - Ambisyllabicity, consonant gemination
- Word boundaries are not consistent

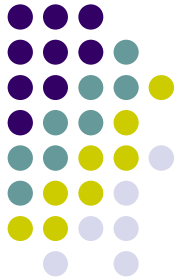
Process: as for vocalization

- Transliterate
- Transduce to produce instance vectors
  - 31 orthographic features
- Outcomes are letter sequences, generally more complicated
  - Perform vocalization and romanization at once

Sample vectors

```
B   , = = = = = = = = = = = = = = = b d x C A n = = = = = = = = = ,
ad  , = = = = = = = = = = = = = = = b d x C A n = = = = = = = = = ,
akh , = = = = = = = = = = = = = = = b d x C A n = = = = = = = = = ,
sh  , = = = = = = = = = = = = = = = b d x C A n = = = = = = = = = ,
a   , = = = = = = = = = = = = = = = b d x C A n = = = = = = = = = ,
n   , = = = = = = = = = = = = = = = b d x C A n = = = = = = = = = ,
B   , = = = = = = = = = = = = = = = = b d x C A n i = = = = = = = = ,
ad  , = = = = = = = = = = = = = = = = b d x C A n i = = = = = = = = ,
akh , = = = = = = = = = = = = = = = = b d x C A n i = = = = = = = = ,
sh  , = = = = = = = = = = = = = = = = b d x C A n i = = = = = = = = ,
a   , = = = = = = = = = = = = = = = = b d x C A n i = = = = = = = = ,
n   , = = = = = = = = = = = = = = = = b d x C A n i = = = = = = = = ,
i   , = = = = = = = = = = = = = = = = b d x C A n i = = = = = = = = ,
B   , = = = = = = = = = = = = = = = = b E A j + z A d h = = = = = = = ,
E   , = = = = = = = = = = = = = = = = b E A j + z A d h = = = = = = = ,
haa , = = = = = = = = = = = = = = = = b E A j + z A d h = = = = = = = ,
j   , = = = = = = = = = = = = = = = = b E A j + z A d h = = = = = = = ,
+   , = = = = = = = = = = = = = = = = b E A j + z A d h = = = = = = = ,
Z   , = = = = = = = = = = = = = = = = b E A j + z A d h = = = = = = = ,
a   , = = = = = = = = = = = = = = = = b E A j + z A d h = = = = = = = ,
d   , = = = = = = = = = = = = = = = = b E A j + z A d h = = = = = = = ,
h   , = = = = = = = = = = = = = = = = b E A j + z A d h = = = = = = = ,
```

Sample raw output



```
::::::::::::::
]it+_...bhbhAni
::::::::::::::
91.11 Ayat+Allah+Bahbahaani
91.11 Ayat+Allah+Bahbahani
88.89 Ayat+Allah+Bahbahanee
88.89 Ayat+Allah+Bahbahaanee
88.89 Aayat+Allah+Bahbahaani
88.89 Aayat+Allah+Bahbahani
88.89 Aayat+Allah+Bahbahaani
88.89 Ayat+Allah+Bahbahaanee
86.67 Aayat+Allah+Bahbahaanee
86.67 Aayat+Allah+Bahbahanee
86.67 Aayat+Allah+BahbahAnee
```

Sample output

حافظ
450.000000 Hafizee
450.000000 Hafeezee

جمشید
399.414000 Jamsheed
396.716000 Jamshid
394.940000 Jamshaid
384.322000 Jamasheed

شاهپور
450.164000 Shaahpur
395.169000 Shaah+Pur

بهنام
436.044000 Bahnaam
402.424000 Behnaam

Syllabification is an issue

- Even in English
  - Merriam Webster: *si.lly*, *ho.llow*, *ba.lance*
    Cambridge: *sill.y*, *ho.llow* or *holl.ow*, *bal.ance*
- People vary in their perceptions, practices
- This has implications for doubled consonants (ambisyllabicity)
- Frequently observed in the data
  - Hessari / Hesaari
- Syllable boundary in vectors would help

Performance and evaluation

- Why not simply transduce?
  - Only one possible realization provided; many are possible and desirable to identify
  - Generate all possible realizations, with scores
- Rote recall of forms provided
- Analogy applied to generate, score, rank alternative possibilities
- Human evaluation of alternatives necessary

Conclusions

- Interesting issues in Arabic-script name processing
- Widely varying practices in romanization of names
- Analogy (and AM) provide good account
- Techniques can be used for other languages (source and target) if training data available