TallVocabL2Fi

A Tall Dataset of 15 Finnish L2 Learners' Vocabulary

Frankie Robertson[†] Li-Hsin Chang[♣] Sini Söyrinki[†]

University of Jyväskylä[†]

University of Turku[♣]

May 10, 2022

Definitions

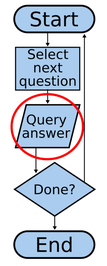
- *Computer Adaptive Testing (CAT)*
  - Testing technique
  - Questions given to a testee are selected online during the test from a question bank
    - Based on a provisional ability estimate
    - Which is based on their previous responses
  - Active learning on an ordered difficult/ability scale
- *Vocabulary Inventory Prediction (VIP)*
  - Prediction task
  - Input: a sample of a learner's vocabulary
    - e.g. self-rating known/unknown responses for a small set of words
  - Output: a function which given an input word predicts known/unknown
- *The Finnish language*
  - Rich inflectional & derivational morphology and word formation
- *Tall dataset*
  - Dataset with lots of data for each participant

Prior vocabulary inventory datasets

- Tall vocabulary dataset from Ehara with answers from L2 English learners in Japan[1]
  - 15 learners self-assess knowledge of 12 000 words on 5-point scale
  - The only prior tall dataset the authors are aware of
- There have also been non-tall vocabulary data collected in other fields
  - Second Language Vocabulary Acquisition
    - None or very little released as open data
  - Psycholinguistics
    - Tend to be only interested in surface forms

[1] `yoehara.com/esl-vocabulary-dataset/`

Why a tall dataset?



- A tall dataset gives better evaluation of VIP
  - Need a lot of information about the vocabulary of each individual
- Lesser reason: validity of the evaluation metrics
  - Increase statistical power
  - Not biased to a few benchmark words
- Main reason: allow evaluation of CAT-based VIP via simulation
  - Normally the "query answer" step of the CAT would normally solicit a response directly from the testee
  - Replace with retrieval of a response from a tall vocabulary dataset
    - "select next question" is limited to the words in the vocabulary dataset
    - Good lexical coverage is needed to prevent limiting the CAT algorithm

Design

- Basic design owes a large debt to Ehara's resource
  - Some changes to make a new, complimentary resource
- Word list
  - Custom word list rather than pre-existing
    - Ehara used a word list for language learners
  - Include also low frequency words
    - Evaluate predictions on these too, avoid circularity
- Triangulation
  - 100 word translation test to estimate self-rating reliability
- More diverse participants
  - Distributed across 3 L1s
    - and 4 CEFR levels
  - Conducted online $\implies$ participants spread across Finland

Creating the word list

- Input data
  - Lemma frequencies from corpora across different genres
  - The new Finnish word list (published by the Institute for the Languages of Finland)
  - Categories of ordinal-like words from Kaikki.org (from Wiktionary)
  - Derivational data from OMorFi and parsed Wikitionary data
- Take some amount from high frequency words unconditionally
- Then create a filtered word list
  - Filter out word derivations estimated to be compositional
  - Filter out certain classes of words such as ordinal-like ones
- Stratified sampling from different frequency bands within different genres
- More detail in the paper

Word knowledge self-rating scale

1. I have never seen the word before
2. I have probably seen the word before, but don't know the meaning
3. I have definitely seen the word before, but don't know the meaning / I have tried to learn the word but have forgotten the meaning
4. I probably know the word's meaning or am able to guess
5. I absolutely know the word's meaning

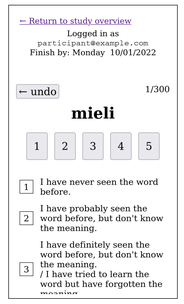
Getting participants

- As a first step expressions of interest were collected
- Recruitment was mainly from Facebook
  - Groups such as "Foreigners in Finland", "Foreigners in Jyväskylä", and "International Working Women of Finland"
  - Once native languages with almost enough interest emerged, the study was advertised in more specific groups
    - e.g. "Brits in Finland" for native English speakers
- Interested participants from relevant groups were then invited
- Participants paid €200 on completion of task
  - Thanks to funding from the Faculty of IT at the University of Jyväskylä
- 154 expressions of interest
- 31 invited
- 15 completed

Participants by CEFR level and L1

	B1	B2	C1	C2	Total
English	2	2	1	0	5
Hungarian	0	1	2	1	4
Russian	2	1	2	1	6
Total	4	4	5	2	15

Collection website



- Written in Python + Quart (like Flask) + htmlx
- Responsive + progressive (good mobile support)
- Hosting thanks to CSC's (Finnish national computing provider) Rahti service
- Open source on "as-is" basis
  - `frankier/finnvocabcollect`

Translation test

- For each participant, group words by responses
  - Sample 20 random words from each group = 100 words

Responses

1. Translate the word
2. Topic of the word
3. Don't know

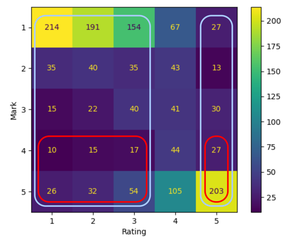
- Native language, English or Finnish

Marks

1a. Completely incorrect answer
1b. No answer
2. Partially correct but misleading
3. Correct enough to help understanding
4. Not quite correct, but fully understandable
5. Completely correct

- Marking done by two markers with help of machine translation + dictionary
  - Initial Quadratic Weighted Kappa of 0.86
  - Later conferred on disagreements to produce final labelling
  - Full details in paper

Enrichment

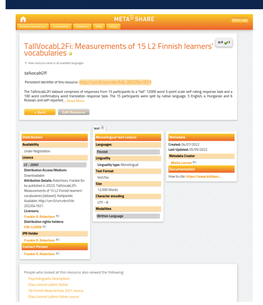


$$\text{reliability} = P(\textit{mark} \geq 4 \mid \textit{rating} \geq 5)$$
$$\text{underrating} = P(\textit{mark} \geq 4 \mid \textit{rating} \leq 3)$$

Balanced reliability measure summarises reliability and underrating similarly to balanced accuracy:

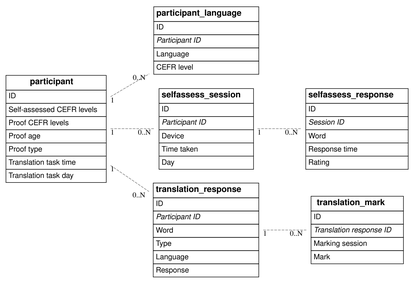
$$\text{reliability}_{\text{bal}} = \frac{1}{2}(\text{reliability} + (1 - \text{underrating}))$$

Release



- Deposited in the Language Bank of Finland
- `http://urn.fi/urn:nbn:fi:lb-2022041921`
- Cross referenced
  - IRIS Database
  - ELRA LRE map
- Public domain/CC0

Format



- Distributed as a series of TSV files
- Datatype information exported from DuckDB
- Fully documented in README