An Investigation into the uses of Factotum and extending its capabilities

By Claire Abu-Hakima

ABSTRACT

I. Introduction

In 1993, Robert Uzgalis came up with Factotum 90, a software tool to help create taxonomies for different sets of data, with no natural language restriction.

a. What is factotum
- i. Written by Uzgalis
- ii. Original development and purpose
  - 1. Provide data models for researchers
  - 2. Language independent
  - 3. Force to formalize facts
- iii. what ended up happening to it

b. What I did with Factotum
- i. Want to demonstrate it's usefulness
  - 1. Given my interest/study of linguistics, want to demonstrate the potential use of factotum as a research tool
    - a. Created a data set of about 20 languages with different language attributes
    - b. All data in one place/data base (generally very spread out online)
    - c. Possible new connections formed/recognized
- ii. Extend its capabilities/what worked on
  - 1. Brief intro to vocabulary parser
  - 2. Brief intro to fact checker
- iii. Possible improvements for factotum
  - 1. Vocabulary generation:
    - a. Less of a cycle of 'massaging by hand' then running it through the software,
      - i. Either: more/entirely automated
      - ii. Or mainly human created (which was in fact my first impression)

II. How Factotum Works & What I worked on

a. Vocabulary parser
- i. Describe what is a vocabulary

      1. Vocabulary can be thought of as a set of ‘rules’ by which the data must abide
      2. But the vocabulary is in itself also a set of facts, since there are generating rules, implying rules, and since in general it give us information about the data (even if it is in terms of formatting)

   - ii. How is the vocabulary presented
      1. Several forms of the vocabulary: (brief & clarifying)
         - a. If blocks/ “rule restrictions”
         - b. Type trees
         - c. Tags and strings
         - d. Strings
         - e. rule is
         - f. phrase is
         - g. attribute
         - h. nattribute
         - i. generation rule
         - j. implication rule
      2. within these, there may be “objects” which refer entities
         - a. clarify how they are presented label, type, token type, etc.
   - iii. PARSER
      1. What my parser did
         - a. Dealt with the strings
            - i. Why
            - ii. How
         - b. Sample code
         - c. Point out where called upon existing factotum functions (eg Lex, entities etc)
         - d. What design decisions I made
            - i. E.g. how to hold the data
            - ii. How to traverse it
- b. FACT CHECKER
   - i. What is the FactChecker
      1. Purpose – make sure facts obey the vocabulary rules (e.g. format) prescribed to them
      2. Necessity—don’t want fact that can’t fit the model
         - a. Can adjust the rules/fact format and in return get better results
         - b. Will force user to look at data and rules more intenselyl
   - ii. Code
      1. My code samples

            2. Code Functionality
                a. Control flow
                b. “internal structure of data”
            3. Design decisions I made
                a. Helper functions
                b. errors it chooses to show

III. Using Factotum –Linguistic Data
    c. Why use Factotum?
        i. Possible linguistic data model
        ii. Collection of attributes all in one place
        iii. Linguistic knowledge currently quite spread out
        iv. No centralized data base/ source of information
        v. Perhaps patterns that have not yet been recognized will become apparent
    d. Design Decisions
        i. What I chose to focus on
            1. Linguistics → languages
            2. Natural languages
            3. Written language
        ii. Which attributes – all available facts
            1. Language family (Genetic)
                a. Can make connections well
                b. Represented as type (< >)
            2. Alphabet
                a. Phrasal, with option of a –variant
            3. grammar
                a. cases?
                b. Gender?
                c. Tense
                d. Moods
                e. Pronouns
                f. Inflection of adj/nouns/verbs/etc
            4. Typology
                a. Word order
                b. Unusual syntactic structure
            5. Phonetics/phonology?
                a. Unusal features??
                b. Eg clicks
            6. Stats/numbers [need better sources if going to include]
                a. Number of total speakers
                b. where official lang
                c. dialects

    - d. types of dialects – details?
- iii. what attributes to skip
  - 1. historical
    - a. not doing a historical linguistics analysis
    - b. origins of writing systems etc interesting, but not our scope
  - 2. political /religious aspects
    - a. many nuances involved with this, would want to do it justice but don’t have time to do thorough discussion/ analysis of politics
    - b. only mention will be:
      - i. names – for example if essentially same lang but different names/dialects then will provide aliases for them
      - ii. geography
        - 1. if language was introduced to a region due to imperialism,
          - a. will just be noted that language X is official/spoken in country Y
    - c. examples
      - i. cumbersome
      - ii. complicated
      - iii. need to veryify
      - iv. can be added by experts in the different languages → potential growth

- iv. Issues with attributes
  - 1. Competing theories,
    - a. If there is a generally accepted theory, then that is the one that is noted
    - b. if there are competing theories
      - i. no clear front runner
      - ii. so I’ve made the decision to mention both of the contending/leading theories
  - 2. Ambiguity
    - a. Lang doesn’t follow pattern all the time
    - b. Lang has some deviations from attribute (e.g. word order)
- v. Case Studies:
  - 1. Arabic
    - a. How represented data

- b. What additional data decided to use
- c. What learned from that data
- d. What learned that only Factotum demonstrated

2. Serbo-Croatian
   - a. How represented data
   - b. What additional data decided to use
   - c. What learned from that data
   - d. What learned that only factotum demonstrated

- vi. Results of running data set with factotum
  1. Excerpt/printout of data
  2. How exactly I ran it
     - a. So that in future it will be easier to reproduce
  3. Excerpt/print out of results
     - a. Description of what each section of results demonstrates
     - b. How do these results help us/ be useful to others
  4. After running fact checker with the results,
     - a. How did I have to adjust vocabulary
       - i. why did we adjust it
     - b. How did I have to adjust facts
       - i. Why did I adjust them
       - ii. What results did that yield

- IV. Improvements to Factotum
  - a. Vocabulary structure after mkvocab
  - b. Either more automation or less automation of vocabulary, combination makes things tricky
  - c.

- V. Conclusion
  - a. What learned about factotum/parsing through data
  - b. What learned from linguistic data set
  - c. What learned about data models

Sources