

# How to Collect, Create, and Implement Collation Information (for technical guys)

Roozbeh Pournader  
Sharif FarsiWeb, Inc.

[roozbeh@farsiweb.info](mailto:roozbeh@farsiweb.info)

# Why is this necessary?

- Make the readers be able to use lists without hassle
  - Any list: dictionaries, phonebooks, employee lists, ...
- Make your requirements known to vendors
- Unification is required, customization is a good plus

# Step by Step

- Learn about the standards
- Create a first table
- Go decode dictionaries
- Create the best table you can
- Contact language authorities
- Finalize it, and get it approved
- Publish and Publicize it

# Learning about the standards

- ISO/IEC 14651 and UTS #10
  - Read them, again and again
  - Check the Unicode default
- Local standards?
  - “Introduction” part of dictionaries
  - Librarian manuals
  - Guides on indexing
  - National standards

# Learning about the standards

- Existing practice
  - Check existing software
  - But don't contact the vendors
- Participate on Unicode mailing list
  - Don't ask them to fix the default

# Create the first table

- You know the order of the alphabet, don't you?
- Get a Unicode chart, choose all the characters needed in your language (letters, diacritics, punctuation, ...)
  - Beware of pitfalls: similar but non-identical characters
  - Publish that list to online communities, for feedback

# Create the first table

- Decide about the levels
  - Main letters on first level
  - Variants on second level
  - Diacritics on third level
  - Punctuation on fourth level
- Decide about the ordering of characters inside levels

# Test that first table

- Get a tool
  - GNU's C library (glibc)
  - IBM's International Classes for Unicode (ICU)
    - There is a online web-based demo
    - You can also download it

# Test that first table

- Create test data
  - Not necessarily meaningful words or phrases
- Create the table (or the delta/tailoring)
- Test it
- Fix it
- Go back to testing until you're satisfied

# Decoding dictionaries

- Go find every authoritative dictionary
  - Don't buy them now, first check if they're usable for collation purposes
  - Possible other sources (Hafez)
- Find about edge cases
  - Browse through the dictionary
  - Check rare cases (دائرة المعارف)
  - Check diacritics and controls (ZWJ, ZWNJ)

# Decoding dictionaries

- Create a list of differing opinions
  - Mostly edge cases
    - Uncommon letters
    - Diacritics
    - Controls
  - Also frequent cases, sometimes
- Find about similar languages
  - Urdu: Persian, Pashto, Arabic

# Create your best table

- Create the best table and test data you technically could:
  - Put a lot of effort into it
  - Work in a team
  - Revise frequently, retest
  - Ask for advice from Unicode and i18n experts
  - Have some decision mechanism
    - Dictatorship is best

# Create your best table

- Decide about everything
  - Based on the references, of course
  - Back them using test data
  - Be ready to create alternative tables for special cases (but try to avoid this)
    - Personal names
- Publish the work online, as “the best we could do”
  - Post installation and testing guides
  - Get feedback

# Contact the authorities

- Academy
  - Language Academies
  - Science Academies
  - Try to create a special small and controlled committee for this

# Contact the authorities

- Government
  - Ministry of Culture
  - National standards body (ISO representative)
  - Ministry of Information Technology
  - Unicode and W3C representatives

# Contact the authorities

## ■ Experts

- Authors of dictionaries
- Linguists understanding computers or logic
- Go for the best possible, but be ready to be turned down

# Finalize it, and get it approved

- Make them decide on everything
  - Best is to make them accept your table
  - But you *must* be very accepting
- Get a stamp under it, also a number if possible
  - “Persian Academy’s Official Sorting Specification”
  - “Iranian National Standard for Collation”

# Publish it

- Write a detailed description
  - As solid as possible
  - Write it for dummies (both technical and linguistics)
  - Don't explain your decisions
  - Don't mention internal disagreements
  - In both English and the language
  - Both on paper and PDF

# Publish it

- Don't forget:
  - Weight tables, both on paper and electronically
  - Vast and detailed test data, both on paper and electronically
  - Your sources, committee members, and authorities

# Publicize it

- Put it on a good server
- Post the link to Unicode list and other linguistic/computing communities
- Send paper and electronic copies to the vendors you know
- Send electronic copies to Free Software/Open Source communities
- Refer to it frequently

# Final recommendations

- Don't invent stuff
  - But feel free to “suggest”
- Don't ignore details
- Ask for advice
  - Language authorities (dead or alive)
  - Unicode and i18n experts
- Do your absolute best
  - But remember that you may need to revise

# Questions

