



CENTER FOR ADVANCED
STUDY OF LANGUAGE

Discovery and Processing of Persian Language Resources

Presented by

Corey Miller

Senior Research Scientist

September 28, 2015



Overview

- *Did You Mean?* (DYM)
- Dialects
- Corpus
- Taarof
- Colloquial and Cyber




Did You Mean? (DYM)

- Lexical Resources
 - Haim
 - Steingass
 - UNO Dari
 - Kabul Phonebook
 - Corpus Wordlist
- Pinglish
- Morphological Analysis
- Talking Dictionary



Morphology



Did You Mean...?

[DYM Main]

Enter a Farsi or Dari word in Persian script to search for it in the lexicons.

Farsi & Dari کتابها Submit Reset

Displaying results for کتابها.

Entry	Phonetic	Similarity	Is This It?	Inflection	Definition	Part of Speech	Source
کتاب	ketab	★★★★☆	This is it	citation ⊕	<ul style="list-style-type: none"> book treatise sacred-book scripture letter 	noun	Haim
کتابها		★★★★☆	This is it	derivation ⊖	Morphological analysis: کتابها is a derivative of کتاب <ul style="list-style-type: none"> Plural noun Plural noun with ezafe 		
کتابهای		★★★★☆	This is it	derivation ⊕			
کتابهائی		★★★★☆	This is it	derivation ⊕			
کتابهائی		★★★★☆	This is it	derivation ⊕			

Pinglish

Did You Mean...?

[DYM Main]

Enter a Pinglish word to search for it in the lexicons.

Pinglish ▼ 4sat

Displaying results for 4sat.

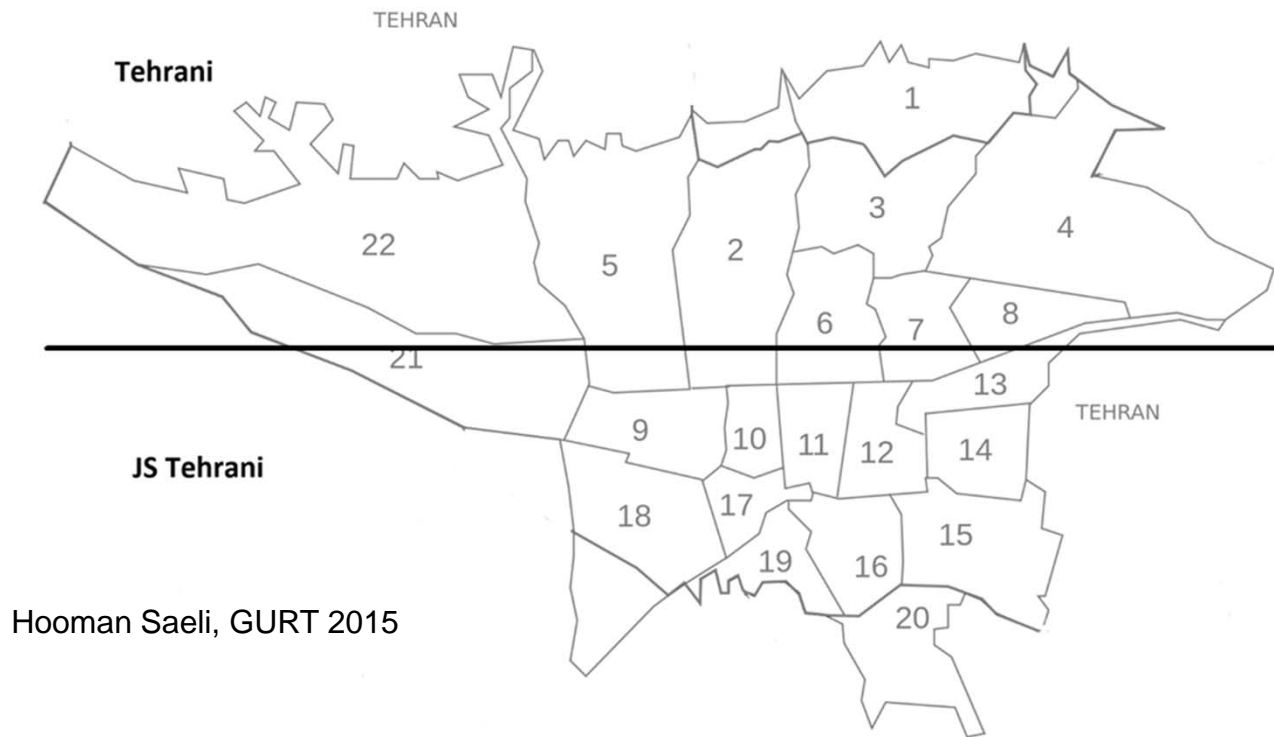
Entry	Phonetic	Similarity	Is This It?	Inflection	Definition	Part of Speech	Source
فرصت	forsæt	★★★★★	This is it!	citation ⊕	<ul style="list-style-type: none"> • opportunity • occasion • leisure 	noun	Haim
فرصات		★★★★☆	This is it!	derivation ⊕	<ul style="list-style-type: none"> • time • chance • leave 		
فرصت		★★★★★	This is it!	word ⊕	Glossary: opportunity Pronunciation: forsæt Frequency Rank: 485		Wordlist
فرصات		★★★★☆	This is it!	derivation ⊕			
فرس	færæs		This is it!	citation	<ul style="list-style-type: none"> • horse 	noun	Haim
فرست		★★★★☆	This is it!	derivation ⊕	<ul style="list-style-type: none"> • Pegasus 		
فرسات		★★★★☆	This is it!	derivation ⊕			



Persian dialects



Tehrani and Jonub Shahri Tehrani



Hooman Saali, GURT 2015

https://umdsurvey.umd.edu/jfe/form/SV_6JydnZbB4NBZ9Pf

Persian Corpus

- Hamshahri
- Bijankhan
- Dari
- Blogs
- LDC (conversation)
- 5000 most frequent words arranged by word family
- Example sentences



Taarof

- Routine politeness strategies
- Specialized phraseology
- Learning modules
 - Sample videos
 - Reference vocabulary



Coming soon: Colloquial and Cyber

- Diglossia
- Colloquial language characterized by specialized
 - Pronunciation
 - Morphology
 - Syntax
 - Vocabulary
- Cyber
 - Vocabulary, abbreviations, technical



For more information

Contact

Corey Miller

Senior Research Scientist

301 226 8834

cmiller@casl.umd.edu

www.casl.umd.edu

