Definitions

- TSV: Tab Separated Values; A tabular file format with one row per line, and with columns/fields separated by tab characters
- CSV: Comma Separated Values, as above but with a comma as the field separator
- DuckDB: An embedded analytical-relational database available at DuckDB.org
- SQL: Structured Query Language; A programming language for querying and manipulating relational databases. Supported by DuckDB.
- Primary key: Relational database terminology referring to a unique identifier for a row in the current table.
- Foreign key: As above but a reference to a row in another table
- Synthetic primary key: A "made up" identifier for each row, e.g. consecutive numbers 1, 2, 3, . . .
- CEFR: The Common European Framework of Reference for Languages. In CEFR, the skills of speaking, writing, listening comprehension and reading comprehension are given coarse-grained levels from A1 to C2. The scale can be understood through a CEFR self-assessment grid.
- YKI: Short for *Yleinen kielitutkinto*; The main general language certificate administered in Finland. We refer by default to the Finnish language certificate here.

Description

This resource has been deposited at the Language Bank of Finland, where it has been assigned the identifier urn:nbn:fi:lb-2022041921.

The TallVocabL2Fi dataset comprises of responses from 15 participants to a "tall" 12000 word 5-point scale self-rating response task and a 100 word confirmatory word translation response task.

The dataset is unique in its combination of the tall data collection set up, where responses are collected for many words, the varied backgrounds of the learners, the use of Finnish prompt words, and the triangulation with a word translation test.

The dataset can be used for vocabulary acquisition research in general, but it is particularly suited to evaluation of the task of Vocabulary Inventory Prediction (VIP) including techniques based on Computer-Adaptive Testing (CAT).

The dataset is relational/tabular. It is distributed as a series of TSV files along with a SQL schema exported from DuckDB.

The 15 participants were split by native language, 5 English, 4 Hungarian and 6 Russian, and self-reported CEFR reading level, 5 B1, 4 B2, 5 C1 and 2 C2. The data was gathered through a website from paid participants resident in Finland over a period of 3 months from September and November 2021. In total there are 180 thousand word knowledge self-rating responses and 1.5 thousand word translation responses.

Dataset format, schema and coding

Formats

There are two formats available. The *simple format* contains only the response data to the self-rating test, and the marks from the translation task, and is meant as an quick analyses. The detailed release format, which contains the full dataset, is the *relational format*. Both formats are TSV based. All text is encoded as UTF-8.

Coding

5-point self-assessment scale

The 5-point self-assessment scale was presented to the respondents as follows:

- 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

Translation task marking scale

The marking scale for the translation task is as follows:

- 1: Completely incorrect answer
- 1b: No answer
- 2: In some way partially correct but also incorrect and misleading with regards to the meaning it would provide within a text. Maximum score for partial compound.
- 3: Correct enough that it may help understanding a text. Maximum for a response with the wrong part-of-speech or which seems to result from parsing a compound with the wrong head.
- 4: Not quite correct, but unlikely to impede understanding
- 5: Completely correct

(For further detail on both scales, refer to the 2022 LREC publication.)

CEFR levels

All CEFR levels in the data set are coded as integers, following an extended version of the conventions of YKI. The coding is as follows:

- 1: A1

- 2: A2
- 3: B1
- 4: B2
- 5: C1
- 6: C2
- 7: Native speaker

Languages

Languages are encoded with a 2-character ISO 639-1:2002 language code.

Sessions, dates and times

Responses are divided into sessions. Sessions are designed as streams of consecutive events which include responses, form input events such as key presses, and window focus and blur events. Whenever there is a gap exceeding 5 minutes between events, the session is considered timed out and a new session is created.

All dates are measured as offsets from the date on which first response is received from the participant. For example, if the participant gives their first response on the 1st, all activity on that date is indicated as day 0, and all activity on the 2nd, day 1 and so forth. All dates are recorded in the webserver's time zone, which was UTC (*not* Helsinki time).

All times are measured from the webserver. Only time intervals are given in the dataset. All times are either given in seconds (secs) or microseconds (usecs, millionths of a second). In both cases rounding is to the nearest integer.

Simple format

The simple format has the following columns:

- `participant`: Numerical identifier for the participant (synthetic primary key)
- `word`: The word the participant was asked to self-assess their knowledge of or to translate (always lower case)
- `time_usecs`: The time from the word being sent to the participant until receiving the response in microseconds
- `rating`: The rating on the 5-point self-assessment scale
- `mark`: The mark on the translation task marking scale. This is only available for 1 in 150 words.

Relational format

The relational format has been exported from DuckDB. It can be used most easily be reimporting back into a DuckDB database. Note that DuckDB has good interoperability including with Python & Pandas and R Data Frames. The following command in the same working directory as this file will import the dataset into a DuckDB database called `tallvocabl2fi.duckdb`:

```
$ duckdb tallvocabl2fi.duckdb "IMPORT DATABASE 'relational'"
```

Since the data is stored in TSV files, can also be loaded by any other software supporting TSV[1].

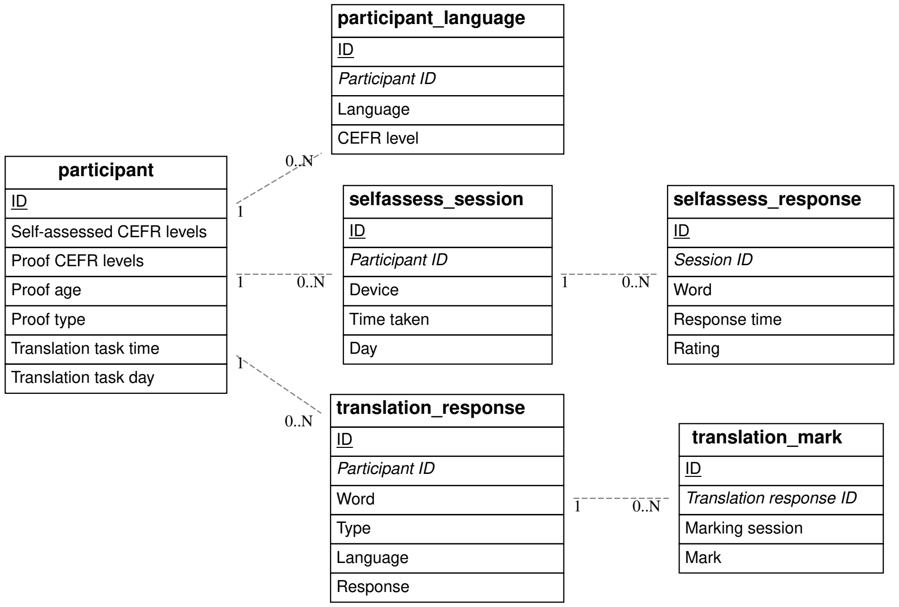

Figure 1: An entity-relationship diagram showing the relationship between the tables, and a selection of their columns.

The schema can be viewed in `erd.svg` and the data types can be found in `schema.sql`, which will be used by DuckDB if you choose to load the data there.

The tables making up the relational format of the dataset are:

- Participant, `participant.csv`
  - Purpose: Gives the basic information about each participant.
  - Columns:
    - `id`: Synthetic primary key
    - `cefr_selfassess_speaking`: Self-assessed speaking CEFR level in Finnish, 1-6
    - `cefr_selfassess_writing`: Self-assessed writing CEFR level in Finnish 1-6

---

[1] Note that the data is in TSV, not CSV even though the files end with `.csv`. This is a quirk of DuckDB's export functionality.

        - `cefr_selfassess_listening_comprehension`: Self-assessed listening comprehension CEFR level in Finnish, 1-6
        - `cefr_selfassess_reading_comprehension`: Self-assessed reading comprehension CEFR level in Finnish, 1-6
        - `cefr_proof_speaking`: Speaking CEFR level in Finnish according to proof document, 1-6
        - `cefr_proof_writing`: Writing CEFR level in Finnish according to proof document, 1-6
        - `cefr_proof_listening_comprehension`: Listening comprehension CEFR level in Finnish according to proof document, 1-6
        - `cefr_proof_reading_comprehension`: Reading comprehension CEFR level in Finnish according to proof document, 1-6
        - `lived_in_finland`: Years lived in Finland as a whole number
        - `proof_age`: Age of proof. Coding:
            - `lt1`: Less than one year, < 1yr
            - `lt3`: Less than three years, < 3yr
            - `lt5`: Less than five years, < 5yr
            - ‘gte5: Greater than or equal to 5 years, >= 5yr
        - `proof_type`: The type of proof. Coding:
            - `yki_intermediate`: The intermediate YKI qualification which confers the levels less than 3, 3 or 4
            - `yki_advanced`: The advanced YKI qualification which confers the levels less than 5, 5 or 6
            - `course_english_degree`: Completion of a course completed as part of an international degree programme taught in English
            - `completed_finnish_upper_secondary`: Completion of upper-secondary school level education in Finnish
            - `completed_finnish_degree`: Completion of a higher or further education qualification in Finnish
            - `other`: Another type of proof
        - `miniexam_time_secs`: The time for the miniexam/translation task to be completed as measured from as the sum of the times of all miniexam sessions. Given in seconds.
        - `miniexam_day`: The day the miniexam was completed on
- Participant language, `participant_language.csv`
    - Purpose: Gives each language known by the participant, and the level at which it is known, including their native language, but not Finnish which is given in the “Participant” table.
    - Columns:
        - `participant_id`: Foreign key to “Participant” table
        - `language`: Language encoded with an 2-character ISO 639 code.
        - `level`: Estimated overall CEFR level, 1-7. Native language is included here as 7.
- Self-assessment session, `selfassess_session.csv`
    - Purpose: Gives information about the sessions in which the self-

    assessment was completed.
  - Columns:
    - `id`: Synthetic primary key
    - `participant_id`: Foreign key to “Participant” table
    - `device`: The device the session was completed on. Coding:
      - `mobile`: Mobile phone
      - `tablet`: Tablet computer
      - `pc`: Personal computer
      - `unknown`: The detection process failed
    - `time_secs`: The time for the session to be completed as measured from the first to last event in the session event stream
    - `day`: The day on which the session started
- Self-assessment response, `selfassess_response.csv`
  - Purpose: Gives the word-level response for each participant
  - Columns:
    - `session_id`: Foreign key to “Self-assessment session” table
    - `word`: The word the participant was asked to self-assess their knowledge of (always lower case)
    - `time_usecs`: The time from the word being sent to the participant until receiving the response in microseconds
    - `rating`: The rating on the 5-point self-assessment scale
- Mini-exam (translation task) mark, `miniexam_mark.csv`
  - Purpose: Gives the mark for participants' responses to the translation task. Please refer to the 2022 LREC publication for the full details of the marking process and the combination of marks.
  - Columns:
    - `miniexam_response_id`: Foreign key to “Mini-exam (translation task) response” table
    - `marker`: The marker/marking session. Coding:
      - `ann1`: The marking from annotator 1
      - `ann2`: The marking from annotator 2
      - `corr`: A corrected/agreed mark between the two annotators (only for some of the disagreeing marks)
      - `final`: A combination of all three of the above to reach a final mark. Given for every response. **Typically only this mark is used.**
    - `mark`: Mark given according to the marking scale
- Mini-exam (translation task) response, `miniexam_response.csv`
  - Purpose: Gives the response of the participant to each word in the translation task
  - Columns:
    - `id`: Synthetic primary key
    - `participant_id`: Foreign key to “Participant” table
    - `word`: Gives the word-level response for each participant
    - `type`: The type of response the respondent opted to give. Coding:
      - `trans_defn`: A translation or definition

        - `topic`: The topic of the word
        - `donotknow`: The respondent specifies that they cannot give any response because they do not know the word at all
    - `lang`: Language of response. Either:
        - `en`: English
        - `fi`: Finnish
        - `hu`: Hungarian (only available for native Hungarian speakers)
        - `ru`: Russian (only available for native Russian speakers)
    - `response`: The plain-text response to the task

Source data

The creation of the word list (see 2022 LREC publication) is based on data sourced from the following other resources:

- Huovilainen, T. (2018). *Psycholinguistic Descriptives* [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2018081601
- Ylönen, T., Wiktionary contributors (2021). *Kaikki.org.* Retrieved from http://kaikki.org/

Publication

Further information is available in the accompanying publication:

Robertson, F., Chang & L., Söyrinki, S. (2022). TallVocabL2Fi: An Extensive Mapping of 15 Finnish L2 Learners' Vocabulary. In *Language Resources and Evaluation Conference* (LREC 2022)

License

This resource is licensed under the CC0 (CC-ZERO) 1.0 license, available at https://creativecommons.org/publicdomain/zero/1.0/ and also included in license.txt

If you make direct use of this resource in academic work, please cite the above publication.