Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6377–6386
Marseille, 20-25 June 2022

TallVocabL2Fi: A Tall Dataset of 15 Finnish L2 Learners' Vocabulary

Frankie Robertson[†], **Li-Hsin Chang**[♣], **Sini Söyrinki**[♠]
[†]Faculty of Information Technology, University of Jyväskylä, Finland
[♣]TurkuNLP group, Department of Computing, Faculty of Technology, University of Turku, Finland
[♠]Centre for Applied Language Studies, University of Jyväskylä, Finland
frankie.r.robertson@jyu.fi, lhchan@utu.fi, sini.t.soyrinki@jyu.fi

Abstract

Previous work concerning measurement of second language learners has tended to focus on the knowledge of small numbers of words, often geared towards measuring vocabulary size. This paper presents a "tall" dataset containing information about a few learners' knowledge of many words, suitable for evaluating Vocabulary Inventory Prediction (VIP) techniques, including those based on Computerised Adaptive Testing (CAT). In comparison to previous comparable datasets, the learners are from varied backgrounds, so as to reduce the risk of overfitting when used for machine learning based VIP. The dataset contains both a self-rating test and a translation test, used to derive a measure of reliability for learner responses. The dataset creation process is documented, and the relationship between variables concerning the participants, such as their completion time, their language ability level, and the triangulated reliability of their self-assessment responses, are analysed. The word list is constructed by taking into account the extensive derivation morphology of Finnish, and infrequent words are included in order to account for explanatory variables beyond word frequency.

Keywords: word knowledge, word response data, mental lexicon, Finnish, learner data

1. Introduction

This paper presents TallVocabL2Fi (Robertson, 2022), a word knowledge response dataset collected from 15 paid L2 Finnish participants. The dataset triangulates word knowledge between a "tall" self-rating test of 12 000 words and a smaller translation test of 100 words, chosen based on the results of the self-rating test. The self-rating test consists of participants giving their own assessment on a 5-point scale of how well they know a given prompt word. The full raw response data is deposited in the Language Bank of Finland and made available under the CC0 combination public domain dedication and license.

There is emerging interest in the field of Computational Linguistics in the task of Vocabulary Inventory Prediction (VIP). In the VIP task, the input is a sample of a learner's vocabulary typically as known/unknown responses to a relatively small set of words, and the output is a function which can predict known/unknown for any input word. However, thus far, there is very little data which can be used for evaluation of the task. To evaluate VIP, we need as much information about the vocabulary of each participant as possible. The first reason for this is the validity of the evaluation metrics. Having many words is not only likely to increase statistical power, but also to ensure inclusion of a wide variety of words, rather than selecting only a few benchmark words, which could bias the evaluation. The second reason is that a large vocabulary allows for simulation of VIP methods based on active learning, referred to as Computerised Adaptive Testing (CAT) in this context. Simulation of CAT-based VIP refers to replacing the query step of the CAT, which would normally solicit a response directly from the testee with a retrieval of a response from a tall vocabulary dataset such as TallVocabL2Fi. Since this retrieval is limited to words actually in the vocabulary dataset, the quality of the simulation hinges on the dataset's lexical coverage. The Finnish language, with its large derivational morphology and word compounding, provides a rich target for VIP research. Such word formation processes provoke an argument to move beyond the frame of vocabulary size prediction familiar from Second Language Vocabulary Acquisition (SLVA). Consider a learner who knows the Finnish word 'isä' (English: *father*) and also the Finnish genitive morpheme '-n' (English: *'s*), then by extension they know also 'isänisä' (English: *grandfather*) and 'isänisänisä' (English: *great grandfather*), and so on. Thus the learner's vocabulary size becomes infinite by induction, making comparison between learners' vocabulary size meaningless. In the VIP frame we learn a classifier which, given a sample of learner responses, produces a prediction for whether any given word is in their vocabulary, so that we do not need to restrict ourselves to a closed word list.

2. Related Work

In this section, we first review studies where word knowledge data has been collected, paying particular attention to open data, but also commenting upon the fields of research which deal with vocabulary data. Next, we look at previous work in VIP from the perspective of their usage of different datasets.

2.1. Vocabulary Knowledge Datasets

The starting point for the design of TallVocabL2Fi is the dataset of Ehara et al. (2010) (Ehara, Y., 2009). Beyond changing the target language from English to Finnish, three notable divergences are made here with the aim of creating a complimentary resource: 1. a

translation test is added to triangulate the results of the self-assessment test 2. rather than coming from a word list published for language learners[1], and therefore containing mainly high-frequency words, low-frequency words are also included here to evaluate predictions of low-frequency word knowledge 3. the participants are more diverse when compared with Ehara's dataset, in which all respondents were University of Tokyo students with 13 out of 15 having Japanese as an L1. In this study, the language background of the participants is balanced between three different L1 languages and the data was collected online as as to avoid undesirable correlations arising from selecting participants living in a single place, e.g. participants having the same language teacher. As well improving VIP evaluation, this helps to avoid overfitting on a small subpopulation when using the dataset as training data, as in Robertson (2021), when models trained on Ehara's dataset did not generalise well to other Japanese L2 English datasets.

Within the field of SLVA, apart from studies which aim to validate vocabulary tests themselves, most studies using a vocabulary test, e.g. a version of the VLT (Vocabulary Levels Test) (Nation, 1983; Schmitt et al., 2001; Webb et al., 2017) are aiming to estimate so-called vocabulary size, a single dimension of vocabulary knowledge. While response-level data is generated by these studies, it is most often seen as a means-to-an-end rather than representing useful or interesting data in and of itself. This may explain why despite the high volume of data collected within the field of SLVA of as of writing, very little — if any — has been released openly. To demonstrate this, we randomly sampled 20 publications form recent years where vocabulary knowledge data had been collected from the Vocabulary Acquisition Research Group Archive (VARGA) (Meara, 2021), representing about 20% of recent yearly output. Of these, none appears to have deposited their data in a repository, as per the FAIR principles (Wilkinson et al., 2016) which emphasises data reusability. Only one, that of Guan and Fraundorf (2020), mentions that "datasets generated for this study are available on request to the corresponding author".

Word recognition tasks have long been of interest in Psycholinguistics. An archetypal example is the task of lexical decision. In this task, participants are shown a real word or a pseudo word: a prompt word designed to appear to be an admissible word in the target language, while not actually being so. They must then choose whether they believe it to be a true word or not. As well as their response, the response time is also often assessed. From a vocabulary perspective, this type of data seems to only test knowledge of the words' surface forms, however psycholinguists do tend to be interested in response and item level aspects of this data since it is seen to be related to issues such as memory and cognitive processing efficiency. The Center for Reading Research at the University of Ghent maintains a list of so-called megastudies, including many which are released as open data[2]. Of the lexical decision tasks listed, response level data is available for roughly half. The largest, such as the Dutch Crowdsourcing Project (Brysbaert, 2019), English Crowdsourcing Project (Mandera, 2019), and SPALEX (Aguasvivas, 2018) employ crowdsourcing techniques, where with a relatively uncontrolled set up where responses are collected from many unpaid participants.

In terms of data gathered from Finnish L2 learners, Salmela et al. (2021) gathered data from 117 L1 and 159 L2 Finnish speakers through crowdsourcing as part of the development of their vocabulary size measurement instrument, Lexize. Their approach largely follows the LexTALE (Lemhöfer and Broersma, 2012) instrument, in which a score is created based on a lexical decision task, but approached from a language acquisition angle, emphasising its usage as a vocabulary size estimation tool. The data used in the study has not be deposited, but is available "upon reasonable request". Salmela et al. (2021) also reviews previous empirical vocabulary research of Finnish.

2.2. Vocabulary Inventory Prediction

The most obvious application of TallVocabL2Fi is evaluation of Vocabulary Inventory Prediction (VIP). In this emerging task, a relatively small word knowledge sample is taken from a learner, which is then used to train a classifier which can predict whether a learner knows any given word. Reports of systems approaching this tasks still vary widely in the exact way the data is split up and the results are evaluated.

Avdiu et al. (2019) approach VIP through feature engineering, benchmarking on the dataset of Ehara et al. (2010) (Ehara, Y., 2009). They associate learners with a mixture of different genre-specific subcorpora of the commercial COCA corpus (Davis, 2020) according to the agreement between their training responses and the frequency distribution of the subcorpora. Next, predictions are made from the mixture of the subcorpora, together with their frequency distributions, as well as other features. The evaluation frame is similar to missing data imputation, where a large section of the data is used for a single round of training, and the rest is predicted based upon this.

Ehara (2019) approaches VIP from the perspective of 1-parameter Item Response Theory (IRT) using the dataset of of Ehara et al. (2010) (Ehara, Y., 2009). Despite the IRT perspective, they instead create an equivalent neural network mapping pretrained Glove (Pennington et al., 2014) word embeddings to a single dimension of word difficulty. As with (Avdiu et al., 2019), a single stage of training was performed so that the ability of the learners was learnt simultaneously with the weights of the prediction network. Robertson

[1]Ehara's word list comes from 究極の英単語 *(Ultimate English Words) SVL 12000* published by ALC Press.

[2]http://crr.ugent.be/programs-data/megastudy-data-available

(2021) also uses the data of Ehara et al. (2010) (Ehara, Y., 2009), but instead fits a 2-parameter logistic IRT model as an initial step, then uses this as training data for a neural network regressor, which represents words using ConceptNet NumberBatch embeddings (Speer et al., 2019). A CAT (Computerised Adaptive Testing) setting was also investigated. While a word frequency baseline was beaten, results did not to generalise to other datasets such as EVKD1 (Ehara, Y., 2018).

3. Resource Design & Creation

TallVocabL2Fi is designed to contain many items in the vocabulary inventory of a few learners. Thus, it is a "tall" dataset, in comparison to a dataset which contains information about less words from more people (a "wide" dataset). The main sampling considerations for the dataset are the selection of words and participants. The measurement considerations are the scale with which participants self-assess their knowledge of the word items, and the response types possible in the translation test and how they are marked.

3.1. Word List

Although the word list was constructed with the aim of avoiding excessive bias towards particular vocabulary, nevertheless some words were systemically excluded to attempt to increase the validity of the responses and make efficient use of the word budget of 12 000 words. The procedure is an approximate one, and some of the filtering stages may remove a few useful words along with removing many less useful ones. As a first step, only content words were included, here meaning those from open Parts-Of-Speech (POSs). Within the Universal Part-Of-Speech (UPOS) specifically this means noun, verb, adjective, adverb or adposition[3]. Other POSs can be highly grammatical and contextually dependent, potentially causing uncertainty and incorrect responses in participants. Another major consideration is from the point of view of word knowledge, responses to *purely compositional* compounds and derived words provide redundant information. We learn no new information from these items, except that the learner knows all of the compound's constituent words. In order to avoid wasting too many items on these words, the list is therefore trimmed of highly compositional words.

In terms of frequency, it would seem natural to give priority to high frequency words, since they provide better value to language learners. However, we also need to give some coverage also to very low frequency words. The reason for this is to avoid circularity. While it is known that frequency correlates with knowledge, basing our sampling entirely on high frequency words according to some background corpus may cause systematic bias which limits analysis of variance in knowledge of low frequency words. This extra variance could arise from, for example how difficult certain word forms and meanings are to learn, as well as the difference between the frequency distributions of our background corpora and frequencies distributions of learners' comprehensible input.

The word list construction starts from the frequency list of Huovilainen (Huovilainen, 2018), which includes lemma frequencies from six Finnish language corpora spanning different registers such as text from a web forum, news website, subtitles, and an online encyclopedia derived by running the Turku Universal Dependencies pipeline (Kanerva et al., 2018; Kanerva et al., 2020) over the corpora. After the list was filtered by UPOS, POS information was removed, and lemmas occurring with a different POS had their occurrence counts summed. Next, lemmas were filtered by their inclusion in The New Finnish Word List (Institute for the Languages of Finland, 2011) published by the Finnish language body. After this, ordinal and proper noun like word such as "Days of the week", "Months", "Languages" and "Countries" were removed according to the structured data from Kaikki.org (Ylonen, 2021), ultimately derived from Wiktionary to obtain the master list. The justification for removing these words starts from the understanding that some words have to be removed. Since ordinal-like words are often learnt by-rote as a single unit, they provide less useful information in the sense of each extra word within each class beyond the first having very low surprisal. Language names, on the other hand, are very often loans in Finnish and so do not tend to give such useful information about Finnish knowledge.

At this point, a second trimmed word list is created from the master list. The filtered word list removes unadapted loans, here defined as any word with an English entry on Kaikki.org. Only English homographs were filtered for a few reasons. Firstly, apart from Finnish, English is the only language that all participants had exposure to. Furthermore, including more languages would increase the number of semantically unrelated homographs, i.e. false positives. Finally, in practice looking at only English gives good coverage for Finnish since it is one of the biggest sources of unadapted loans in Finnish, and the list of Finnish-English unadapted loans also includes many interlingual homographs which are borrowed in many languages, e.g. "pasta". The filtered word list also removes any word with a dash or determined to be a suffix according to Kaikki.org. Again, the justification for removing dash starts from acknowledging some words have to be removed to fit within the word budget. Dashed compounds are relatively rare in Finnish, and are generally either inserted to split unadapted loans from Finnish words or to break up character combinations which would otherwise form a diphthong. In the former case, having the word in the list is somewhat undesirable, but

[3]The inclusion of adpositions may be surprising since English prepositions are a closed class, but as Huumo (2021) explains: "Finnish adpositions are a semi-open class including items which are grammaticalized to a greater or lesser extent".

in other cases the compound is visually unusual enough that it is removed in case it causes learners to accidentally answer less accurately.

Derived words with suffixes determined to be highly productive and compositional are dropped at this stage, and frequencies redistributed to the head. The suffixes are -ja, -sti, -ton, -uus, -mpi, -in and -nen. Derivations were found based on OMorFi (Pirinen, 2015) as well as data derived from etymology information from English Wiktionary dumps using wikiparse (Robertson, 2020), described by Robertson (2020).

Compounds are then trimmed according to their compositionality. The compositionality of derived words is defined following Cordeiro et al. (2019). In this framework, if we have a word which can be split into parts $W = w_1, w_2, ..., w_n$, then we can define a measure of compositionality as cosine similarity between a word vector corresponding to W and some aggregation of its parts w_i. The $\mathrm{pc}_{\mathrm{uniform}}$ measure is used, which was robust for noun compound compositionality across different languages according to the results of Cordeiro et al. (2019). In this measure, each compound part is weighted equally, so that compositionality is defined as $\mathrm{c}(W)$:

$$\mathrm{c}(W) = \mathrm{sim}_{\cos}\left(\mathbf{v}(W), \sum_{i=1}^{n} \frac{\mathbf{v}(w_i)}{\|\mathbf{v}(w_i)\|}\right)$$

Where $\mathbf{v}$ denotes lookup from a word embedding space, in this case ConceptNet NumberBatch 19.08 (Speer et al., 2019), described by (Speer et al., 2017). Here, the derived items include compounds and derived words. Words with a compositionality of more than 0.9 were dropped, while those with a value of less than 0.7 were kept. Words with a compositionality between 0.7 and 0.9 were kept with probability linearly decreasing from 1 to 0 along this interval. These values were determined by inspection of a small number of random words and their compositionality values.

There are now two word lists, the master list and the trimmed list. The final word list is then constructed. First, we ensure there are high frequency words in the output list. First we sort the master list by frequencies from each subcorpus (e.g. only from forums, only from subtitles), make sure each time the top 2000 words are included in the output list. Next we do the same for the trimmed list.

The next step is to include words uniformly from different frequencies. We divide the words from each subcorpus up into frequency bands. To begin with, each subcorpus is divided into frequency buckets 0.1 zipfs[4] wide starting from 2 zipfs. As well as the subcorpora of Huovilainen there is a "virtual" subcorpus made from the minimum frequency across all subcorpora to ensure that words that are rare across all subcorpora are included. We then consider the set of all frequency buckets across all subcorpora together. The final word list is then constructed by following a greedy process of iteratively taking a random word from the frequency bucket with the least words included in the current state of the word list until we reach 12,000 words.

[4]The zipf scale is defined based on the word frequency f by Van Heuven et al. (2014) in logarithmic frequency space as $\log_{10} 10^9 f$.

3.2. Self-Assessment Test

The format of the self-assessment portion of the dataset follows Ehara et al. (2010). This scale synthesises the well known Vocabulary Knowledge Scales (VKS) test (Wesche and Paribakht, 1996) and Dale's (1965) scale into a 5-point scale, which unlike either of the aforementioned, does not include any free text answers to check the vocabulary knowledge directly. In comparison, in the VKS, a respondent answering 5 is expected to define the word. Excluding such a requirement is more appropriate when collecting "tall" data since each item can be responded to with a single interaction, potentially allowing many items to be completed in a limited amount of time.

The scale is reproduced below:

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

Wolter (2005), summarised by Milton (2009), noted that although the VKS was designed as a test of vocabulary depth rather than one of breadth, it does not tackle many aspects of vocabulary depth. The perspective taken here is that the aim is primarily to measure vocabulary breadth. The use of the ordinal scale is more a device to account for a degree of uncertainly from the learner, rather than potentially losing information by forcing a yes/no answer, and to take into account a single, narrow aspect of depth: having encountered a word's form versus knowing its meaning. It is worth noting, that there is some scope for different interpretation of the scale and in practice people do use the scale differently. For this reason the scale is calibrated with a translation test.

3.3. Translation Test

After completing the self-assessment, participants were asked to complete a translation test, with words chosen based upon the participants' response to the self-assessment test. For each participant, a word list was created, by first grouping words into 5 buckets according to the participant's response on the 5-point scale. Next, 20 words were selected at random from each

bucket to make a total of 100 words. These words were then presented in a random order.
For each item, participants had the possibility of giving one of the following types of answers:

1. Translate/define the word
2. Give the topic of the word i.e. completing the sentence "this word has something to do with..."
3. Saying they don't know the word at all

For either translation or topic answers, answers could be given in the learners' native language, or English if it's not their native language, or in Finnish. Responses were then marked on a 5-point scale:

1a. 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

3.4. Participant Selection

Each participant was offered €200 upon successful completion of study. Given the available budget, this allowed for 15 paid participants, the same as in Ehara et al. (2010). In the first phase, expressions of interest were gathered through a Google Form. These responses enabled identifying L1 languages with enough speakers of various Finnish proficiency levels interested in participation. Participants gave their proficiency level by self-rating their reading according to the Common European Framework of Reference for Languages (CEFR) self-assessment grid[5]. Upon acceptance into the study, participants had to additionally provide proof in the form of a language certificate that they are at least at a B1 level.
The first attempt at soliciting expressions of interest was to contact Finnish as a second language teachers in various educational facilities via email. However, out of an initial five emails, no responses were received. Instead, potential participants were reached directly through Facebook groups such as "Foreigners in Finland", "Foreigners in Jyväskylä", and "International Working Women of Finland". Once the native languages of interest were chosen, the study was advertised in more specific groups, e.g. announcing the study in "Brits in Finland" to find native English speakers.
Due to the small number of participants, careful selection was performed to capture both diversities and similarities between learners. As far as was possible given the available pool of candidate participants, stratified sampling was performed in 2-dimensions: across 3 groups of native languages and 4 CEFR levels. These were English, Russian, and Hungarian, and B1, B2 and C1 and C2, respectively. Being such a high level, C2 was given less weight. Participants from the targeted groups were invited into the study initially in the same order in which they responded. Later, when inactive participants were removed from the study, replacements were found by looking at the newest expressions of interest first, with the hope that these were more likely to become active. Further detail about those expressing interest versus completing the study is given in Section 4.3. Table 1 shows the make-up of the participants who completed the study.

	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

Table 1: Breakdown of participants by L1 and self-reported reading CEFR level

3.5. Collection Process

The final instrument was delivered remotely via a custom web application written in Python with the Quart web framework and the htmlx client-side framework. Before the participant gave any responses, they were given a description of the study including the rating scales and the expected length of the study, along with advice on how to schedule self-rating sessions to complete within the deadline. Participants agreed to answer both parts honestly and to the best of their ability, as well as to avoid as far as possible any deliberate Finnish vocabulary learning during the test period. Participants were given a three-week deadline, also in order to try and control the amount of Finnish they could learn in this time. This deadline was extended in a few cases due to reasons such as website downtime, or in cases where someone was almost finished in time, but had fallen slightly behind.
Since all participants had some level of English skills, the website itself, as well as all correspondence and instructions were presented in English.
In the self-assessment stage, a word would be shown on screen, and the participant could press a number 1-5 either using a button or their keyboard. To help participants pace themselves, there was the option of choosing a fixed batch of self-assessment words to answer, after which the website returns the user to the homepage. The participants are then instructed to complete the 100-word translation test described in 3.3 in a single session.

[5] https://www.coe.int/en/web/portfolio/self-assessment-grid

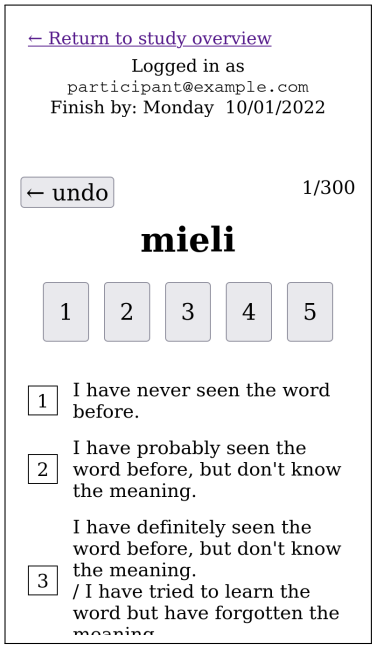

Figure 1: Screenshot of self-assessment stage as seen by a participant using a mobile phone

3.6. Marking of the Translation Test

Two markers marked the translation responses. The first was a native Finnish speaker, with B2 English and A1 Hungarian and no Russian, and the second was a native English speaker with L2 Finnish at the B1 level and no Hungarian or Russian. It may be that marks from a native Hungarian or Russian speaker would better reflect the respondent's knowledge level, but it is hoped that the usage of multiple markers mitigates this. In addition, the raw response data is included in the released resource so it is possible for any user of the resource to check or correct the marks.

In order to mark the responses given their language skills, bilingual dictionaries were used in cases where the markers could not otherwise make a decision. The first marker used the MOT English-Finnish, Russian-Finnish and Hungarian-Finnish dictionaries, while the second used entries on English Wiktionary in combination with the emMorph Hungarian morphological analyser (Novák, 2018) described in Novák et al. (2016), falling back to examples parallel sentences retrieved using Glosbe[6] and then Google Translate[7], pivoting via English when possible.

Initially, the first two markers marked the responses separately. At this point, inter-annotator agreement was measured. In terms of Cohen's Kappa this is 0.62, giving rather low value which nevertheless indicates moderate agreement. On the other hand the Quadratic Weighted Kappa value is rather higher at 0.86. Figure 2 shows a confusion matrix between the two markers.

Next, the two markers conferred on items with a score

[6] https://glosbe.com/

[7] https://translate.google.com/

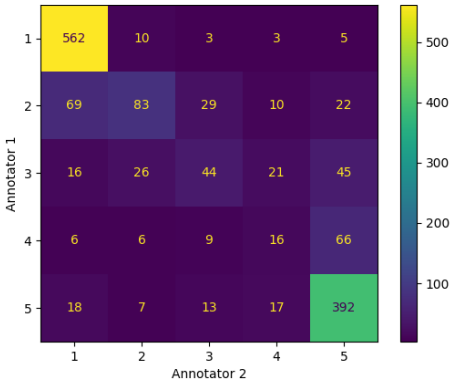

Figure 2: Confusion matrix of marks between marker 1 and marker 2 summed across all participants

difference of 2 or more: 96 out of 401 disagreeing items (out of 1500 items in total). Each item was discussed with reference to the scale until the markers reached a consensus. One common source of disagreements was words where participant had interpreted a word as a word form or a proper noun, even though these were not included in the word list e.g. "joensuu" (prompted all lowercase rather than in correct proper noun case) is a Finnish city, but it also literally means "river's mouth". The markers had made different decisions on these cases, but it was agreed that since this was not explained to the learners in advance, any correct word form or proper noun answers should be accepted. For a future resource, these ambiguous cases could be filtered from the word list. Some other disagreements were due to rare translation responses which were not included in any dictionaries, which were nevertheless resolved using Glosbe or Google Translate.

After a common decision was reached on all 96 extreme disagreements, the final mark was created by using these resolved marks when possible, otherwise taking the minimum from the other two marks. All marks, including the initial marks before resolution are included in the released resource.

4. Evaluation & Enrichment

In this section, we present some data descriptive statistics drawn from the resource. Here, data is grouped by participant, but due to the low number of participants (15), the trends reported here are more qualitative than quantitative, representing mainly the individuals who participated in the study rather than the populations they are drawn from. This information is primarily likely to be useful for anyone intending to perform similar data collection themselves.

Each participant is given a reliability index based on the agreement of their self-assessment and translation test, which is included in the dataset. We inspect correlations, such as with completion time, so as to try and determine whether lower quality answers may be due to

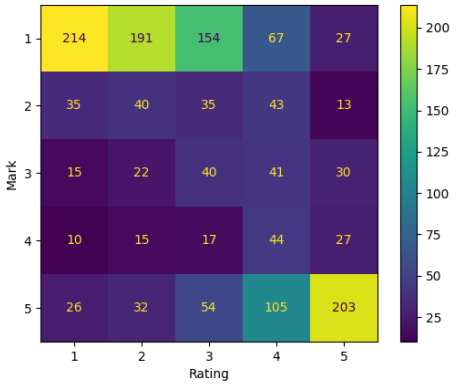

Figure 3: Confusion matrix of self-rating versus mark

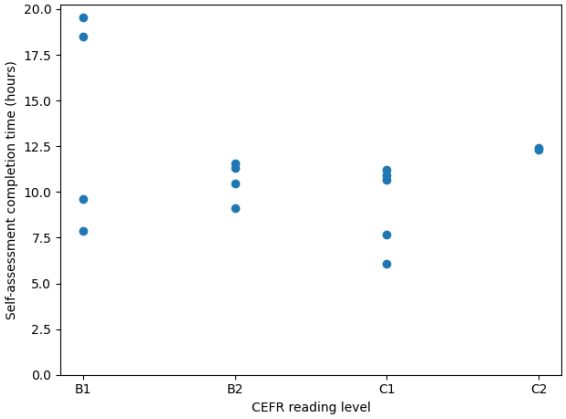

Figure 4: Completion time varying with CEFR level

some participants rushing through the self-assessment versus other who took a more deliberative approach.

4.1. Participant Reliability

Figure 3 gives a confusion matrix between self-ratings and the the final marks for all participants combined. We would like to quantify the reliability of individual participants on the basis of the agreement between participant's responses. To this end we construct several summary statistics of the confusion matrix. Reliability correspond to the true positive rate of a participant's self-rating procedure. Underrating on the other hand, corresponds to the false negative rate. The partial variants set a lower mark threshold for an item on the translation test to be considered correct.

$$\begin{aligned}\text{reliability} &= P(rating \geq 5, mark \geq 4)\\ \text{partial-reliability} &= P(rating \geq 5, mark \geq 2)\\ \text{underrating} &= P(rating \leq 3, mark \geq 4)\\ \text{partial-underrating} &= P(rating \leq 3, mark \geq 2)\end{aligned}$$

The final balanced reliability measure used for the rest of this paper summarises reliability and underrating similarly to balanced accuracy:

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

4.2. Relationship Between Completion Time, Language Level and Reliability

Questions arise to whether there is any systematic relationship between participants' CEFR level, the time they spent completing the self-assessment, and the reliability of their responses. In particular 1) if any group is systematically less reliable, this is interesting because they might be eliminated from any future data collection, 2) groups that complete the test quickly can be given more words. Observe Figures 4, 5, and 6. It is fairly apparent that in terms of completion, there is a cluster at 10±3 with 1 outlier at around 6 hours and 2 at around 20 hours.

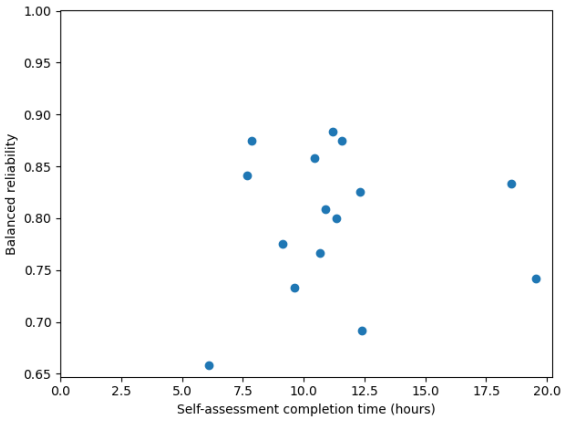

Figure 5: Reliability varying with completion time

Overall there does not seem to be any obvious correlations, however, outliers do occur in extreme situations. For example, the two participants who took around 20 hours to complete were both at the low end of ability being B1, and the least reliable responses were received from the participant who completed the assessment most quickly.

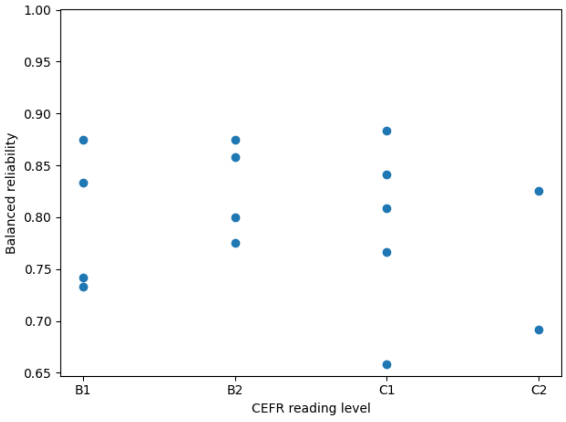

Figure 6: Reliability varying with to CEFR level

4.3. Participant Experience

Overall, out of 154 expressions of interest, 29 were invited to participate in the study. Out of these 29, 11 did not start, while 3 dropped out part way through after responding to 2314, 2600, and 2031 words from the self-assessment test in 90 minutes to 4 hours. It looks like this is a critical point for drop-offs, but it is not clear how to either encourage completion at this stage or discourage participation altogether from this group – who all self-rated their reading comprehension at a B2 level.

At the end of the study, participants were invited to give feedback with two prompts. The first concerned their experience completing the self-assessment, and whether they thought they were able to concentrate and answer accurately. The second concerned their experience completing the translation test, namely whether they thought it was fair and reflected their abilities.

Two respondents noted being confused by repeating words at the self-assessment stage. While there were no repeating words, but there were some very similar words. One possibility for future word list compilation would therefore be to make sure not to pick too many words from the same word family, or too many compounds using the same parts.

Multiple people noted high satisfaction with the possibility of completing the self-assessment during time which might otherwise be spent waiting such as while waiting for a bus, or while on a ferry. Replicating this in a country without ubiquitous mobile internet would require the data collection process to work offline.

Multiple people reported finding the translation exam to be quite difficult, or even demoralising. From the perspective of producing valid data for TallVocabL2Fi, this is not necessarily a problem. Given also that the participants were compensated relatively well, frustration comparable to, for example, working at an office job during part of the process does not seem to pose ethical problems. It should be noted however that 60% of the words on the translation test were ones which the learner had said they did not know. It may be that it more resolution is needed at the higher end of the self-rating scale, since this is the part which really concerns word knowledge. The best balance to use for selecting the words for the translation test could be revisited in future work.

Another possibility which could both improve participant experience and possibly increase the richness and validity of the resulting resource is to add more triangulation tasks. These tasks could in principle use almost any vocabulary measurement device, including those covered by Milton (2009), for example, free production, cloze (gap filling) and word association. To further improve things, sections from each task could be alternated and overall session length limited to control for cognitive load, fatigue, and boredom.

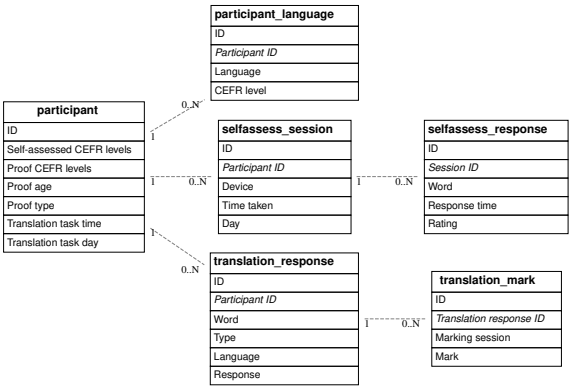

Figure 7: An entity-relationship diagram of TallVocabL2Fi's tables and a selection of their columns

5. Resource Description

The resource is primarily made available as a series of TSV (Tab-Separated Values) files. These have been exported from the DuckDB[8] (Raasveldt and Mühleisen, 2020) relational-analytical database. The recommended way to use the resource is to reimport it into DuckDB since DuckDB has support for joins and analytical queries and can interoperate with Pandas and R. However, the resource could also be imported into spreadsheet software.

Figure 7 shows an entity-relation diagram with the schema of the database. Noteworthy is that relatively detailed data is preserved, including for example the division of participants' responses into the sessions in which they gave them.

6. Conclusion

We have presented TallVocabL2Fi, a dataset consisting of word knowledge responses from 15 L2 Finnish participants. Each participant has rated their knowledge of 12,000 words, and given their response to a translation test of 100 words. Although primarily intended for Vocabulary Inventory Prediction, we hope this dataset can also be of use to SLVA researchers. One way in which this could manifest is as part of a larger combined dataset for comparison between groups, or as data for developing or teaching quantitative methods for vocabulary acquisition.

The resulting language resource (Robertson, 2022) has been deposited in The Language Bank of Finland, and it is licensed under the liberal CC0 license to encourage maximal reuse. The release includes a "readme" file with a detailed description of all tables and columns and the coding of different data types. It is cross-referenced in the IRIS Database and the LREC Language Resource Map. The source code for data collection website is also made available under the Apache v2 license[9].

[8] https://duckdb.org/

[9] Available at http://github.com/frankier/finnvocabcollect/

Bibliographical References

Avdiu, D., Bui, V., and Klimčíková, K. P. (2019). Predicting learner knowledge of individual words using machine learning. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, pages 1–9, Turku, Finland, September. LiU Electronic Press.

Cordeiro, S., Villavicencio, A., Idiart, M., and Ramisch, C. (2019). Unsupervised Compositionality Prediction of Nominal Compounds. *Computational Linguistics*, 45(1):1–57, 03.

Dale, E. (1965). Vocabulary measurement: Techniques and major findings. *Elementary English*, 42(8):895–948.

Ehara, Y., Shimizu, N., Ninomiya, T., and Nakagawa, H. (2010). Personalized reading support for second-language web documents by collective intelligence. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10, page 51–60, New York, NY, USA. Association for Computing Machinery.

Ehara, Y. (2019). Neural rasch model: How do word embeddings adjust word difficulty? In *PACLING*.

Guan, C. Q. and Fraundorf, S. H. (2020). Cross-linguistic word recognition development among chinese children: A multilevel linear mixed-effects modeling approach. *Frontiers in psychology*, 11:544.

Huumo, T. (2021). On the gradable nature of the search domain: A study of degree modifiers and the scalar semantics of finnish spatial grams. *Cognitive Semantics*.

Kanerva, J., Ginter, F., Miekka, N., Leino, A., and Salakoski, T. (2018). Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics.

Kanerva, J., Ginter, F., and Salakoski, T. (2020). Universal lemmatizer: A sequence to sequence model for lemmatizing universal dependencies treebanks. *Natural Language Engineering*, pages 1–30.

Lemhöfer, K. and Broersma, M. (2012). Introducing lextale: A quick and valid lexical test for advanced learners of english. *Behavior research methods*, 44(2):325–343.

Meara, P. (2021). Vocabulary acquisition research group archive.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Matters.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, 5:12–25.

Novák, A., Siklósi, B., and Oravecz, C. (2016). A new integrated open-source morphological analyzer for hungarian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1315–1322.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*.

Pirinen, T. A. (2015). Development and use of computational morphology of finnish in the open source and open science era: Notes on experiences with omorfi development. *SKY Journal of Linguistics*, 28:381–393.

Raasveldt, M. and Mühleisen, H. (2020). Data management for data science-towards embedded analytics. In *CIDR*.

Robertson, F. (2020). Word sense disambiguation for finnish with an application to language learning. Master's thesis, University of Jyväskylä.

Robertson, F. (2021). Word discriminations for vocabulary inventory prediction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1188–1195, Held Online, September. INCOMA Ltd.

Salmela, R., Lehtonen, M., Garusi, S., and Bertram, R. (2021). Lexize: A test to quickly assess vocabulary knowledge in finnish. *Scandinavian journal of psychology*.

Schmitt, N., Schmitt, D., and Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language testing*, 18(1):55–88.

Speer, R., Chin, J., and Havasi, C. (2017). ConceptNet 5.5: An open multilingual graph of general knowledge. In *AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Van Heuven, W. J., Mandera, P., Keuleers, E., and Brysbaert, M. (2014). Subtlex-uk: A new and improved word frequency database for british english. *Quarterly journal of experimental psychology*, 67(6):1176–1190.

Webb, S., Sasao, Y., and Ballance, O. J. (2017). The updated vocabulary levels test. *ITL – International Journal of Applied Linguistics*, 168:33–69.

Wesche, M. and Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, 53(1):13–40.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. O. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S. C., Evelo, C. T. A., Finkers, R., González-Beltrán, A. N., Gray, A. J. G., Groth, P., Goble, C. A., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R. W. W., Kuhn, T., Kok, R. G., Kok, J. N., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R. C., Sansone, S.-A., Schultes, E. A., Sengstag, T., Slater, T., Strawn, G. O., Swertz, M. A., Thompson, M., van der Lei, J.,

van Mulligen, E. M., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3.

Wolter, B. (2005). *V-Links: A new approach to assessing depth of word knowledge*. Ph.D. thesis, University of Wales Swansea.

Language Resource References

Aguasvivas, J. (2018). *SPALEX*. Figshare. Published at https://figshare.com/projects/SPALEX/29722.

Brysbaert, M., Mandera, P. (2019). *Dutch Crowdsourcing Project*. Center for Open Science. Published at https://osf.io/5fk8d/.

Davis, M. (2020). *Corpus of Contemporary American English*. Retrieved from https://www.english-corpora.org/coca/.

Ehara, Y. (2009). *ESL Vocabulary Dataset*. Published on personal website at http://yoehara.com/vocabulary-prediction/.

Ehara, Y. (2018). *EVKD1 Dataset*. Published on personal website at http://yoehara.com/evkd1/.

Huovilainen, T. (2018). *Psycholinguistic Descriptives*. The Language Bank of Finland. Published at http://urn.fi/urn:nbn:fi:lb-2018081601.

Institute for the Languages of Finland. (2011). *Modern Finnish Word List*. Institute for the Languages of Finland. Published at http://urn.fi/urn:nbn:fi:lb-2021092006.

Mandera, P., Keuleers, E., Brysbaert, M. (2019). *English Crowdsourcing Project*. Center for Open Science. Published at https://osf.io/rpx87/.

Novák, A. (2018). *emMorph*. GitHub. Published at https://github.com/nytud/emMorph.

Robertson, F. (2020). *wikiparse*. GitHub. Published at https://github.com/frankier/wikiparse.

Robertson, F. (2022). *TallVocabL2Fi: Measurements of 15 L2 Finnish learners' vocabularies*. The Language Bank of Finland. Published at http://urn.fi/urn:nbn:fi:lb-2022041921.

Speer, R. and Chin, J. and Havasi, C. (2019). *ConceptNet Numberbatch 19.08*. Luminoso Technologies. Published at https://github.com/commonsense/conceptnet-numberbatch.

Ylonen, T. (2021). *Kaikki.org*. Retrieved from https://kaikki.org/.