A Tall Dataset of 15 Finnish L2 Learners’ Vocabulary

Frankie Robertson1 Li-Hsin Chang2 Sini Säyninki1 University of Jyväskylä University of Turku May 10, 2020

Definitions

- Computer Adaptive Testing (CAT)
  - Testing technique
  - Questions given to a tester are selected online during the test from a question bank
  - Based on a provisional ability estimate
  - Active learning on an ordered difficulty scale

Vocabulary Inventory Prediction (VIP)

- Prediction task
  - Input: a sample of a learner’s vocabulary
  - Example self-rated known/unknown responses for a small set of words
  - Output: a function which given an input word predicts known/unknown
- The Finnish language
  - Rich inflective & derivational morphology and word formation

Definitions

- Tall vocabulary dataset from Ehara with answers from L2 English learners in Japan
  - 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

Definitions

- 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
  - Tall vocabulary dataset from Ehara with answers from L2 English learners in Japan
  - Distributed across 3 L1s
  - Conducted online — participants spread across Finland
  - Includes also low frequency words
  - Some changes to make a new, complimentary resource

Design

- 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)
  - Take some amount from high frequency words unconditionally
  - Then create a filtered word list
  - Filter out words estimated to be computational
  - Filter out certain classes of words such as ordinal-like seen
  - Stratified sampling from different frequency bands within different genres
  - More detail in the paper

Creating the word list

- 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)
  - Take some amount from high frequency words unconditionally
  - Then create a filtered word list
  - Filter out words estimated to be computational
  - Filter out certain classes of words such as ordinal-like seen
  - Stratified sampling from different frequency bands within different genres
  - More detail in the paper

Creating the word list

- 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
  - 4. I probably know the word’s meaning or am able to guess
  - 5. I absolutely know the word’s meaning

Word knowledge self-rating scale

- Participants by CEFR level and L1

Participants by CEFR level and L1

- Collection website

Collection website

- Translation test

Translation test

- Enrichment

Enrichment

- Why a tall dataset?

Why a tall dataset?

- Getting participants

Getting participants

- Release

Release

- Format

Format

- Definitions

Definitions

- Prior vocabulary inventory datasets

Prior vocabulary inventory datasets

- Design

Design

- Creating the word list

Creating the word list

- Word knowledge self-rating scale

Word knowledge self-rating scale

- Participants by CEFR level and L1

Participants by CEFR level and L1

- Collection website

Collection website

- Translation test

Translation test

- Enrichment

Enrichment

- Why a tall dataset?

Why a tall dataset?

- Getting participants

Getting participants

- Release

Release

- Format

Format