Persian
(ISO 639-1:2002)

Naghmeh Fazeli
Summer Semester 2016
Persian Spoken Countries

- Persian: **Iran** (spoken by 45 million people in Iran out of a total population of 81 million)

- Dari/Afghan Persian: **Afghanistan** and Pakistan (7.6 million)

- Tajiki/Tajiki Persian: **Tajikistan and Uzbekistan** (4.5 million)

- But also in some of the countries in the south of the Persian Gulf....

- Totally spoken by **134** million people as first or second language
Persian or Farsi?

- **How should we call the official language of Iran?** Persian is the English translation/version of Farsi.

- **How would we call the other two variants of the language, namely Dari and Tajiki?** If we label the three as Persian, Dari, and Tajiki we will miss the tight genetic relationship between them, given the fact that all three are derived from Middle Persian.
First Solution

One solution is to employ *Persian for Farsi*, and call the other two variants *Tajiki-Persian* and *Dari-Persian*. This type of labeling clearly shows the relationship between the three variants, given the fact that all three variants are derived from *Middle Persian* or *Farsi-ye Miyane*.

The problem with this solution:

The unmarked usage of *Persian*, referring to *Farsi*, gives the impression that this variant is the main version, and the other two have derived from it. This is not true linguistically though: the syntactic properties of *Tajiki* and *Dari*, for example, are closer to the syntax of the earlier stages of *Persian* than *Farsi* is.
A second solution is to follow the existing tradition used for other languages: the two major variants of Portuguese, for example, are called *European Portuguese* versus *Brazilian Portuguese*. *British English* versus *Standard American English* is another example. Given this analogy, we will arrive at *Iranian Persian* (for Farsi), *Tajiki Persian*, and *Dari* (Afghani) Persian.
Statistics for the Languages of Iran
Statistics from CIA World Factbook

- **Persian, Gilaki** and **Mazandarani** and **Luri** 66%
- **Azerbaijani** and other **Turkic** dialects 18%
- **Kurdish** 10%
- **Arabic** 2%
- **Balochi** 2%
- **Tati** and **Talysh** 1%
### Old Persian

**Independent vowels**
- 103A0 _OLD PERSIAN SIGN A
- 103A1 _OLD PERSIAN SIGN I
- 103A2 _OLD PERSIAN SIGN U

**Consonants**
- 103A4 _OLD PERSIAN SIGN KA
- 103A5 _OLD PERSIAN SIGN KU
- 103A6 _OLD PERSIAN SIGN GA
- 103A7 _OLD PERSIAN SIGN GU
- 103A8 _OLD PERSIAN SIGN XA
- 103A9 _OLD PERSIAN SIGN JA
- 103AA _OLD PERSIAN SIGN JI
- 103AB _OLD PERSIAN SIGN TA
- 103AC _OLD PERSIAN SIGN TU
- 103AD _OLD PERSIAN SIGN DA
- 103AE _OLD PERSIAN SIGN DI
- 103AF _OLD PERSIAN SIGN DU
- 103B0 _OLD PERSIAN SIGN THA
- 103B1 _OLD PERSIAN SIGN PA
- 103B2 _OLD PERSIAN SIGN BA
- 103B3 _OLD PERSIAN SIGN FA
- 103B4 _OLD PERSIAN SIGN NA
- 103B5 _OLD PERSIAN SIGN NU
- 103B6 _OLD PERSIAN SIGN MA
- 103B7 _OLD PERSIAN SIGN MI
- 103B8 _OLD PERSIAN SIGN MU
- 103B9 _OLD PERSIAN SIGN YA
- 103BA _OLD PERSIAN SIGN VA
- 103BB _OLD PERSIAN SIGN VI
- 103BC _OLD PERSIAN SIGN RA
- 103BD _OLD PERSIAN SIGN RU
- 103BE _OLD PERSIAN SIGN LA
- 103BF _OLD PERSIAN SIGN SA
- 103C0 _OLD PERSIAN SIGN ZA
- 103C1 _OLD PERSIAN SIGN SHA
- 103C2 _OLD PERSIAN SIGN SSA
- 103C3 _OLD PERSIAN SIGN HA

**Various signs**
- 103C5 _OLD PERSIAN SIGN AURAMAZDA
- 103C6 _OLD PERSIAN SIGN AURAMAZDA-2
- 103C7 _OLD PERSIAN SIGN XSHAAYATHYA
- 103C8 _OLD PERSIAN SIGN DHAHYAAAUSH
- 103C9 _OLD PERSIAN SIGN DHAHYAAAUSH-2
- 103CA _OLD PERSIAN SIGN BAGA
- 103CB _OLD PERSIAN SIGN BUJUMISH

**Punctuation**
- 103D0 _OLD PERSIAN WORD DIVIDER

**Numbers**
- 103D1 _OLD PERSIAN NUMBER ONE
- 103D2 _OLD PERSIAN NUMBER TWO
- 103D3 _OLD PERSIAN NUMBER TEN
- 103D4 _OLD PERSIAN NUMBER TWENTY
- 103D5 _OLD PERSIAN NUMBER HUNDRED
55 Endangered Languages in Iran
List of Available Text Corpora for Research

• Hamshahri Corpus (Ehsan Darrudi-DBRG Group)
• TEP: Tehran English-Persian parallel corpus
• TMC: Tehran Monolingual Corpus (NLP Lab of University of Tehran)
• Persian today corpus (Hamid Hassani, ILI)
• Bijankhan corpus (University of Tehran)
• Persica
Challenges in Persian Electronic Text Analysis

- Ambiguities while Characters Manipulation
- Ambiguity at Word Boundaries
- Ambiguity in morphology
- Ambiguity while Detecting Proper Nouns in Persian
- Ambiguity in Persian Syntax analysis
Projects

• PersianSMT
• Persica
PersianSMT

Goals:
2. Creation of largest English Persian parallel corpus by the use of movie subtitles.

Result: SMT system strongly outperforms the Google translator in translating both in-domain (movie subtitle) and out-of-domain sentences.
Persica (a multiplication text corpus)

This project presents a new corpus for NEWS articles analysis in Persian called Persica.

NEWS analysis includes NEWS classification, topic discovery and classification, category classification and many more procedures.

Dealing with NEWS has special requirements and first of all a valid and reliable corpus to perform the experiments on them.
References

- [http://www.endangeredlanguages.com](http://www.endangeredlanguages.com)
- [https://sourceforge.net/projects/persica/](https://sourceforge.net/projects/persica/)
- [http://aboutworldlanguages.com](http://aboutworldlanguages.com)
- [https://wikipedia.org/](https://wikipedia.org/)

- Challenges in Persian Electronic Text Analysis
  Behrang QasemiZadeh, Saeed Rahimi and Mehdi Safaee Ghalati
  a Text and Speech Technology Ltd., Tehran, Iran
  b Tehran University, Faculty of Literature and Humanities, Enqelab, Tehran, Iran
از توجه شما ممنونم