Language Testing in Asia

RESEARCH Open Access

An intelligent vocabulary size measurement method for second language learner

Tian Xia[1†], Xuemin Chen[2†], Hamid R. Parsaei[3†] and Feng Qiu[4*†]

†All authors contributed equally to this work.

*Correspondence: qxf@shnu.edu.cn

[1] School of Computer and Information Engineering, Shanghai Polytechnic University, 2360 Jinhai Road, Shanghai 201209, China
[2] Department of Engineering, Texas Southern University, 3100 Cleburne St, Houston 77004, TX, USA
[3] Department of Industrial & Systems Engineering, Texas A & M University, 750 Agronomy Road, College Station 77843, TX, USA
[4] Institute of Artificial Intelligence on Education, Shanghai Normal University, 100 Guilin Road, Shanghai 200233, China

Abstract

This paper presents a new method for accurately measuring the vocabulary size of second language (L2) learners. Traditional vocabulary size tests (VSTs) are limited in capturing a tester's vocabulary and are often population-specific. To overcome these issues, we propose an intelligent vocabulary size measurement method that utilizes massive robot testers. They are equipped with randomized and word-frequency-based vocabularies to simulate L2 learners' variant vocabularies. An intelligent vocabulary size test (IVST) is developed to precisely measure vocabulary size for any population. The robot testers "take" the IVST, which dynamically generates quizzes with varying levels of difficulty adapted to the estimated tester's vocabulary size in real-time using an artificial neural network (ANN) through iterative learning. The effectiveness of the IVST is factually verified by their visible vocabularies. Additionally, we apply a long short-term memory (LSTM) model to further enhance the method's performance. The proposed method has demonstrated high reliability and effectiveness, achieving accuracies of 98.47% for the IVST and 99.87% for the IVST with LSTM. This novel approach provides a more precise and reliable method for measuring vocabulary size in L2 learners compared to traditional VSTs, offering potential benefits to language learners and educators.

Keywords: Vocabulary size test, Computerized adaptive testing, Intelligent vocabulary size measurement, Artificial neural network, Long short-term memory, Robot testers

Introduction

Vocabulary size is essential for developing second language (L2) skills, as it accounts for 72%, 52%, and 39% of the tester's scores above average in reading, listening, and writing tests, respectively (Alahmadi & Foltz, 2020). Speaking proficiency is also known to be significantly impacted by vocabulary knowledge (Uchihara & Clenton, 2022), and vocabulary size is a critical predictor for L2 knowledge development (Enayat & Derakhshan, 2021). Vocabulary expansion affects different learning aspects, such as informal conversations, spoken proficiency, and academic comprehension (Masrai & Milton, 2021). Making a precise L2 vocabulary size measurement for L2 learners is essential for individualized instruction, customized pedagogy design, and vocabulary learning tools development (Hsu & Ou Yang, 2013).

In addition, vocabulary size strongly correlates with the foundation of L2 learning, such as phonological representations (Georgiou et al., 2020), morphosyntactic regularities (Mahr & Edwards, 2018), and even learning efficiency (Chang et al., 2018). Statistical research concluded that the most frequent 2000 words cover 79.7% of written text (Nurmukhamedov & Webb, 2019), and a higher result of 96% was achieved in informal oval texts (Honig, 2007). This finding suggests that a vocabulary of 2000 to 3000 words is adequate for the demands of English conversation (Adolphs & Schmitt, 2003), although spoken proficiency may require a more extensive vocabulary than expected for fluent expressions. Other research also agrees on the vocabulary size of 3000 words for fluent conversation (Li et al., 2023) and suggests 5000 words for academic text comprehension, as 5000 words produce 95% coverage (Bian et al., 2021).

Furthermore, vocabulary size expands progressively as the learner grasps new words. Among the possible variables that influence word learning progress, word frequency is often considered the most important and even equal to word difficulty by researchers and teachers when designing L2 learning materials (Hashimoto & Egbert, 2019). As widely cited in teacher guidebooks and research studies (Laubscher & Light, 2020), the first 2000 words in the most frequent English word list are considered the general core vocabulary. These words are mainly chosen from the General Service List (GSL), the most popular word corpus for vocabulary size test (VST). The GSL has since been updated as the New GSL (NGSL) (Brezina & Gablasova, 2015). The Academic Word List (AWL) comprises a collection of 570 words extracted fracademic texts in addition to the NGSL (Coxhead, 2000). There has been much research on L2 vocabulary size measurement in the past few decades. These studies usually involve a vocabulary size test predesigned for a particular population of L2 learners (Li & Deng, 2018). The test usually contains fixed form-meaning quizzes or words with yes/no options indicating whether a tester can solve them (Zhang et al., 2020). The vocabulary size of a learner is measured by interpreting their testing results based on statistical models. Among these methods, the Rasch measurement (Holster & Lake, 2016) is part of the family of Rasch models, which assesses how closely the results of a measurement instrument in a VST align with their probabilistic expectations. Other statistical models for estimation include conditional, joint, and marginal maximum likelihood estimation (Nicklin & Vitta, 2022).

However, the effectiveness of the methodology for vocabulary size measurement has been questioned (Hashimoto, 2021). Since conducting a comprehensive vocabulary test is not feasible for L2 learners, the full extent of their vocabularies remains hidden from researchers. In comparison to the actual vocabulary size, testing with VSTs involves sampling with a limited sample size. These statistical interpretations require further empirical validation to draw definitive conclusions from these studies. Furthermore, vocabulary size tests are typically designed for specific, often small populations of L2 learners, and some tests are inflexible, making their application to larger or different populations challenging. Additionally, VSTs are often inefficient due to their reliance on paper-and-pencil (P&P) formats (Tseng, 2016).

The research question (RQ) for this study is "Given the condition that vocabulary of an L2 learner cannot be traversed, how to design an effective and factually verifiable vocabulary size measurement method for L2 learners?" The research question is divided into the following sub-questions:

RQ1: How to develop a factual verification method for vocabulary size measurement?
RQ2: How to design an intelligent vocabulary size test that can be tailored to any population of L2 learners and can measure vocabulary size adaptively, accurately, and efficiently?
RQ3: Can machine learning techniques be employed to assist the IVST to achieve high accuracy of assessing vocabulary size?

To address these issues, this study proposes a novel and systematic approach, namely an intelligent vocabulary size measurement method for L2 learners, which is adaptive, efficient, and factual verifiable. We first propose using massive robot testers to simulate variant vocabularies of L2 learners. The widely accepted notion shows that the more frequently a word is used in a language, the more likely it is to be acquired by a second language learner (Hashimoto, 2021). Statistical evidence suggests that higher word frequency indicates a higher probability of encountering, recognizing, and mastering the word in various contexts and daily speech (Hadley & Mendez, 2021). Low-frequency words are quickly forgotten, even if remembered at one time, due to their absence in written texts and everyday language. Consequently, word frequency is often assumed to be the natural order of word learning (Teng, 2019; González-Fernández & Schmitt, 2020). The robot testers are created to have the same word-frequency-based vocabularies with L2 learners. They can form a solid foundation for convincing verification with their visible vocabularies. Additionally, to address the issue that VSTs are often designed for and tested among specific population groups, an intelligent vocabulary size test (IVST) is put forward to estimate any tester's vocabulary size in real-time by an artificial neural network (ANN) through iterative learning. The IVST adaptively generates quizzes with appropriate difficulties according to the estimation and converges the estimated size to one's true vocabulary size efficiently within 60 quizzes. Furthermore, a long short-term memory (LSTM) model is applied to further improve measurement accuracy. Suppose that L2 learners with a specific vocabulary size will likely produce similar testing data in a VST. The robot testers are first grouped by different vocabulary sizes and "take" the IVST. The IVST produces testing data for testers of different groups. The LSTM model extracts group-related latent features from the testing data and measures any tester's vocabulary size by classifying one's testing data into a correct group.

The contributions of this study are as follows:

1. This study is the first one to propose and implement robot testers to simulate L2 learners with word-frequency-based and randomized vocabularies. The robot testers with their visible vocabularies form a factual verification foundation for the proposed vocabulary size measurement.
2. This study proposes the intelligent vocabulary size test. It generates quizzes dynamically and adaptively based on the tester's vocabulary size estimated in real-time by an artificial neural network model. It includes an efficient measurement strategy to complete the measurement within 60 quizzes. The IVST is designed for any population.
3. A LSTM model is applied to vocabulary size measurement to further improve the accuracy. Trained by the extensive testing data of grouped robot testers collected in

the IVST, the LSTM model measures the vocabulary size of any newcomer precisely by classifying one's testing data into a correct group.

The paper is organized as follows. In the "Introduction" section, the background, motivation, research questions, and contribution of this study are presented. The related work is provided in the "Literature review" section. In the "Methodology" section, the proposed methodology is introduced in detail, including robot testers, the intelligent vocabulary size test, and the LSTM model. In the "The experiments and results" section, the extensive experiments are conducted to testify the effectiveness of the proposed method. The "Conclusion and future work" section is attained in the end.

Literature review

The research of vocabulary testing traces to the time before 1970s (Nizonkiza & Van den Berg, 2014). Although the vocabulary lists were commonly provided for learners for writing and translation, vocabulary tests remained absence. In 1970s, standard synthesized tests emerged mainly for lexical and language proficiency purpose instead of vocabulary (Spolsky, 1995). Vocabulary was peripheral because grammar attracts L2 learners and instructors the most. People at that time believed that grammar strengthened L2 learners to generate infinite L2 sentences and obtain language proficiency (Nizonkiza, 2011).

The 1980s are widely considered a turning point in the measurement of L2 vocabulary size (Read & Dang, 2022) because vocabulary became an integral and predominant component of L2 language acquisition and proficiency (Pawley & Syder, 1983). Lewis (1993) announced the famous principle that "language consists of grammaticalised lexis, not lexicalised grammar." Several vocabulary studies emerged at this time, including the first widely used conventional vocabulary size measurement test, the vocabulary levels test (VLT) (Nation & Chung, 2009). The VLT is a pedagogical diagnostic tool and used to measure vocabulary by levels. The lower-level vocabulary contains words with high frequency (Coxhead, 2010). The VLT was soon superseded by the VST (Beglar, 2010), which was free and available online for various purposes and populations. For example, Coxhead et al. (2015) designed tests for native English speakers in New Zealand secondary schools, while Karami et al. (2020) created tests for Iranian English learners, and Zhao and Ji (2018) developed a Mandarin version. However, the validity of these tests was based on the modern theory of interpreting scores into vocabulary sizes (Read & Dang, 2022). The validation could have been more convincing due to the tests' lack of robustness in accounting for differences in learners' vocabularies. Such tests were fixed and designed for specific populations who cannot represent the diversity of global L2 learners, such as age, native language, educational background, and even L2 learning materials. The limitations were also acknowledged in these papers (Stoeckel et al., 2019). To address this issue, Tseng (2016) highlighted the flaws of conventional paper-and-pencil tests for vocabulary size measurement, which are fixed, uniform, and contain many test items. They proposed computerized adaptive testing (CAT) as an alternative. They divided 1536 participants into five groups to take the CAT of 240 vocabulary test items. The results were validated through Rasch analysis to suggest its potential efficiency and precision in measuring English vocabulary size. However, the validation for the specific test does not guarantee validity in other examinees. A test showing high validity through

Rasch analysis for a specific population may have low validity for other examinees (Gökcan & Aktan, 2022).

In addition, the notion of frequency-based acquisition was first derived from psychological research which recognized frequency as one of the three major experiential factors (Ellis, 2013). High frequency of a word indicates more experience conjunctions, including perception, context association, learning, practice and memory. The power law of learning (Anderson, 1982) also supports the relationships between practice and performance in the language acquisition. Therefore, frequency-based vocabulary size tests have been put forward to measure the vocabulary sizes of any L2 learners. Milton and Treffers-Daller (2013) utilized the frequency-based vocabulary size test to investigate the relationship between vocabulary size and the academic achievement of undergraduates in three British universities. Schmitt (2014) analyzed the pedagogical challenge raised by high, medium, and low-frequency vocabulary corpora and pointed out the importance of the most frequent words in vocabulary expansion. These two studies demonstrated the well-accepted notion that word frequency is crucial in vocabulary expansion. Conversely, Hashimoto (2021) argued that frequency-based vocabulary size tests might not reflect the variance of words that learners actually know. This statement was asserted based on the small Pearson correlation value between Rasch item difficulty and word frequency. The value was calculated based on the result of an experiment on a specific population of 403 English learners who took a pre-designed yes/no VST containing 10% random sample of the first 5000 most frequent words in the corpus of contemporary American English and sampled pseudowords. de Groot (2006) suggests the factors that may also be considered, including necessity, coverage, semantic neutrality, length, part of speech, polysemy, morphological regularity, cognateness, and orthographic transparency. Therefore, a VST should be flexible to accept additional factors.

Furthermore, Segbers and Schroeder (2017) created frequency-based virtual vocabularies to represent the L2 learners. These established virtual lexicons were collected from specific corpora and varied in size. The virtual lexicons were repeatedly sampled and "took" a VST to find the probability that a quiz could be solved. Vocabulary sizes were estimated by interpreting their testing scores with the probabilities of quizzes. However, the VST is fixed and pre-designed for specific examinees.

In summary, the abovementioned research is lack of solid validation or is not flexible enough. The words grasped by L2 learners may not exist in the VST quizzes, from which vocabulary size can be interpreted (Stoeckel et al., 2021). Specifically, although the CAT contained adaptive test items, the vocabularies of the 1536 human test-takers remained invisible and may not be qualified to represent any L2 learner, which weakened the validation of its effectiveness. On the other hand, frequency-based vocabulary size tests were not flexible enough to accept other factors. Thirdly, the virtual lexicons were not adaptive as they only took a VST with fixed test items. Finally, the mentioned VSTs are inefficient because they usually contain hundreds of vocabulary test items.

This research initiates artificial intelligence (AI) technologies in vocabulary measurement. To the best of our knowledge, the proposed intelligent vocabulary size measurement method is the first one designed for any population group learning any second language The method introduces robot testers to take the adaptive IVST. The quizzes are adaptively and dynamically generated with appropriate difficulties. Its effectiveness is also factually verified by the robot testers which accurately simulate L2 learners. Verified

by the substantial experiments, 60 quizzes are robust enough for the IVST to estimate vocabulary size accurately. More importantly, a long short-term memory model is applied to accompany the IVST to further improve its accuracy. And furthermore, the method is flexible to accept any merit factor without modification when applied in the real world.

Methodology

The overview of proposed methodology is shown in Fig. 1. To simulate L2 learners, the robot testers have randomized and word-frequency-based vocabularies of different sizes. The IVST consists of an initialization process for rough vocabulary size estimation, an ANN for real-time estimation, and an efficient testing strategy and generates appropriate quizzes adaptively. The robot testers "take" the IVST. Their sequential testing data are collected and used to train a LSTM model, which improves the performance further.

The robot testers: modeling the L2 learners

This section presents the method for creating robot testers to accurately model L2 learners from the vocabulary standpoint.

The creation of robot testers

A robot tester is a piece of running code combined with its data. Each robot tester has its own vocabulary which has various words but specific size. According to the well-accepted notion that frequent words are more likely to be grasped, we deduce the assertion that a robot tester's vocabulary should meet the following constraints:

1. The vocabulary size restriction: a robot tester has a known word list, e.g., its vocabulary. A specific group of robots has a matching number of known words, e.g., equal-sized vocabulary;
2. The commonly known words restriction: particular words are known by all robots of a specific group;
3. The vocabulary differentiation restriction: particular words are known only to some robots of a specific group; and
4. The variants restriction: the known words of the robots obey the variants that merit consideration, such as word frequency, part of speech, and syntactic complexity.

In this study, the vocabulary of a robot tester satisfies the constraints simultaneously. Any robot testers have a specific vocabulary size. Those having the same size are categorized

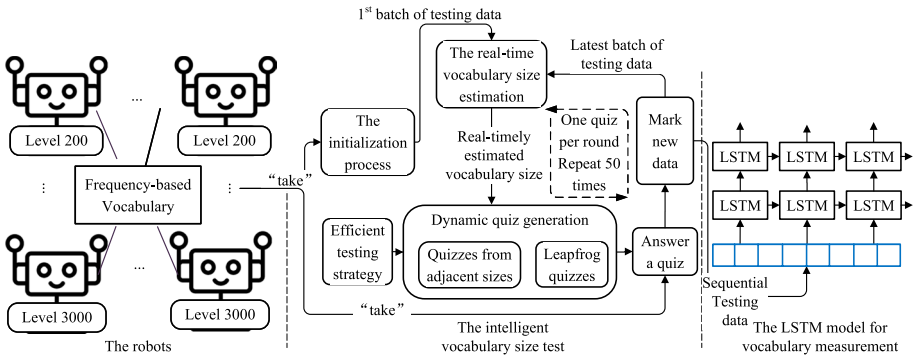

Fig. 1 The overview of the methodology

in a group, denoted as G_{size}. For example, a robot tester who has an 800-sized vocabulary, denoted as V_{800}, should belong to G_{800}. To simulate the gradual expansion of vocabulary by L2 learners, an interval of 200 words is specified between the adjacent sizes of groups. The minimum vocabulary size is 200 words, and the maximum vocabulary size is 3000 words, resulting in 15 robot tester groups. Their vocabulary words are randomly selected from the NGSL (2801 words) and AWL (570 words) with probabilities dynamically calculated based on word frequency.

In addition, the most frequently used words are usually retained by L2 learners in a group and are thus collected as the commonly known words, denoted as V_{size}^{common}. We suggest that the robot testers in a group of larger sizes, namely, the upper group, always include the commonly known words of those in groups of smaller sizes, namely, the lower group, into their vocabulary in advance to randomly select new words during creation. For instance, the commonly known words of robot testers in the smallest group, G_{200}, are collected and denoted as V_{200}^{common}. Then, when creating the robot testers in the next upper group, G_{400}, the commonly known words of the lower group are first included, denoted as $V_{400} \supset V_{200}^{common}$. Then, other words are randomly selected for any robot testers in G_{400} to fill their 400-sized vocabulary. This process repeats iteratively for any robot testers in subsequent upper groups until all robot testers are created.

Therefore, the vocabulary V_{size} of a robot tester in a group G_{size} can be represented as a union of two subsets, the set of commonly known words and the set of random selected words, as shown in Eq. (1).

$$\begin{aligned} V_{size} &= V_{size}^{common} \cup V_{size}^{freq} \\ V_{size}^{freq} &= \{word_i\}, 1 \le i \le n \\ word_i &= random(P), \end{aligned} \tag{1}$$

where V_{size}^{freq} is the word-frequency-based subset of the vocabulary which has n vacancies other than the common words. It selects new words randomly based on the set P, the probability coefficients of the *random* function. P consists of the probabilities of the remaining words calculated by the words' frequencies labeled in NGSL and AWL. A more frequent word has a higher probability. In addition, the labeled frequencies do not decline gradually. The words at the top have large frequency values, whereas the words at the bottom have a relatively low frequency (Vongpumivitch et al., 2009). Therefore, the probability for one of the remaining words, denoted as $word_j$, is calculated by

$$P_j = \frac{Frequency_j}{\sum Frequency}, \tag{2}$$

where $\sum Frequency$ denotes the total frequency of the remaining words that have yet to be chosen.

To conclude, the created robot testers meet the aforementioned constraints. The vocabulary size constraint is assured by the word selection process. In addition, guaranteed by the probability coefficient, the robot testers in a group are likely to be familiar with the most frequent words and share a collection of known words. Therefore, the commonly known words and variants restrictions are satisfied. Meanwhile, the random process generates different words for each robot tester to guarantee the vocabulary differentiation constraint.

The creation of a robot follows Algorithm 1.

procedure ROBOT CREATOR($size$) ▷ **Output:** a robot with the vocabulary size
    Collect the common words of the robot testers in any lower groups;
    Initialize the candidate words as those in NGSL or AWL that are not in the common words;
    for $j = size_{common} + 1$ to $size$ **do**
        Update the probabilities P of the remaining words using Eq. (2);
        Randomly select a word from the candidate words based on P;
        Add the word to the robot's vocabulary;
        Remove the word from the candidate words;
    end for
end procedure

Algorithm 1 The Algorithm for Creation of a Robot Tester

The intelligent vocabulary size test

The intelligent vocabulary size test consists of an initialization process, real-time vocabulary size estimation, dynamic quiz generation, and efficient testing strategy, as shown in Fig. 1. The IVST:

1. Estimates the vocabulary size of a tester in real-time based on the ANN model which is trained iteratively by one's latest six testing data;
2. Dynamically and adaptively generates quizzes of appropriate difficulties based on the estimation;
3. Efficiently and precisely converges the estimation to a tester's vocabulary size within 60 quizzes.

The initialization process

The initialization process roughly estimates a tester's vocabulary size and prepares the first testing data batch of six quizzes for the ANN model training.

The words in corpus are first sorted based on word frequency and then separated into collections without intersection. Each collection contains a specific number of words, i.e., 200, with different frequencies representing varying difficulty levels. Then, the initialization process randomly generates quizzes from these collections. The score of the previous quiz determines the difficulty of the next quiz. The steps of the initialization process are as follows:

1. For efficiency, any tester is supposed to have a middle-sized vocabulary of 1400. Thus, the IVST generates the first quiz from a 1400-sized word collection representing medium difficulty.
2. If the answer to the quiz is correct, the estimated vocabulary size increases by a certain amount, known as the stride. Since the tester's vocabulary size is unknown, a larger stride value of 400 words is chosen to speed up the rough estimation process.

3. If the answer is incorrect, the estimated vocabulary size decreases by the stride.
4. Subsequent quizzes are generated from the collection corresponding to the estimated size.
5. Ten quizzes are generated in the initialization process.

The ten quizzes' estimated vocabulary sizes are the initialization process's testing data, denoted as l_i and $1 \leq i \leq 10$. $l_5, l_6, \ldots, l_{10}$ are the first batch of testing data used to train the ANN model.

The real-time vocabulary size estimation

ANNs are designed to simulate the human brain's biological information processing. They learn knowledge, such as patterns, from input data during the training process. An ANN consists of several layers of artificial neurons. Each neuron has inputs, a transfer function, and an output. The neurons between layers are connected by coefficients, also known as weights. The outputs of neurons in the previous layer typically serve as inputs for neurons in the next layer. The weighted sum of these inputs forms the activation signal for the neuron, which then undergoes the transfer function to produce an output. Knowledge is acquired by updating the weights, guided by a loss function during the training process.

The architecture of the proposed ANN is shown in Fig. 2. It contains three hidden layers with ten neurons in each and is finally connected to a Sigmoid function. The loss is calculated by the mean squared error (MSE) loss function and is propagated backwards to update the parameters of the ANN model. All testing data are not used to train the ANN together because the data at the beginning is usually noisy, abundant, and inaccurate. A batch of the latest six data is suitable for driving the real-time estimation to the tester's true vocabulary size and guiding the IVST to generate quizzes with appropriate difficulty. The ANN is first trained by the testing data collected in the initialization process. Then, when the tester answers a quiz and a new piece of data l_m is marked, the latest six data, denoted as l_i and $m - 5 \leq i \leq m$, are used to train the ANN model. It estimates the tester's vocabulary size in real-time, denoted as $V_{estimated}$. Its value indicates the division point between the sizes of adjacent groups where the tester's actual vocabulary size may lie.

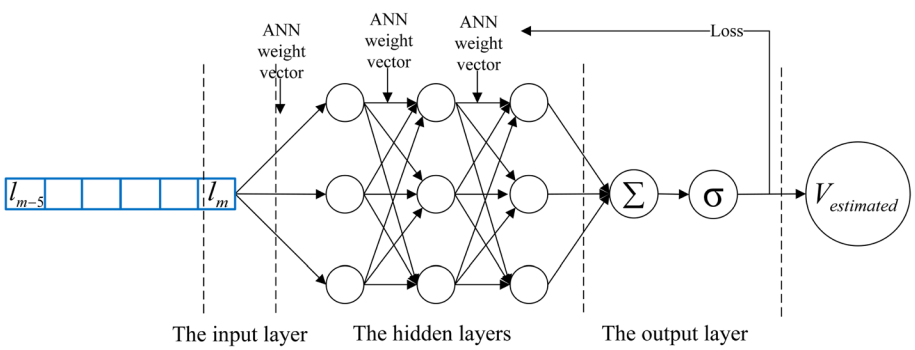

Fig. 2 The architecture of the ANN model

Dynamic quiz generation

After the initialization process, every quiz is generated based on the ANN model's real-time estimation of the tester's vocabulary size. As the estimated size value usually lies between the sizes of adjacent groups, the lower and upper group sizes, represented as $\lfloor V_{estimated} \rfloor$ and $\lceil V_{estimated} \rceil$, respectively, are used as the difficulties to generate the quizzes. For instance, if the estimated vocabulary size is 578, the lower and upper adjacent sizes would be V_{400} and V_{600}, respectively.

Quiz generation on adjacent sizes A subsequent quiz is generated according to the estimated vocabulary size.

When the estimated size $V_{estimated}$ is greater than the average of two adjacent vocabulary sizes, denoted as μ, the next quiz is generated with the difficulty of the upper size $\lceil V_{estimated} \rceil$. If the estimated size is lower than μ, the next quiz is generated according to the lower size $\lfloor V_{estimated} \rfloor$. When a tester answers a quiz, a new piece of testing data is generated by marking data of vocabulary sizes. If the answer is correct, the upper vocabulary size above the estimated vocabulary size is marked as the new piece of testing data, denoted as $l_m = \lceil V_{estimated} \rceil$. If the answer is incorrect, $l_m = \lfloor V_{estimated} \rfloor$ is marked. For example, suppose a tester usually answers quizzes generated with the difficulty of the lower size correctly but incorrectly for the ones according to the upper size. In that case, the new data are marked on either upper or lower sizes. In this condition, the ANN estimated real-time vocabulary size is always between the two adjacent sizes. Thus, a horizontal trend appears as the ANN perceives that the estimated vocabulary size is correct.

Leapfrog quiz generation A leapfrog quiz allows a tester's real-time estimated vocabulary size upgrading or downgrading across adjacent groups.

The ANN estimated vocabulary size most likely increases when the tester answers a quiz correctly. If the tester usually answers quizzes correctly, an increasing trend accumulates. When the estimated vocabulary size inclines and surpasses the preset upgrading threshold, denoted as $\hat{\theta}_{\uparrow}$, an upgrading leapfrog quiz is generated with the difficulty of the second upper vocabulary size. If the answer is still correct, the vocabulary size data is marked at the second upper size, $l_m = \lceil V_{estimated} \rceil + 1$. If the answer to the leapfrog quiz is incorrect, the data is marked as before, $l_m = \lfloor V_{estimated} \rfloor$. Alternately, if the incorrect answers accumulate enough and the estimated vocabulary size level is below the preset downgrading threshold $\hat{\theta}_{\downarrow}$, a downgrading leapfrog quiz from the collection of the second lower size is presented. If the answer is still incorrect, the new data is marked at the second lower size, $l_m = \lfloor V_{estimated} \rfloor - 1$. Otherwise, $l_m = \lceil V_{estimated} \rceil$ is marked as before.

Efficient testing strategy

An increase in the number of commonly known words often correlates with expanding vocabulary. Groups of L2 learners with larger vocabularies tend to know more commonly known words. To reduce the time needed for estimation, we suggest collecting the commonly known words of robot testers from each group and using their

differences as a key of efficiency. An efficient testing strategy can take advantage of this to speed up the adaptive testing process and the convergence of estimation by generating quizzes from adjacent groups' different commonly known words. Specifically, when a quiz is generated with the difficulty of a specific vocabulary size, the testing word is randomly selected from the commonly known words of the group of that size but not from the adjacent smaller one. For example, suppose a tester's estimated vocabulary size, $V_{estimated}$, is greater than the average size μ of the adjacent groups. In that case, the following quiz is generated from the commonly known words in the upper group but not in the lower group, denoted as $\lceil V_{estimated} \rceil^{common} - \lfloor V_{estimated} \rfloor^{common}$.

The algorithm of the intelligent vocabulary size test

The intelligent vocabulary size test runs according to Algorithm 2.

procedure VOCABULARY SIZE MEASUREMENT(A robot) ▷ **Output:** Marked testing data of 60 quizzes and 50 real-timely estimated vocabulary sizes

  Follow the initialization process to get the tester's latest six training data $l_5, l_6, \ldots, l_{10}$
  for all quiz from No.11 to No.60 **do**
    Feed the ANN model with the latest six training data for training.
    Estimate the real-time tester vocabulary size $V_{estimated}$ by the ANN model.
    if $V_{estimated} \geq \hat{\theta}_{\uparrow}$ **then**
      Generate a leapfrog quiz from the second upper vocabulary size $\lceil V_{estimated} \rceil + 1$.
      Mark the second upper size as new data $l_m = \lceil V_{estimated} \rceil + 1$, if the answer is correct.
    else if $V_{estimated} \geq \mu$ **then**
      Generate a quiz from the upper adjacent vocabulary size $l_m = \lceil V_{estimated} \rceil$.
      Mark the upper adjacent size as new data $l = \lceil V_{estimated} \rceil$, if the answer is correct.
    else
      Mark the lower adjacent size as new data $l_m = \lfloor V_{estimated} \rfloor$, if the answer is incorrect.
    end if
    if $V_{estimated} \leq \hat{\theta}_{\downarrow}$ **then**
      Generate a leapfrog quiz from the second lower vocabulary size $\lfloor V_{estimated} \rfloor - 1$.
      Mark the second lower size as new data $l_m = \lfloor V_{estimated} \rfloor - 1$, if the answer is incorrect.
    else if $V_{estimated} \leq \mu$ **then**
      Generate a quiz from the lower adjacent vocabulary size $\lfloor V_{estimated} \rfloor$.
      Mark the lower adjacent size as new data $l_m = \lfloor V_{estimated} \rfloor$, if the answer is correct.
    else
      Mark the upper adjacent size as new data $l_m = \lceil V_{estimated} \rceil$, if the answer is correct.
    end if
    Record $V_{estimated}$ and the marked data.
  end for
end procedure

Algorithm 2 The Algorithm of the Intelligent Vocabulary Size Test

The LSTM model for vocabulary size measurement

The intelligent vocabulary size test relies on an ANN to build a convergent path of the estimated vocabulary size through iterative learning. However, this method may be vulnerable when handling some exceptional cases containing coincidental outliers. For example, a second language learner may know a few hard words, which may also happen to be used as the IVST quizzes by chance. Furthermore, if these quizzes appear in concentration, the marked data responding to the correct answers may form an inclining trend. Given enough subsequent quizzes, the IVST still stands a chance of driving the trend back, but, if these quizzes happen to appear near the end, an incorrect measurement may appear.

The long short-term memory model is applied to address this issue because robot testers who have the same vocabulary size are likely to respond similarly to the IVST quizzes. Therefore, their testing data should contain group-related features which can be extracted and analyzed by deep learning models. The LSTM model extracts such latent features from the testing data of robot testers grouped by specific vocabulary sizes. Then, the LSTM model can precisely predict the vocabulary size for any new testers, including any second language learners, by classifying their testing data into the correct groups. In its entirety, the LSTM model makes the IVST more robust and insensitive to a few outliers.

The testing data for training

In this study, massive robot testers with a specific vocabulary size are created for each group so that the LSTM has enough knowledge to handle exceptional randomization. Their testing data of the last 50 IVST quizzes are marked as a sequence for each robot tester, denoted as $(l^r_{11}, l^r_{12}, \ldots, l^r_t, \ldots, l^r_{60})$ for the r^{th} robot tester. The sequential testing data of the grouped robot tests are then used to train the LSTM model. The group-related latent features are extracted from the massive testing data and represents human testers with a specific vocabulary size accurately.

The LSTM model for vocabulary size measurement

LSTM is put forward to replace the recurrent neural network (RNN) to address the known vanishing gradient problem. The LSTM is relatively insensitive to gap length which constitutes its advantage to extract long dependencies and make predictions based on sequential data. Therefore, it is an ideal choice for learning from sequential testing data of 50 items and making precise predictions. In detail, LSTM is a gating mechanism that controls memory by storing, reading, and forgetting information through gates. An LSTM cell is composed of an input gate, a forget gate, and an output gate, as shown in Fig. 3. The input gate controls the information that can be sent into the LSTM cell; the forget gate determines how much information should be removed. The output gate is responsible for outputting the proper information. Each gate contains parameter matrices that are updated through the training process and extract the long-term dependencies of any 50-item sequence of the testing data. Generally, given a pair

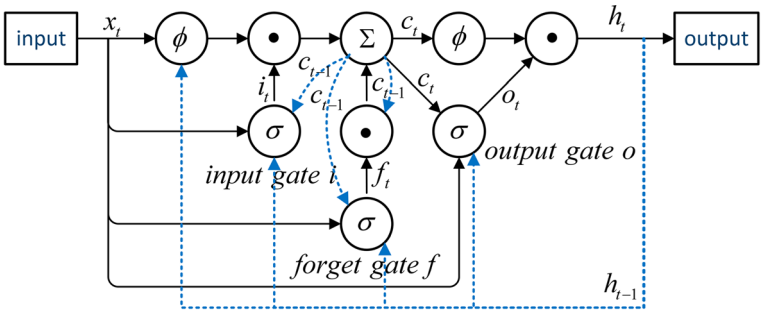

Fig. 3 An LSTM cell

of an input sequence $x = (x_1, x_2, \ldots, x_t, \ldots x_k)$, the calculations in an LSTM cell are as follows:

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + W_{ic}c_{t-1}), \tag{3}$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + W_{fc}c_{t-1}), \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \phi(W_{cx}x_t + W_{ch}h_{t-1}), \tag{5}$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + W_{oc}c_t), \tag{6}$$

$$h_t = o_t \odot \phi(c_t), \tag{7}$$

where σ and ϕ represent the sigmoid and tanh functions, respectively. The operators $\odot$ and $+$ are the element-wise multiplication and element-wise addition, respectively. As shown in Eq. (3), the input gate of the LSTM model reads the current input x_t, the previous hidden state h_{t-1} and the previous cell state c_{t-1}, and uses a sigmoid function (σ) to determine the importance of the input at time t. Similarly, the forget gate f_t determines which information needs to be attended to or ignored. The cell state c_t is calculated based on the input and forget gate information by the tanh function (ϕ) and element-wise multiplication ($\odot$) with Eq. (5). Consequently, the output gate calculates the hidden state h_t carried over to the next time step along with the cell states.

As illustrated in Fig. 4, the LSTM model has two layers, each consisting of 50 LSTM cells. The testing data from quizzes 11 to 60 are used as the input sequence to feed the LSTM cells, respectively, allowing the model to learn the whole convergent trend of each robot tester from their testing data. The gating mechanism of the LSTM can extract and memorize the long-term dependencies inside the testing data sequence. Finally, a linear classifier is connected to the LSTM to make predictions. In comparison, the artificial neural network only focuses on real-time vocabulary size estimation by extracting local short-term dependencies of the latest six testing data. The LSTM, however, can learn the whole sequences of each group thoroughly and comprehensively, extracting group-related latent features of the massive grouped testing data. Thus, when a new tester takes the IVST and produces new testing data, the LSTM can accurately predict one's vocabulary size by classifying the testing data into the correct group.

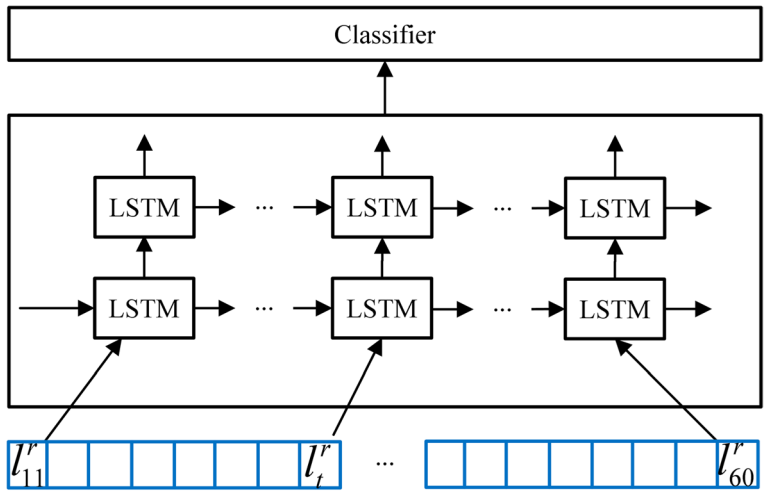

Fig. 4 The LSTM classifier

The experiments and results

This section presents the experiments and results which demonstrate high reliability and effectiveness, achieving accuracies of 98.47% for the IVST and 99.87% for the IVST with LSTM.

The experiments settings

The code of the experiments are developed in Python 3.7. The mainly used packages include pytorch and pyplot, responsible for ANN, LSTM and plots. The code runs on a computer with Intel Core i7-7820 CPU and 16 GB memory.

The corpus for the vocabulary size measurement experiments consists of the NGSL (2801 words) and AWL (570 words), for a total of 3371 words. Assuming that no one can grasp all these words, the maximum vocabulary size is set at 3000, with 200 words as the interval between adjacent groups, resulting in 15 groups from G_{200} to G_{3000}. One hundred robot testers are created for each group.

The effectiveness of the method is validated by accuracy metrics. The accuracy for group G_i, denoted by $Accuracy_i$, is defined by $Accuracy_i = \frac{N_i^{correct}}{N_i}$, where $N_i^{correct}$ denotes the number of correctly estimated cases, and N_i denotes the number of cases in group G_i. The overall accuracy is the fraction of all correct cases among all cases in the 15 groups, which is calculated $Overall\ Accuracy = \frac{\sum_1^{15} N_i^{correct}}{\sum_1^{15} N_i}$.

The accuracy values are calculated based on the robot testers' visible vocabularies. For the IVST, its estimation test is deemed accurate only if the estimated vocabulary size is within a range of plus or minus 100 of the true vocabulary size. For example, a prediction for a robot tester in group G_{400} is considered accurate only if the predicted vocabulary size fell within the range of (300, 500). For the LSTM model, a estimation is considered correct only when the model successfully classified a new robot tester into the correct group.

Experiment 1: Evaluation on sufficient quantity of robot testers

Experiment 1 evaluated how many robot testers are guaranteed for the LSTM to eliminate the influence caused by outliers. Human testers may know a few difficult words, which is simulated by outliers of randomization during robot tester creation. We test different quantities of robot testers in each group, i.e., 10, 30, 50, 70, 90, and 110. The results of the IVST alone and with the LSTM model are collected.

Experiment 2: Validation of the intelligent vocabulary size test

Experiment 2 verifies the effectiveness of the IVST. One thousand five hundred robot testers grouped by specific vocabulary sizes undergo the IVST individually. The accuracy is calculated based on the results of the 1500 robot testers.

Experiment 3: Validation of the LSTM model for vocabulary size estimation

One thousand five hundred sequences of testing data are collected in experiment 2. They are then used to train the LSTM model in experiment 3. Additionally, 1500 new robot testers (100 for each group) are created and used to simulate other human testers and validate the LSTM model. These new robot testers also take the IVST which generates the new testing data. The vocabulary size of a new robot tester is determined by the group in which one's testing data are classified.

Results

The creation mechanism of robot testers enabled the collection of commonly known words for each group. Table 1 shows the percentage of commonly known words in the robot tester vocabulary sizes of each group, confirming that the efficient testing strategy is statistically sound.

Results of experiment 1

The code ran several times to calculate the average value of the accuracy which is sometimes unstable for small quantities. The accuracy calculated from quantities more than 90 only change very slightly during the experiment. The results are shown in Table 2.

Results of experiment 2

Figures 5, 6, and 7 illustrate the testing data of three robot testers collected in the IVST. The blue *round* markers represent the real-time estimated vocabulary size based on the latest six testing data, beginning with the first one on the 11th quiz. The green *square* markers indicate the difficulty of the vocabulary size from which the quizzes are

Table 1 The percentage of the commonly known words of the robot testers in each group

Group	200	400	600	800	1000
Percentage	76%	73%	74%	68%	71%
Group	1200	1400	1600	1800	2000
Percentage	66%	61%	66%	65%	64%
Group	2200	2400	2600	2800	3000
Percentage	69%	61%	65%	63%	64%

Table 2 The results of the IVST alone and with the LSTM model for different quantities of robot testers in a group

Robot tester quantities	10	30	50	70	90	110
The IVST	67.36%	73.38%	81.74%	93.35%	97.94%	98.49%
The IVST with LSTM	66.93%	75.64%	89.38%	96.45%	99.83%	99.87%

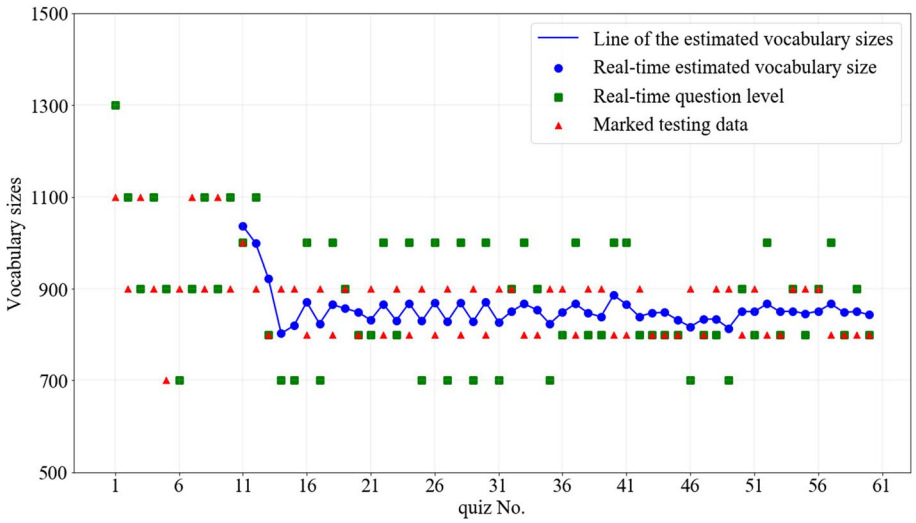

Fig. 5 Results of a robot tester in group: 800

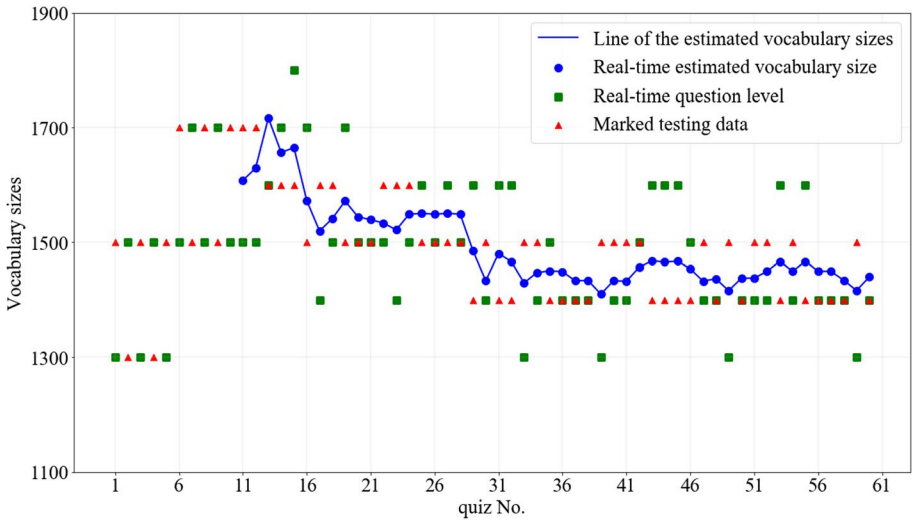

Fig. 6 Results of a robot tester in group: 1400

generated, referred to the real-time question level in the figure. The markers outside the adjacent sizes around the blue *round* markers are the difficulty of leap frog quizzes. The red *triangle* markers represent the marked testing data obtained from a robot tester in

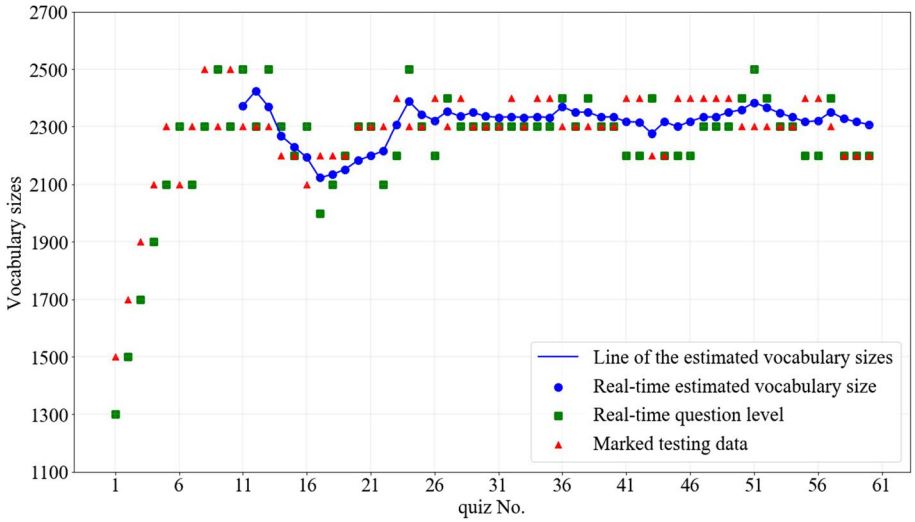

Fig. 7 Results of a robot tester in group: 2200

response to a quiz. The estimated vocabulary size finally produced at the 60th quiz is the result of a tester in the IVST.

Figure 5 shows one of the vast majority of cases. The robot tester has a frequency-based, randomized 800-sized vocabulary. The initialization process provides a rough estimation, and the IVST then real-timely estimates the tester's vocabulary size based on the last six testing data. Quizzes are appropriately provided from quizzes 11 to 14, leading the estimation to converge to the tester's actual vocabulary size quickly. The trend continues variably, with leapfrog quizzes offered many times, such as quizzes No. 14, 18, 23, and 25–31. As the robot tester usually answers correctly for downgrading quizzes and incorrectly for upgrading quizzes, the convergent trend becomes stable after quiz No. 40 until the end. This typical case illustrates the stability of the intelligent vocabulary size test, as the whole trend is controlled quickly within the two adjacent vocabulary sizes where the true vocabulary size lies.

Figure 6 demonstrates the robustness of the intelligent vocabulary size test. In the figure, the estimation for the robot test with a vocabulary size of 1400 begins with an undulating trend. The trend climbs as the robot answers correctly from quizzes 10 to 12, resulting in an upward leap. Fortunately, the trend then turns down and begins converging as the robot answers incorrectly from quizzes 14 to 16. As the estimated vocabulary size is close to the actual size, it takes more quizzes, from quizzes 16 to 28, to converge. From quiz 30 onwards, when the estimation is correct, the remaining trend is typical and similar to the majority of cases in Fig. 5. This shows that, with enough quizzes, the IVST can drive the trend and make a correct estimation in the end.

The experiment results confirm that failure cases appear rarely and the margin of error is always minimal, as shown in Fig. 7. The robot tester has an extensive vocabulary size of 2200 words. An ascending trend forms as they answer correctly in the initial process. Then, they answer incorrectly from quizzes 11 to 16 and correctly from 17 to 23. This small probability event drives the trend up and down, causing the IVST to overestimate

the vocabulary. In the end, the estimated vocabulary size is around 2300, only slightly higher. With additional sufficient quizzes, an accurate estimation is expected.

The results of experiment 2 are presented in Table 3, showing an impressive overall accuracy of 98.47%. This overall percentage is underscored by the individual group accuracies ranging from 94 to 100%, indicating that failure cases are extremely rare.

Results of experiment 3

In experiment 3, the classification results of the testing data of the 1500 new robot testers validated the model. The accuracy results of the vocabulary sizes estimated by the IVST with the LSTM model are presented in Table 4. The even higher overall accuracy value of 99.87% showed that the LSTM model can successfully mitigate failure cases and was able to predict the vocabulary size more accurately by classifying the testing data of the new testers into correct groups.

Parameter sensitivity

Parameters impacted the results significantly. Robot testers can be employed to investigate further the sensitivity of these parameters and optimize them. The most important parameters to consider are:

- β, which is the number of quizzes when the ANN model takes effect;
- $\hat{\theta}_{\uparrow}$ and $\hat{\theta}_{\downarrow}$, which are the decimal values that represent thresholds for upgrading and downgrading, thus triggering leapfrog quizzes.

Figure 8 shows the accuracy results of the IVST with the LSTM model in a sensitivity experiment of the parameter β when setting $\hat{\theta}_{\uparrow} = 0.6$ and $\hat{\theta}_{\downarrow} = 0.4$. The accuracy of the model changes significantly with different parameter β values. The lowest accuracy is observed when $\beta = 10$, and the accuracy increases with larger β values. However, when β equals 40, the increasing trend flattens. Therefore, the value of 50 is chosen for β as it yields the highest overall accuracy of 99.97%, while larger values lead to heavier computing and more quizzes for testers.

Figure 9 shows that, with a static β value of 50, the sensitivity experiment starts from the parameters $\hat{\theta}_{\uparrow} = 0.9$ and $\hat{\theta}_{\downarrow} = 0.1$. The large difference between the two parameters results in a significant decrease in accuracy. This occurrence is due to the strict values limiting the generation of leapfrog quizzes, thus restraining the convergence process. The best accuracy is achieved when setting $\hat{\theta}_{\uparrow} = 0.6$ and $\hat{\theta}_{\downarrow} = 0.4$, as these values provide higher flexibility for the LSTM model to guide the convergence trend and achieve better results.

Table 3 The result of the intelligent test in each group

Group	200	400	600	800	1000
Accuracy	100%	100%	99%	99%	98%
Group	1200	1400	1600	1800	2000
Accuracy	99%	94%	99%	96%	96%
Group	2200	2400	2600	2800	3000
Accuracy	98%	99%	100%	100%	100%
Overall accuracy	98.47%				

Table 4 The result of the IVST with the LSTM model in each group

Group	200	400	600	800	1000
Accuracy	100%	100%	99%	100%	99%
Group	1200	1400	1600	1800	2000
Accuracy	100%	100%	100%	99%	99%
Group	2200	2400	2600	2800	3000
Accuracy	100%	100%	100%	100%	100%
Overall accuracy	99.87%				

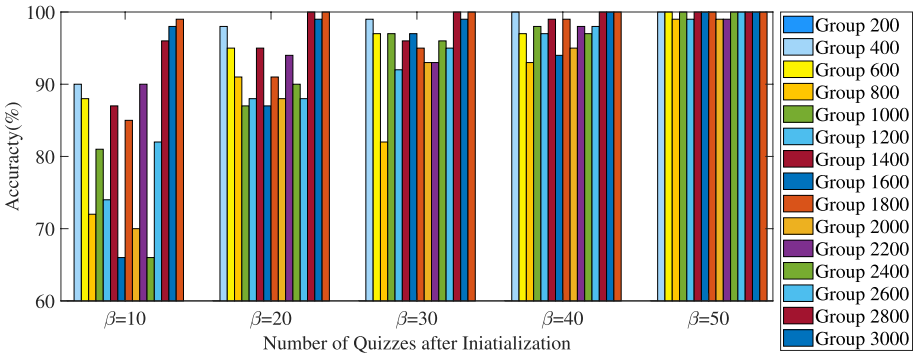

Fig. 8 Sensitivity of parameter β

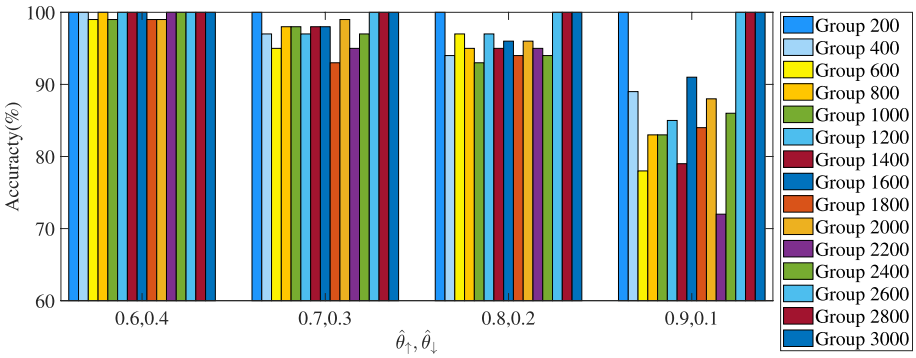

Fig. 9 Sensitivity of parameter $\hat{\theta}_{\uparrow}$ and $\hat{\theta}_{\downarrow}$

Discussions

Implementations and reflections

The first experiment showed that 1500 robot testers are adequate for correctly simulating L2 learners' vocabularies because the result value of accuracy became stable when processing more robot testers than quantity of 90 in each group. It also verified that the IVST along and with the LSTM model can handle the outliers, given sufficient robot testers and their testing data for training.

The second experiment demonstrated the IVST's capability to estimate vocabulary size for any individual L2 learner of any population. The difficulty of precise estimation lies in the variance of vocabularies between L2 learners. The IVST handled this issue effectively, as most cases followed a similar trend of Fig. 5 to converge to the robot tester's actual vocabulary size directly and smoothly. Occasionally, in some cases, the robot tester may continue answering challenging quizzes correctly and building up an increasing trend.

However, the IVST was robust enough to pull the trend back and make a correct estimation, as long as the remaining quizzes were sufficient. The accuracy value of 98.47% confirmed that the IVST could make precise vocabulary size estimations in most cases.

The third experiment showed an even higher precise estimation capability of the LSTM model by classifying testers' testing data into correct groups. Compared to the IVST, the LSTM model could extract the latent features of the grouped testing data and amend the rare failure cases, which had a minimal deviation from the correct value in experiment 2. The LSTM model effectively identified these failure cases and increased the accuracy to 99.87%.

Answers to the research questions

This study addresses the research questions by proposing a systematic approach for vocabulary size measurement, including robot testers, the intelligent vocabulary size test, and the LSTM model. The proposed methods are confirmed effective and can be verified in the experiments. The sub-questions are answered as follows:

> RQ1: How to develop a factual verification method for vocabulary size measurement methods?

The current method for measuring vocabulary size cannot be convincingly evaluated due to the invisibility of any L2 learner's vocabulary. To address this, we deduce the four constraints based on the notion that words with higher frequency are more likely to be met and then learned. Then, we first proposed and developed robot testers to best model L2 learners in terms of vocabulary and use them as a factual verification method. Created based on word frequency and randomization, the robot testers meet the four restrictions and model L2 learners accurately. With visible vocabularies, the robot testers then "take" the IVST and convincingly verify the effectiveness of the proposed vocabulary size measurement methods. Furthermore, the robot testers are flexible enough to incorporate other factors that may be deemed relevant for future VST research.

> RQ2: How to design a intelligent vocabulary size test that can be tailored to any population of L2 learners and can measure vocabulary size adaptively, accurately and efficiently?

A VST towards specific population is inadequate for accurately measuring vocabulary size due to its inability to accommodate all L2 learners. To address this, an intelligent vocabulary size test has been proposed that dynamically and adaptively generates quizzes based on the tester's vocabulary estimated by the ANN model in real-time. As the IVST progresses, it converged the estimated vocabulary size to the tester's actual size. Experiment 2 has demonstrated that the IVST can efficiently and accurately measure vocabulary size within 60 quizzes.

> RQ3: Can machine learning techniques be employed to assist the IVST to achieve high accuracy of assessing vocabulary size?

This research assumes that L2 learners with a specific vocabulary size most likely have commonly known words and should generate similar testing data in the IVST.

Therefore, this paper implements a LSTM model to extract latent features from the testing data of grouped robot testers labeled with specific vocabulary sizes. The model accurately classifies the testing data of any new one into correct groups and thus estimate their vocabulary size. The LSTM model achieves higher accuracy in experiment 3.

Applications

This study investigates the feasibility of creating robot testers to simulate L2 learners and verify vocabulary size measurement methods factually. While student models have been studied and discussed for many years, this paper proposes an initiative to implement a kind of student models in robot learners. Word frequency is a widely accepted parameter for L2 development in vocabulary size measurement research. Therefore, using robot testers to represent L2 learners based on word frequency is compelling. This methodology can be applied to other educational studies. Based on the accurate vocabulary size reported by the IVST, more precise teaching strategies can be suggested, more accurate learning plan can be designed and more specific sets of words can be provided for specific L2 learners to acquire. Actually, the proposed method has been integrated in an online English Learning Website. The IVST reported vocabulary sizes for L2 learners are accord with the instructors estimation.

Additionally, a new type of recursive learning, including the IVST testing step and the learning step to acquire artificial intelligence (AI) suggested words, is possible to be provided. With the assistance of AI, more discoveries and intelligent educational systems can be developed, optimized, and verified by robot learners.

Limitations and flexibility

This study is based on the widely accepted idea that the vocabularies of L2 learners are word-frequency-based. However, the study's limitations become apparent if this notion changes significantly. In real-world applications, word acquisition route may be related with the diversity of L2 learners (age, native language, educational background, etc.). To address this, the proposed robot testers are flexible enough to incorporate other factors as variables which may be identified merit and necessary in the future VSTs. For example, the expert knowledge of instructors can be used as a merit factor. The frequency order of words can be manually rearranged by instructors. The probabilities of the words acquired by robot testers are rearranged at the same time. As a result, the IVST adapts to the specific requirement in real-world. Furthermore, no changes are necessary for the IVST or the LSTM model for vocabulary size measurement.

Conclusion and future work

Measuring vocabulary in L2 learning is crucial, but current methods, such as vocabulary size tests (VSTs), are inflexible and often tailored to specific populations, limiting their representation of the diverse human vocabularies. This limitation can lead to discrepancies in item accuracy and validity across different populations, highlighting the need for more comprehensive validation.

To address these issues, this study proposed a systematic vocabulary size measurement method that addressed the issues of traditional vocabulary size measurement methods, including limited factual verification and population specificity. The proposed method, which included robot testers, the intelligent vocabulary size test, and the long short-term memory model, is accurate, efficient, and factually verifiable.

The robot testers accurately modeled variant L2 learners by word-frequency-based and randomized vocabularies, and the proposed intelligent vocabulary size test accurately, adaptively, and efficiently estimated vocabulary size within 60 quizzes based on the artificial neural network. The effectiveness of the method was validated by predicting 1500 robot tester's vocabulary sizes with an accuracy of 98.47%. Furthermore, the accuracy was further improved to 99.87% by applying the long short-term memory model to train the IVST testing data and make estimation through classification.

Future research will focus on exploring and incorporating additional factors related to vocabulary development into robot testers, such as cultural factors and individual learner differences. This study provides an initiative and practice case to incorporate student models into robot learners in educational research. With the help of artificial intelligence, robot learners can be developed to represent human learners for optimization, evaluation, and verification in other educational research areas, including language learning and other subject domains.

Abbreviations

L2	Second language
VSTs	Vocabulary size tests
IVST	Intelligent vocabulary size test
ANN	Artificial neural network
LSTM	Long short-term memory
GSL	General Service List
NGSL	New GSL
P&P	Paper-and-pencil
AWL	Academic Word List
RQ	Research question
VLT	Vocabulary levels test
CAT	Computerized adaptive testing
MSE	Mean squared error

Acknowledgements
Not applicable.

Authors' contributions
T.X. conceived the idea and developed the algorithm and software. All authors conducted the experiments and interpreted of results. T.X. prepared the original manuscript. X.C. and H.P. revised the manuscript. F.Q. was responsible for funding acquisition. All authors reviewed the manuscript.

Funding
This work was supported by Shanghai Engineering Research Center of Intelligent Education and Bigdata and Research Base of Online Education for Shanghai Middle and Primary Schools.

Availability of data and materials
The authors confirm that the data supporting the findings of this study are available from the corresponding author, upon reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 7 August 2023 Accepted: 14 September 2023
Published online: 13 October 2023

References

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics, 24*(4), 425–438. https://doi.org/10.1093/applin/24.4.425

Alahmadi, A., & Foltz, A. (2020). Effects of language skills and strategy use on vocabulary learning through lexical translation and inferencing. *Journal of Psycholinguistic Research, 49*(6), 975–991. https://doi.org/10.1007/s10936-020-09720-9

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*(4), 369. https://doi.org/10.1037/0033-295X.89.4.369

Beglar, D. (2010). A rasch-based validation of the vocabulary size test. *Language testing, 27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Bian, X., Cai, X., & Cai, D. (2021). The contributions of listening and reading vocabularies to listening comprehension of chinese efl students. *International Journal of Listening, 35*(2), 110–122. https://doi.org/10.1080/10904018.2019.1623678

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? introducing the new general service list. *Applied Linguistics, 36*(1), 1–22. https://doi.org/10.1093/applin/amt018

Chang, Y.-H., Liu, T.-C., & Paas, F. (2018). Cognitive resources allocation in computer-mediated dictionary assisted learning: From word meaning to inferential comprehension. *Computers & Education, 127*, 113–129. https://doi.org/10.1016/j.compedu.2018.08.013

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238. https://doi.org/10.2307/3587951

Coxhead, A. (2010). Grabbed early by vocabulary: Nation's ongoing contributions to vocabulary and reading in a foreign language. *Reading in a Foreign Language, 22*, 1–14.

Coxhead, A., Nation, P., & Sim, D. (2015). Measuring the vocabulary size of native speakers of english in new zealand secondary schools. *New Zealand Journal of Educational Studies, 50*(1), 121–135. https://doi.org/10.1007/s40841-015-0002-3

de Groot, A. M. B. (2006). Effects of stimulus characteristics and background music on foreign language vocabulary learning and forgetting. *Language Learning, 56*(3), 463–506. https://doi.org/10.1111/j.1467-9922.2006.00374.x

Ellis, N. C. (2013). Frequency-based accounts of second language acquisition. In S. Gass & A. Mackey (Eds.), The Routledge handbook of second language acquisition (pp.193–210). London: Routledge

Enayat, M. J., & Derakhshan, A. (2021). Vocabulary size and depth as predictors of second language speaking ability. *System, 99*, 102521. https://doi.org/10.1016/j.system.2021.102521

Georgiou, G. P., Perfilieva, N. V., & Tenizi, M. (2020). Vocabulary size leads to better attunement to l2 phonetic differences: Clues from Russian learners of English. *Language Learning and Development, 16*(4), 382–398. https://doi.org/10.1080/15475441.2020.1814779

Gökcan, M., & Aktan, D. Ç. (2022). Validation of the vocabulary size test. *Journal of Measurement and Evaluation in Education and Psychology, 13*(4), 305–327. https://doi.org/10.21031/epod.1144808

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics, 41*(4), 481–505. https://doi.org/10.1093/applin/amy057

Hadley, E. B., & Mendez, K. Z. (2021). Learning words that matter: Selecting vocabulary words for young children. *The Reading Teacher, 74*(5), 595–605. https://doi.org/10.1002/trtr.1978

Hashimoto, B. J. (2021). Is frequency enough?: The frequency model in vocabulary size testing. *Language Assessment Quarterly, 18*(2), 171–187. https://doi.org/10.1080/15434303.2020.1860058

Hashimoto, B. J., & Egbert, J. (2019). More than frequency? exploring predictors of word difficulty for second language learners. *Language Learning, 69*(4), 839–872. https://doi.org/10.1111/lang.12353

Holster, T. A., & Lake, J. (2016). Guessing and the rasch model. *Language Assessment Quarterly, 13*(2), 124–141. https://doi.org/10.1080/15434303.2016.1160096

Honig, A. S. (2007). Oral language development. *Early Child Development and Care, 177*(6–7), 581–613. https://doi.org/10.1080/03004430701377482

Hsu, C., & Ou Yang, F.-C. (2013). A vocabulary learning tool for l2 undergraduates reading science and technology textbooks. *International Journal of Science Education, 35*(7), 1110–1138. https://doi.org/10.1080/09500693.2013.764474

Karami, H., Kouhpaee Nejad, M., Nourzadeh, S., & Ahmadi Shirazi, M. (2020). Validation of a bilingual version of the vocabulary size test: comparison with the monolingual version. *International Journal of Bilingual Education and Bilingualism, 23*(4), 368–380. https://doi.org/10.1080/13670050.2017.1391744

Laubscher, E., & Light, J. (2020). Core vocabulary lists for young children and considerations for early language development: A narrative review. *Augmentative and Alternative Communication, 36*(1), 43–53. https://doi.org/10.1080/07434618.2020.1737964

Lewis, M. (1993). *The lexical approach: The state of elt and the way forward*. England: Longman.

Li, J., & Deng, Q. (2018). What influences the effect of texting-based instruction on vocabulary acquisition? Learners' behavior and perception. *Computers & Education, 125*, 284–307. https://doi.org/10.1016/j.compedu.2018.06.017

Li, J., Ji, L., & Deng, Q. (2023). The heterogeneous and transfer effects of a texting-based intervention on enhancing university english learners' vocabulary knowledge. *Computer Assisted Language Learning, 36*(1–2), 52–80. https://doi.org/10.1080/09588221.2021.1900264

Mahr, T., & Edwards, J. (2018). Using language input and lexical processing to predict vocabulary size. *Developmental Science, 21*(6), e12685. https://doi.org/10.1111/desc.12685

Masrai, A., & Milton, J. (2021). Vocabulary knowledge and academic achievement revisited: General and academic vocabulary as determinant factors. *Southern African Linguistics and Applied Language Studies, 39*(3), 282–294. https://doi.org/10.2989/16073614.2021.1942097

Milton, J., & Treffers-Daller, J. (2013). Vocabulary size revisited: the link between vocabulary size and academic achievement. *Applied Linguistics Review, 4*(1), 151–172. https://doi.org/10.1515/applirev-2013-0007

Nation, P., & Chung, T. (2009). Teaching and testing vocabulary. *The Handbook of Language Teaching,* 543–559. https://doi.org/10.1002/9781444315783.ch28

Nicklin, C., & Vitta, J. P. (2022). Assessing rasch measurement estimation methods across r packages with yes/no vocabulary test data. *Language Testing, 39*(4), 513–540. https://doi.org/10.1177/02655322211066822

Nizonkiza, D. (2011). The relationship between lexical competence, collocational competence, and second language proficiency. *English Text Construction, 4*(1), 113–145. https://doi.org/10.1075/etc.4.1.06niz
Nizonkiza, D., & Van den Berg, K. (2014). The dimensional approach to vocabulary testing: What can we learn from past and present practices? *Stellenbosch Papers in Linguistics, 43*, 45–61. https://doi.org/10.5774/43-0-169
Nurmukhamedov, U., & Webb, S. (2019). Lexical coverage and profiling. *Language Teaching, 52*(2), 188–200. https://doi.org/10.1017/S0261444819000028
Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. *Language and Communication, 191*, 225.
Read, J., & Dang, T. N. Y. (2022). Measuring depth of academic vocabulary knowledge. *Language Teaching Research*. https://doi.org/10.1177/13621688221105913
Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language learning, 64*(4), 913–951.
Segbers, J., & Schroeder, S. (2017). How many words do children know? a corpus-based estimation of children's total vocabulary size. *Language Testing, 34*(3), 297–320. https://doi.org/10.1177/0265532216641152
Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford: Oxford University Press.
Stoeckel, T., McLean, S., & Nation, P. (2021). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition, 43*(1), 181–203. https://doi.org/10.1017/S027226312000025X
Stoeckel, T., Stewart, J., McLean, S., Ishii, T., Kramer, B., & Matsumoto, Y. (2019). The relationship of four variants of the vocabulary size test to a criterion measure of meaning recall vocabulary knowledge. *System, 87*, 102161. https://doi.org/10.1016/j.system.2019.102161
Teng, F. (2019). The effects of context and word exposure frequency on incidental vocabulary acquisition and retention through reading. *The Language Learning Journal, 47*(2), 145–158. https://doi.org/10.1080/09571736.2016.1244217
Tseng, W.-T. (2016). Measuring english vocabulary size via computerized adaptive testing. *Computers & Education, 97*, 69–85. https://doi.org/10.1016/j.compedu.2016.02.018
Uchihara, T., & Clenton, J. (2022). The role of spoken vocabulary knowledge in second language speaking proficiency. *The Language Learning Journal,* 1–18. https://doi.org/10.1080/09571736.2022.2080856
Vongpumivitch, V., Huang, J.-Y., & Chang, Y.-C. (2009). Frequency analysis of the words in the academic word list (awl) and non-awl content words in applied linguistics research papers. *English for Specific Purposes, 28*(1), 33–41. https://doi.org/10.1016/j.esp.2008.08.003
Zhang, X., Liu, J., & Ai, H. (2020). Pseudowords and guessing in the yes/no format vocabulary test. *Language Testing, 37*(1), 6–30. https://doi.org/10.1177/0265532219862265
Zhao, P., & Ji, X. (2018). Validation of the mandarin version of the vocabulary size test. *RELC Journal, 49*(3), 308–321. https://doi.org/10.1177/003368821663976

Publisher's Note