Building a Persian Frequency Dictionary
Corey Miller (MITRE) & Karineh Aghajanian-Stewart
May 24, 2018

The author’s affiliation with The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE’s concurrence with, or support for, the positions, opinions or viewpoints expressed by the author.
Outline

• Motivation for a frequency dictionary of Persian
• Sample entries
• Corpus development
• Text normalization
• Morphological analysis
• Thematic vocabulary
• Lessons learned and future plans
What is a frequency dictionary?

• Instead of ordering words alphabetically, we order them by frequency (most frequent first)
  • How is frequency determined?
    • We collect a representative corpus (assemblage) of documents in a language and count* how many times each word* appears

• Why build a frequency dictionary?
  • Learners and teachers want to know what words are most useful to know
    • By “front-loading” the most frequent words in a curriculum, students gain the most access to the most foreign language content
<table>
<thead>
<tr>
<th>Word</th>
<th>Meaning</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>رفتان</td>
<td>go adj,n,v</td>
<td>128</td>
</tr>
<tr>
<td>—— با آن پسر رفت بیرون چند مرتبه</td>
<td>She went out with that guy several times</td>
<td>71</td>
</tr>
</tbody>
</table>
Our corpus and some definitions

• We distinguish **types** from **tokens**
  • The above sentence contains 5 tokens and 5 types
  • This sentence **has** 10 tokens and it **has** 9 types (**has** is 1 type)
How do we count?

• We can reduce the items to be counted further by using word families (Bauer & Nation 1993)
  • The intuition is that certain words belong to the same families, and therefore should be counted as one
    • Register/dialect/sociolinguistic/spelling variants of a given word (Text Normalization): e.g. night/nite, honor/honour, goin’/going
    • Morphological variants of a given word (Morphological Analysis): e.g. book/books, go/goes/going/went/gone
Text Normalization in Persian

• Spelling variation
  • Hamze followed by yeh, or two yehs
    • امريكائي، امريکائي (‘American’)
  • Persian or Arabic Unicode codepoints
    • ی or ك
  • Bound/unbound/half-space (zero-width non-joiner, ZWNJ)
    • ‘library’
• Register variation (Standard/Colloquial)
  • ‘house’، ‘they are doing’
• Recomposition of multiwords (e.g. English ice cream): وارد شهر شدم ‘she entered the city’
• Decomposition of multiwords: (cf. English isn’t): اينجاست، ‘generally’، ‘it’s here’
Morphological Analysis in Persian

- Iteratively developed **pipeline** consisting of:
  - Override: “Put this word in this family”
  - Morphological Analyzer: part of speech lexica constrain inflection
  - Stemmer: all operations attempted
    • E.g. ‘Azerbaijan’ analyzed as ان + آذرباییان

- Ambiguous forms
  • کتابی is either /ˈkeːtɒbi/ ‘a book’ or /ˈketɒbi/ ‘bookish’
  • Original plan was to consequently merge Noun and Adjective word families, but this would have caused us to eliminate important words like ماهی ‘fish’ and صورتی ‘pink’, so we backed off
    • In the future, we would like to use Word Sense Disambiguation to gain more accurate counts of ambiguous forms
Additional book features

- Alphabetic and part-of-speech indices
- Thematic vocabulary tables
  - Body, Clothing, Colors, Countries
  - Electronics, Emotion, Family
  - Female names, Male names, Last names
  - Food, Health, Materials, Nationalities
  - Nature, Politics, Professions, Religion
  - Sports, Time, Days
  - Transport, War, Weather
  - Islamic/French/Persian/Dari months
  - Light verb constructions, Simple verbs
Lessons Learned & Future work

• Publishers don’t necessarily provide foreign language copy-editing
  • DIY
• Publishers don’t necessarily have foreign language typesetting expertise
  • Punctuation problems
• Audio (TTS or Real)
• Word Sense Disambiguation
  • شیر ‘lion’ or ‘milk’