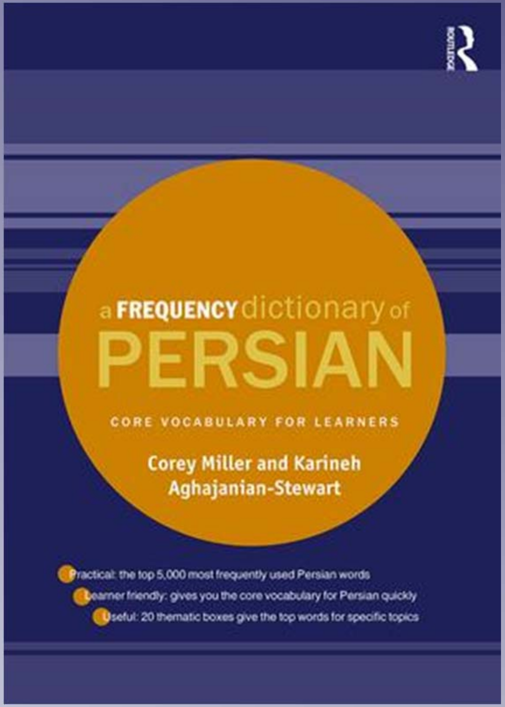
Building a Persian Frequency Dictionary

Corey Miller (MITRE) & Karineh Aghajanian-Stewart

May 24, 2018



a FREQUENCY dictionary of PERSIAN

CORE VOCABULARY FOR LEARNERS

Corey Miller and Karineh Aghajanian-Stewart

- Practical: the top 5,000 most frequently used Persian words
- Learner friendly: gives you the core vocabulary for Persian quickly
- Useful: 20 thematic boxes give the top words for specific topics

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.

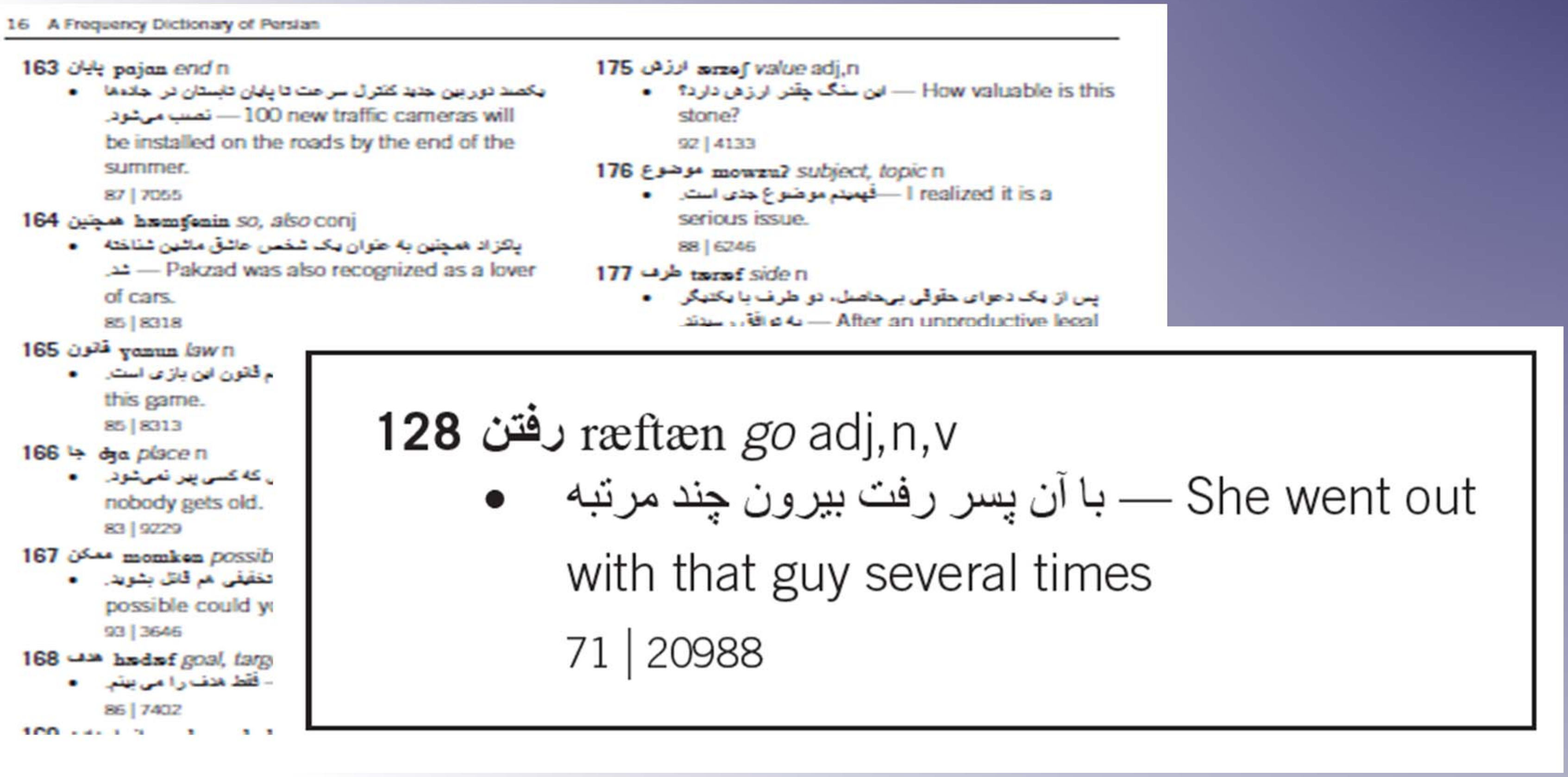
Outline

- 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

Sample page from *A Frequency Dictionary of Persian*



16 A Frequency Dictionary of Persian

163 پایان **pajan** *end* n
- یکصد دوربین جدید کنترل سرعت تا پایان تابستان در جاده‌ها نصب می‌شود. — 100 new traffic cameras will be installed on the roads by the end of the summer.

  87 | 7055

164 همچنین **hæmʧenin** *so, also* conj
- پاکزاد همچنین به عنوان یک شخص عاشق ماشین شناخته شد. — Pakzad was also recognized as a lover of cars.

  85 | 8318

165 قانون **ɣanun** *law* n
- م قانون این بازی است.

  this game.

  85 | 8313

166 جا **dʒa** *place* n
- ی که کسی پیر نمی‌شود.

  nobody gets old.

  83 | 9229

167 ممکن **momken** *possib*
- تحقیقی هم قاتل بشوند.

  possible could y

  93 | 3646

168 هدف **hædæf** *goal, targ*
- فقط هدف را می بینم. -

  86 | 7402

175 ارزش **ærzeʃ** *value* adj,n
- این سنگ چقدر ارزش دارد؟ — How valuable is this stone?

  92 | 4133

176 موضوع **mowzuʔ** *subject, topic* n
- فهمیدم موضوع جدی است.— I realized it is a serious issue.

  88 | 6246

177 طرف **tæræf** *side* n
- پس از یک دعوای حقوقی بی‌حاصل، دو طرف با یکدیگر به توافق رسیدند — After an unproductive legal

128 رفتن ræftæn *go* adj,n,v
- با آن پسر رفت بیرون چند مرتبه — She went out with that guy several times

  71 | 20988

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)

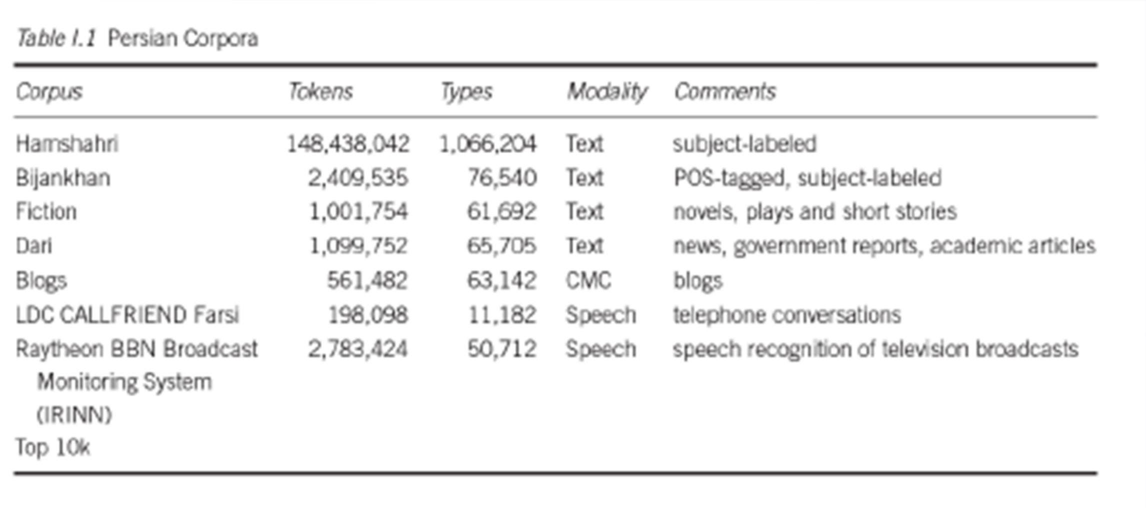


Table I.1 Persian Corpora

Corpus	*Tokens*	*Types*	*Modality*	*Comments*
Hamshahri	148,438,042	1,066,204	Text	subject-labeled
Bijankhan	2,409,535	76,540	Text	POS-tagged, subject-labeled
Fiction	1,001,754	61,692	Text	novels, plays and short stories
Dari	1,099,752	65,705	Text	news, government reports, academic articles
Blogs	561,482	63,142	CMC	blogs
LDC CALLFRIEND Farsi	198,098	11,182	Speech	telephone conversations
Raytheon BBN Broadcast Monitoring System (IRINN)	2,783,424	50,712	Speech	speech recognition of television broadcasts
Top 10k				

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

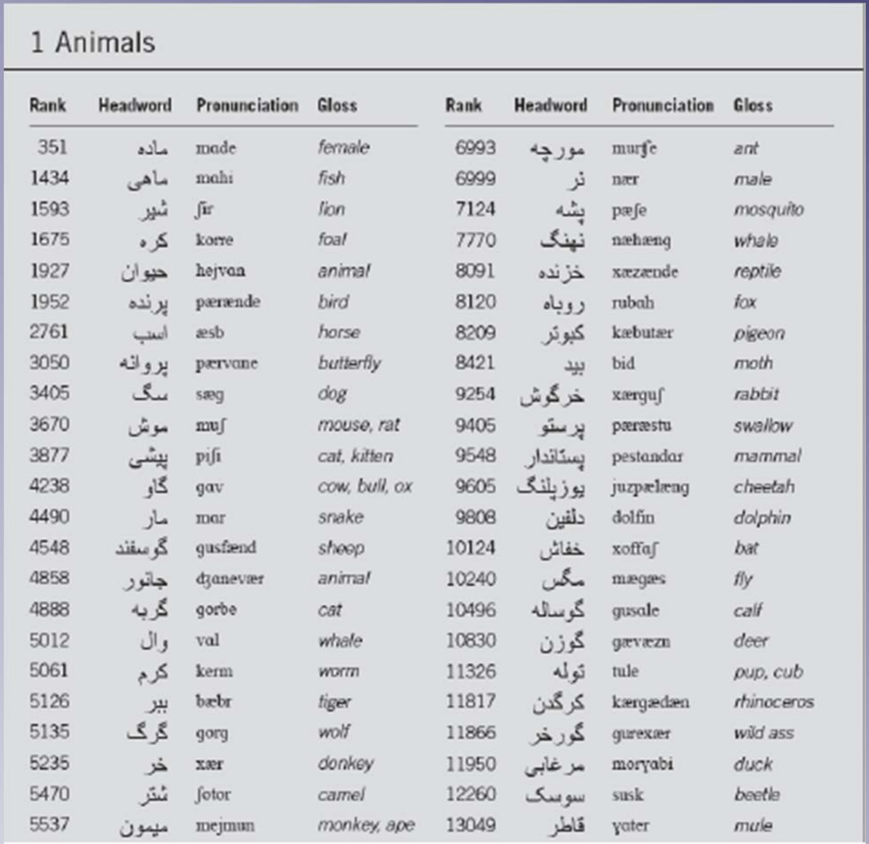
- Spelling variation
  - Hamze followed by yeh, or two yehs
    - آمریکایی، آمریکائی (or امریکایی، امریکائی) ‘American’
  - Persian or Arabic Unicode codepoints
    - ک or ك, ی or ي
  - Bound/unbound/half-space (zero-width non-joiner, ZWNJ)
    - کتابخانه، کتاب خانه، کتاب‌خانه ‘library’
- Register variation (Standard/Colloquial)
  - ‘house’ میکنند/میکنن، خانه/خونه ‘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*): بطور کلی ‘generally’, اینجاست ‘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. آذربایجان ‘Azerbaijan’ analyzed as آذربایج + ان
- Ambiguous forms
  - کتابی is either /keˈtɒbi/ ‘a book’ or /ketɒˈbi/ ‘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



1 Animals

Rank	Headword	Pronunciation	Gloss	Rank	Headword	Pronunciation	Gloss
351	ماده	mɑde	*female*	6993	مورچه	murʧe	*ant*
1434	ماهی	mɑhi	*fish*	6999	نر	nær	*male*
1593	شیر	ʃir	*lion*	7124	پشه	pæʃe	*mosquito*
1675	کره	korre	*foal*	7770	نهنگ	næhæng	*whale*
1927	حیوان	hejvɑn	*animal*	8091	خزنده	xæzænde	*reptile*
1952	پرنده	pærænde	*bird*	8120	روباه	rubɑh	*fox*
2761	اسب	æsb	*horse*	8209	کبوتر	kæbutær	*pigeon*
3050	پروانه	pærvɑne	*butterfly*	8421	بید	bid	*moth*
3405	سگ	sæg	*dog*	9254	خرگوش	xærguʃ	*rabbit*
3670	موش	muʃ	*mouse, rat*	9405	پرستو	pæræstu	*swallow*
3877	پیشی	piʃi	*cat, kitten*	9548	پستاندار	pestɑndɑr	*mammal*
4238	گاو	gɑv	*cow, bull, ox*	9605	یوزپلنگ	juzpælæng	*cheetah*
4490	مار	mɑr	*snake*	9808	دلفین	dolfin	*dolphin*
4548	گوسفند	gusfænd	*sheep*	10124	خفاش	xoffɑʃ	*bat*
4858	جانور	dʒɑnevær	*animal*	10240	مگس	mægæs	*fly*
4888	گربه	gorbe	*cat*	10496	گوساله	gusɑle	*calf*
5012	وال	vɑl	*whale*	10830	گوزن	gævæzn	*deer*
5061	کرم	kerm	*worm*	11326	توله	tule	*pup, cub*
5126	ببر	bæbr	*tiger*	11817	کرگدن	kærgædæn	*rhinoceros*
5135	گرگ	gorg	*wolf*	11866	گورخر	gurexær	*wild ass*
5235	خر	xær	*donkey*	11950	مرغابی	morɣɑbi	*duck*
5470	شتر	ʃotor	*camel*	12260	سوسک	susk	*beetle*
5537	میمون	mejmun	*monkey, ape*	13049	قاطر	ɣɑter	*mule*

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’