Understanding Vocabulary Growth Through An Adaptive Language Learning System

Kerz, E.; Burgdorf, A.; Wiechmann, D.; Meeger, S.; Qiao, Y.; Kohlschein, C.; Meisen, T.

Publication date
2019

Document Version
Final published version

Published in
Proceedings of the 8th Workshop for Natural Language Processing for Computer-Assisted Language Learning

License
CC BY

Link to publication

Citation for published version (APA):
Kerz, E., Burgdorf, A., Wiechmann, D., Meeger, S., Qiao, Y., Kohlschein, C., & Meisen, T. (2019). Understanding Vocabulary Growth Through An Adaptive Language Learning System. In D. Alfter, E. Volodina, L. Borin, I. Pilan, & H. Lange (Eds.), *Proceedings of the 8th Workshop for Natural Language Processing for Computer-Assisted Language Learning: NLP4CALL 2019* (pp. 65-78). (NEALT Proceedings Series; Vol. 39), (Linköping Electronic Conference Proceedings; No. 164). LiU Electronic Press. https://www.aclweb.org/anthology/W19-6307/

Understanding Vocabulary Growth Through An Adaptive Language Learning System

Elma Kerz[1], Andreas Burgdorf[3], Daniel Wiechmann[2], Stefan Meeger[1], Yu Qiao[1], Christian Kohlschein[1], Tobias Meisen[3]

[1]RTWH Aachen University, Germany
[2]University of Amsterdam, Netherlands
[3]University of Wuppertal, Germany

Abstract

Learning analytics and educational data mining have gained an increased interest as an important way of understanding the way humans learn. The paper introduces an adaptive language learning system designed to track and accelerate the development of academic vocabulary skills thereby generating dense longitudinal data of individual vocabulary growth trajectories. We report on an exploratory study based on the dense longitudinal data obtained from our system. The goal is the study was twofold: (1) to examine the pace and shape of vocabulary growth trajectories and (2) to understand the role various individual differences factors play in explaining variation in such growth trajectories.

1 Introduction

Considerable variability is observed in the rate at which individuals (both children and adults) learn language. From the literature on child development and adult second language development we know that some individuals start slow and speed up, others start fast and continue at a steady pace. This variability is particularly apparent in the area of vocabulary acquisition (see, e.g., Hart and Risley, 1995; Pellicer-Sánchez, 2018). Understanding the acquisition of vocabulary knowledge – i.e. the pace of vocabulary growth – is considered to be of key importance for a number of reasons: Vocabulary skills are shown to be strongly related to a variety of academic, vocational and social outcomes (e.g. Rohde and Thompson, 2007; Dollinger et al., 2008; Verhoeven et al., 2011). These skills are a crucial component of language competence and language use (Nation, 1993; Milton, 2013) and their development is found to boost the acquisition of other language domains, such as grammar and phonology (e.g., Goodman and Bates, 2013). Vocabulary skills have been recognized as a strong predictor of reading comprehension ability in both first and second language development (e.g., Muter et al., 2004; Tannenbaum et al., 2006; Verhoeven and Van Leeuwe, 2008; Verhoeven et al., 2011; Cain and Oakhill, 2011).

With the emergence of technology-enhanced language learning systems and automatic analyses of educational data obtained by such systems, many efforts have been directed at facilitating the learning experience (e.g., Becker and Nguyen, 2017). These efforts emphasize the effectiveness of adaptive (personalized) language learning as opposed to traditional cohort-based learning (Ismail et al., 2016). The dense longitudinal data generated by such systems open up new avenues for exploring human learning based on learning analytics and educational data mining, an emerging multidisciplinary field closely linked to statistics and machine learning on the one side and the cognitive and language sciences on the other side (Vahdat et al., 2016). These data make it possible to perform learning behavior analytics at many different granularities and behavior categories.

In this paper we introduce an adaptive language learning system – AISLE (short for *Adaptive Statistical Language Learning*) – that was designed with the aim to track and accelerate the development of vocabulary skills and to generate dense, longitudinal data to understand the dynamics of growth of individual learning trajectories. The design of the system was motivated by recent developments in the language sciences in general, and in the area of language learning and processing in particular. These developments are driven, among other things, by the existence of large databases of

Elma Kerz, Andreas Burgdorf, Daniel Wiechmann, Stefan Meeger, Yu Qiao, Christian Kohlschein and Tobias Meisen 2019. Understanding vocabulary growth through an adaptive language learning system. *Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019)*. Linköping Electronic Conference Proceedings 164: 65–78.

language use (language corpora), the use of NLP techniques and statistical analyses and computational modeling of language data.

The paper is organized as follows. In a first step, we describe the architecture and design principles of the adaptive language learning system. In a second step, we present first results of a study on vocabulary growth based on the dense longitudinal data obtained by the system. The data come from a group of 46 second language (L2) learners of English who engaged with the AISLE system in a laboratory setting for several hours distributed across three sessions over a period of three weeks. Using a within-subject design embedded in an individual-differences (IDs) framework, the same group of participants was administered a battery of tasks assessing a range of experience-related, cognitive and affective IDs factors that may affect second language acquisition. The study is guided by the following two research questions: (1) What is the best longitudinal model that describes participants' vocabulary growth and how much variation is there in growth rates? and (2) What is the role of a range of IDs factors in explaining variation in participants' vocabulary acquisition?

2 Introducing AISLE: Design Principles and Architecture

The AISLE system is characterized by two design features: [1] 'optimal language input' (see Subsection 2.1) and [2] 'optimal repetition intervals' (see Subsection 2.2). The graphical user interface (GUI) was designed to give users automatic feedback during the learning process (see Subsection 2.3). The interface also includes a number of questionnaires and tasks assessing diverse individual differences across experience-related, cognitive and affective domains.

2.1 Extraction and Representation of Vocabulary Items

Since the target population are university students, we were particularly interested in tracking and accelerating the development of academic vocabulary (AV). As it is the case with the general vocabulary skills, AV knowledge is recognized as an indispensable component of academic reading abilities (e.g., Biemiller, 1999; Nagy and Townsend, 2012), which is and has been directly linked to academic success, economic opportunity, and societal well-being (Ippolito et al., 2008; Jacobs, 2008). The key role of academic vocabulary in educational success is true for both native and non-native speakers of English (e.g., Schmitt et al. 2011). In response to this, a substantial amount of research has been devoted to the compilation of vocabulary lists (Gardner and Davies, 2013). A major advance has been in recognizing that language requires not only knowledge of a vast amount of statistically relevant academic vocabulary but also successful extraction of the statistics of academic multiword sequences (MWS), i.e. variably sized continuous or discontinuous frequently recurring strings of words. In fact, in recent models of language MWS (ngrams) are increasingly recognized as the fundamental building blocks that facilitate anticipatory processing and boost language acquisition (Arnon and Christiansen, 2017). Correspondingly, the term 'vocabulary item' is used here as a cover term for both single words and MWS (ngrams of different orders).

To arrive at 'optimal language input', we extracted 'statistically relevant' vocabulary items – words (unigrams) and n-grams for $n \in \{2, 3, 4\}$ – from a Corpus of Contemporary American English[1], approx. 560 million words of text equally divided among spoken, fiction, popular magazines, newspapers, and academic texts. This extraction involved several preprocessing steps that we performed using the spaCy[2] framework for natural language processing. The whole preprocessing pipeline is written with PySpark[3] and executed on a Hadoop[4] cluster. The pipeline involved the following four consecutive steps:

1. *Lemmatization*: This step is only performed to extract unigrams. The outcome of this step is a sequence of lemmas for a given processed document.

2. *Sentence splitting*: The sentence splitting was performed to ensure that n-grams are not extracted across sentence boundaries, and also to increase the degree of parallelism of the following steps.

3. *N-Gram Extraction*: Next, we extracted n-grams for all for $n \in \{2, 3, 4\}$ for each sentence. The result of this step is a collection

[1] https://corpus.byu.edu/coca/

[2] https://spacy.io/

[3] https://spark.apache.org/

[4] https://hadoop.apache.org/

of all n-grams along with the number of documents in which an item occurs and its frequency of occurrence in each document.

4. *Metrics Calculation*: The final step concerns the calculation of more sophisticated metrics, used in the identification of statistically relevant vocabulary items. By applying these metrics, only those words relevant for understanding texts - neither too general, nor to specific - are presented to learners.

As a metric for the distribution of a n-gram in the corpus we use dispersion as defined by Gardner and Davies (2013). Formalized, we used the following metrics for frequency f and dispersion d where i_n defines an arbitrary item with n words (n-gram), T_k defines a subcorpus with $k \in \{(a),(b),(c),(d)\}$, $N_n(T_k)$ defines the list of n-grams in subcorpus T_k and $\#i_n(T_k)$ describes the count of the item i_n in the subcorpus T_k. Further, $\sigma i_n(T_k)$ describes the number of documents of subcorpus T_k, i_n appears in. The abbreviation *gen* stands for 'general' and *ac* for 'academic':

$$f_{gen}(i_n) := \frac{\sum\limits_{k \in \{a,b,c\}} \#i_n(T_k)}{\sum\limits_{k \in \{a,b,c\}} |N_n(T_k)|} \tag{1}$$

$$d_{gen}(i_n) := \frac{\sum\limits_{k \in \{a,b,c\}} \sigma i_n(T_k)}{\sum\limits_{k \in \{a,b,c\}} |N_n(T_k)|} \tag{2}$$

$$f_{ac}(i_n) := \frac{\#i_n(T_{(d)})}{|N_n(T_{(d)})|} \tag{3}$$

$$d_{ac}(i_n) := \frac{\sigma i_n(T_{(d)})}{|N_n(T_{(d)})|} \tag{4}$$

A vocabulary item is considered to be 'statistically relevant' if one of the conditions given in (5) and (6) holds, where k_f and k_d are variable thresholds for the frequency and dispersion ratio, respectively, between academic and general corpora that are determined experimentally, depending on value n:

$$\frac{f_{ac}(i_n)}{f_{gen}(i_n)} > k_f \tag{5}$$

$$\frac{d_{ac}(i_n)}{d_{gen}(i_n)} > k_d \tag{6}$$

Further, we calculate a rank that defines how academic a n-gram is as follows where the parameters MIN_D (minimum academic dispersion) and MIN_R (minimum ratio of academic and general dispersion) have to be evaluated experimentally for each item length:

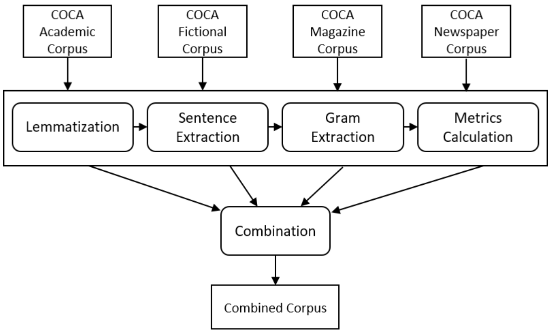


Figure 1: NLP pipeline for extracting statistically relevant vocabulary items

$$rank = \begin{cases} 0 & \text{if } d_{ac} < MIN_D \\ 0 & \text{if } \frac{d_{ac}}{d_{gen}} < MIN_R \\ \frac{f_{ac}}{f_{gen}} \cdot \frac{d_{ac}}{d_{general}} & \text{else} \end{cases} \tag{7}$$

After this pipeline has been executed for all subcollections (a)-(d) of the COCA corpus a combination step is performed that aggregates the results from the four collections and calculates the defined rank for each item.

The extracted items are represented in a Neo4j[5] graph database. The access to the database was realized in a flask API. This enables the interconnection of each n-gram with all its constitutive lemmas, which is especially useful later on to ensure that the basic building blocks of an n-gram are known to a degree necessary to present it to the user during learning. The graph representation of data consists of three different types of nodes:

- *lemma*-nodes contain a lemma as well as related metrics like frequency and dispersion.
- *n-gram*-nodes contain a n-gram and related metrics like frequency and dispersion.
- *user*-nodes represent a user of the learning application together with some information about his current state of learning.

For relations, we define the following two types:

- *contains*: this directed relation connects n-gram nodes to the lemmas of the words it

[5] https://neo4j.com/

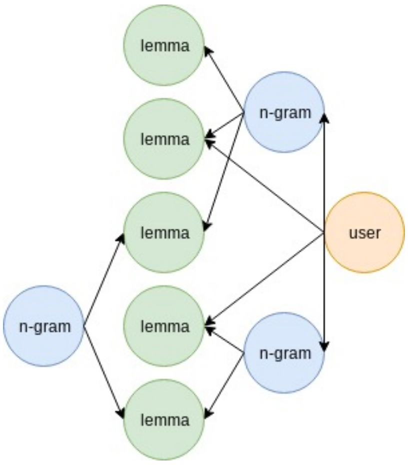


Figure 2: Representation of words and grams in a graph database

consists of. Additionally the relations store the metrics relative item frequency and entropy.

- *has_seen*: this directed relation connects user nodes to the lemmas and n-grams they have already seen while using the learning application. As soon as this relation is established, it further contains information about the current learning progress for this item like learning score, how often it has been presented and how long answers did take.

Lemma and n-gram nodes and their metrics are initially loaded directly from the aggregated corpus list and then connected using *contains*-relations which indicate the position of a lemma in the respective n-gram. Following the equations given in (5) and (6), we only imported academically relevant items into the graph database. The user nodes are populated from the running learning system application and a new node is generated for every user upon registration.

2.2 Learning algorithm

To ensure 'optimal repetition intervals' we developed and implemented an adaptive learning algorithm. The general structure of the learning algorithm is visualized in Figure 3. The algorithm selects a set of items from the graph database, which fall into four different categories: (1) items never seen before, (2) items recently answered incorrectly, (3) items close to be learned and (4) items already learned. The algorithm presents all items to the user one after the other and waits for the user's response. The current knowledge status of a vocabulary item is represented by 'normalized learning score' that takes the user's prior history of a given item into account. Values between 0 and 0.8 indicate that a given item is not yet part of the user's vocabulary repertoire. Once the learner has reached a normalized learning score for a given item that is greater than 0.8 the item is considered to be learned. The scoring of an individual user response to an item depends on whether or not the item has already been presented to the user. If the item is presented for the first time and the answer is correct, a *has-seen*-relation with the value 0.8 is created in the graph database between the user and the item and the item is treated as already known. If the answer is incorrect, the evaluation of the answer depends on the severity of the error, so that a spelling error is punished less than a completely wrong word. To this end, the user's response – a character string – is compared with the target word(s) based on the Levenshtein distance between the two strings. The evaluation takes the length of the target item into account (1 - *Edit distance/Word length* (*in characters*)) and ranges between between 0 (maximally incorrect) and 1 (maximally correct)).

The primary metric of participants' performance is their cumulative number of items learned during the time of engagement with the system. An vocabulary item was taken to be learned if (i) it is was not marked as 'previously known' and (ii) its 'normalized learning score' – the sum of all scores received for an item normalized by the number of presentations – has reached the threshold value of 0.8. Based on the user's prior performance, the algorithm decides on the next set of items based on their statistical relevance and the learner's current knowledge state of the vocabulary items.

2.3 User Interface

Users interact with the system via a web interface based on the *vue.js*-framework[6]. The web interface provides two major functionalities: user interaction and user tracking. After login, the user has access to the vocabulary learning module as well as to a number of tasks and measures geared to assess a range of learner background and IDs factors. During vocabulary learning, the user performs a cloze test (aka fill-the-gap task) where a sentence is presented in which the target item is

[6] `https://vuejs.org/`

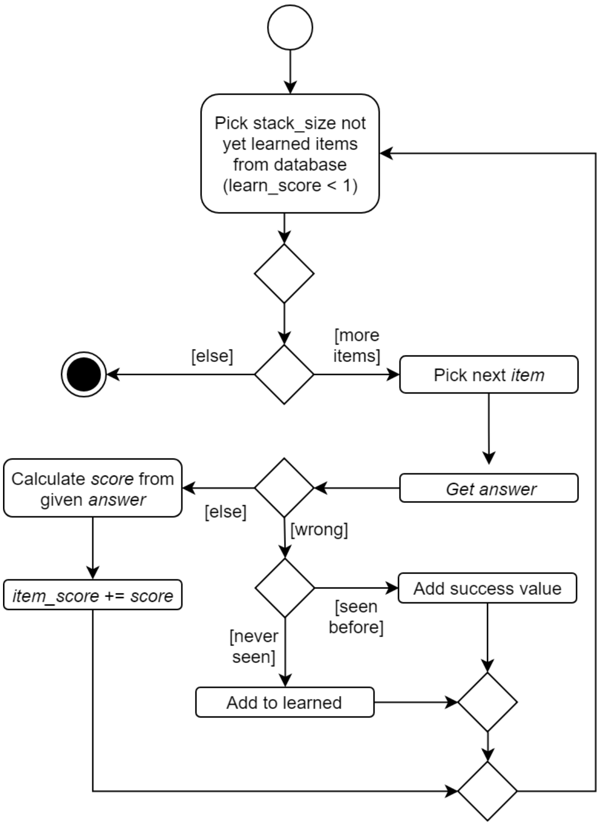
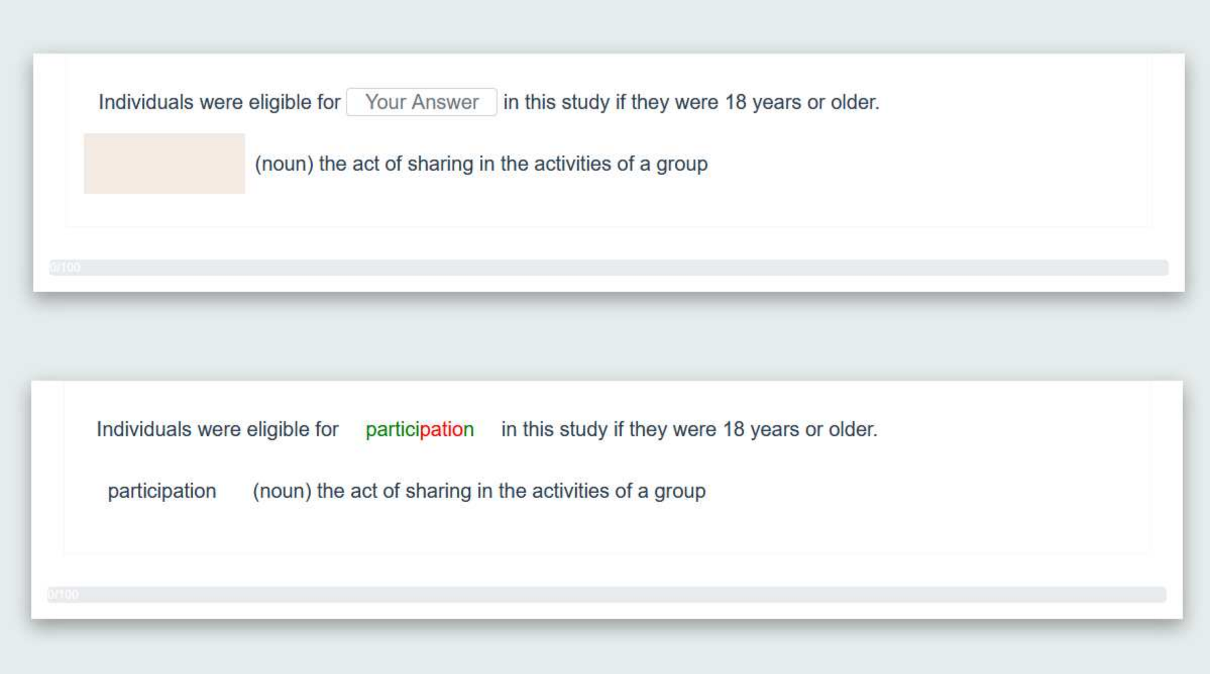


Figure 3: Learning Algorithm

missing and the task is to fill in the gap. The corresponding definition of the target item is presented below the sentence (see Figure 4 (top)). In case the user has entered the target word, the vocabulary item is colored green and displayed for two seconds. In case of a mismatch between the target and the user's input string, the correct string is presented with mismatching characters being highlighted in red font color (see Figure 4 (bottom)). After presentation of the correct answer, the user is prompted to re-enter it and the next item is presented. The interface stores and visualizes multiple relevant performance indicators that are available to the user at any point in time during interaction with the system. These indicators include the number of learned words so far and the longest streaks of consecutive correct inputs in the current session and during the total interaction period with the system (see Figure 5). A number of additional metrics are collected that, while not shown to the user, are useful for subsequent data analyses. These metrics include the number of responses per minute, the average number of repetitions per item, the number of items presented that were already known, the average number of presentations of an items until the item was learned, mean time until item learned in minutes, and the mean number of words per hour (see Table 1).

2.4 Integrated individual differences tasks and measures

The interface features a range of questionnaires and tasks assessing diverse individual differences (IDs) factors across experience-related, cognitive and affective domains. Upon successful registration, learners can currently complete a total of eight questionnaires and tasks. The group of currently integrated measures includes a two standardized tests designed to assess receptive vocabulary, the 'Lexical Test for Advanced Learners of English' (LexTALE, Lemhöfer and Broersma, 2012) and the 'Vocabulary Levels Test' (VLT, Schmitt et al., 2001) as well as a proxy measure of print exposure, the 'Author Recognition Test' (ART, West et al., 1993), and the 'Need for Cognition' test (NFC, Cacioppo et al., 1984), a personality-based measure indicating the degree to which an individual prefers cognitively engaging activities (see Subsection 3.2. for further details). The battery further contains implementations of two language and social background questionnaires – the LEAP-Q questionnaire (Marian et al., 2007) and the LSBQ-questionnaire (Anderson et al., 2018), as well as the Big Five Inventory (BFI, John et al., 2008) designed to assess five personality dimensions (Extraversion, Neuroticism, Conscientiousness, Openness to Experience, and Agreeableness). The web-based integration of tasks gauging additional cognitive abilities is still under development. At present, such tasks can be integrated using separate applications (see Section 3 for details on how these tasks are currently integrated into the system).



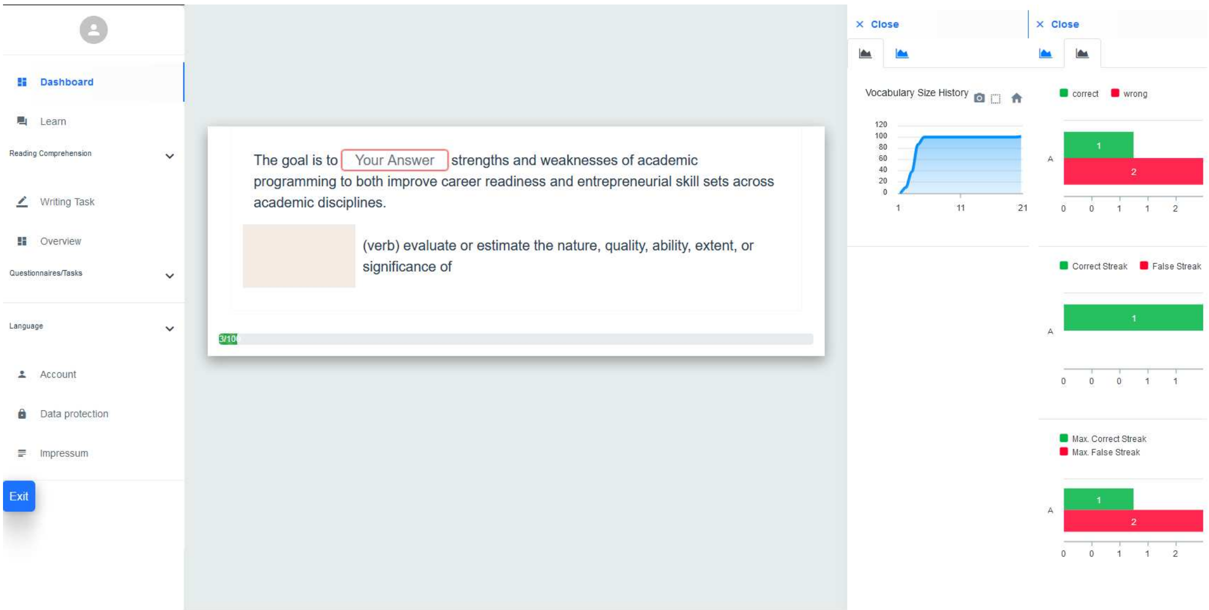
Individuals were eligible for Your Answer in this study if they were 18 years or older.

(noun) the act of sharing in the activities of a group

Individuals were eligible for participation in this study if they were 18 years or older.

participation (noun) the act of sharing in the activities of a group

Figure 4: Item presentation – Users interact with the AISLE system via a fill-the-gap task. Sentences containing an empty slot to be filled with a target word are presented along with a definition of the target word (top). In case of a mismatch between the target and the user's input string, the correct string is presented with mismatching characters being highlighted in red font color (bottom).



The goal is to Your Answer strengths and weaknesses of academic programming to both improve career readiness and entrepreneurial skill sets across academic disciplines.

(verb) evaluate or estimate the nature, quality, ability, extent, or significance of

Figure 5: Performance feedback provided to the user – The interface stores and visualizes multiple relevant performance indicators that are available to the user at any point during interaction with the system. These indicators include the number of learned words so far and the longest streaks of consecutive correct inputs in the current session and during the total interaction period with the system.

3 Modeling Growth Trajectories

In this section, we report on first results of a study on vocabulary growth based on the dense longitudinal data obtained by the AISLE system. As outlined in the Introduction (Section 1), the study addressed the following two research questions: (1) What is the best longitudinal model that describes participants' vocabulary growth and how much variation is there in growth rates? and (2) What is the role of a range of IDs factors in explaining variation in participants' vocabulary acquisition? We focus here on the acquisition of individual words (1-grams). The number of cumulative word types learned within a four-hour engagement with the AISLE system served as the measure of vocabulary growth. Variability in this performance metric was related to a total of 17 individual difference measures: four experience-based measures, five personality indicators and eight cognitive measures (see Subsection 3.2 for details; an overview of these measures is provided in Table 2).

3.1 Participants

The data come from forty-six second language (L2) learners of English (25 female and 21 male, M = 22.98 years, SD = 3.32). All participants were university students from the RWTH Aachen University studying towards a BA or MA degree.

3.2 Materials

L2-Experience measures: Participants were administered two receptive vocabulary tasks: the 'Lexical Test for Advanced Learners of English' (LexTALE, Lemhöfer and Broersma, 2012) and the 'Vocabulary Levels Test' (VLT, Schmitt et al., 2001). The LexTALE is a short yes/no vocabulary test implemented as a lexical decision task. In it particpants are presented a series of letter strings, some of which are existing English words and some of which are not, and are asked to indicate for each item whether it is an existing English word or not. The test consists of 60 items (40 words, 20 nonwords). Performance on the test is assessed as the percentage of correct responses adjusted for the unequal proportion of words and nonwords (averaged % correct).

The VLT assesses vocabulary knowledge at four frequency levels of English word families targeting the top 2000, 3000, 5000, and 10000 most frequent words in a language plus words from the domain of academic language (based on items from the Academic Word List; Coxhead, 2000). Each level consists of 30 items in a multiple matching format in which single words in the left-hand column need to be matched with a meaning presented in the right-hand column. Performance on the VLT is measured as the number of correct matches.

In addition, participants completed the 'Author Recognition Test' (ART, West et al., 1993) and the 'Need for Cognition' test (NFC, Cacioppo et al., 1984). The ART is a proxy measure of print exposure in which test takers are presented with a series of 81 names and foils and are asked to indicate which ones they recognize as authors. Performance on the task is assessed in terms of the number of correctly identified authors minus the number of foils selected.

The NFC is a personality-based measure indicating the degree to which an individual prefers cognitively engaging activities. Test takers indicate their agreement (based on a 5-point Likert scale) with 18 statements such as 'I really enjoy a task that involves coming up with new solutions to problems' (positive polarity item) or 'Thinking is not my idea of fun' (negative polarity item). Scores on the NFC are determined by averaging the responses to all items (with negative polarity items being reverse scored).

Personality-related measures: Participants were also asked to fill in the Big Five Inventory (BFI, John et al., 2008), a 44-item personality-related questionnaire that measures an individual on the Big Five personality dimensions (Extraversion, Neuroticism, Conscientiousness, Openness to Experience, and Agreeableness). Scores on each dimension are assessed in terms of person-centered z-scores adjusted for differences in acquiescent response styles ('yea-saying' vs. 'nay-saying')

Cognitive measures: We administered a total of eight cognitive IDs measures as indicators three aspects of cognition: (1) four indicators of statistical learning ability (the probabilistic Serial Reaction Time (pSRT) task from Kaufman et al. (2010), along with the Visual-Nonverbal-Adjacent (VNA), Auditory-Verbal-Adjacent (AVA), and the Auditory-Verbal-Nonadjacent (AVN) Artificial Grammar Learning tasks described in Siegelman and Frost (2015)), (2) one indicator of verbal working memory (a modified version of the Reading Span (RSPAN) task as described in Farmer et al. (2017)), and (3) three indicators of cognitive control (the variants of the Simon task and the Eriksen-Flanker task used in Wilhelm et al. (2013) as well as the Stroop Color-Word task described in Linnman et al. (2006)). Performance on all measures was scored following standard procedures. A brief description of each measure is provided in Table 2. For further details on these tasks the reader is referred to the cited literature.

3.3 Procedure

The participants engaged with the adaptive language learning system in a laboratory setting for a total of approximately four hours distributed across three sessions within a period of three weeks. Once they had successfully registered participants completed the experience- and personality-related questionnaires and tasks. The cognitive tasks were interspersed with the vocabulary learning sessions. These tasks were administered in a laboratory setting using PsychoPy[7], an open-source application for the creation of experiments in behavioral science (Peirce et al., 2019). The results obtained from these tasks were automatically exported into the graph database.

3.4 Results

Before turning to the modeling results, we first briefly present an overview of the descriptive statistics of the engagement- and performance-metrics tracked by the AISLE system (Table 1). As shown in Table 1, there was considerable variation in the way users interacted with the system as well as in their learning outcomes. For example, the observed range of the number of items learned was 8 to 84 items, with a mean of 19.18 items learned and a standard deviation of 17.90 items. Normalized by the net amount of time that users engaged with the system these differences corresponded to an observed range in mean learning rates of 0.89 to 23.03 words learned per hour, with a mean of 5.16 words per hour and a standard deviation of 4.36 words per hour. The descriptive statistics of all cognitive, personality-related and L2 experience-related individual difference measures investigated in this study is presented in Table 2.

In finding the best model for vocabulary growth, we began with an empirical plot of participants' cumulative number of words learned (Figure 6: left panel). As is evident from this plot, participants varied considerably in their rates of vocabulary growth. Growth curve analysis (Mirman, 2017) was used to analyze the word learning trajectories up to the 678^{th} interaction, which was reached by 75% of the participants (i.e. 25% of the participants responded to fewer than 678 items). To obtain the best fitting within-person model for these data, i.e. the 'unconditional growth model (GCM)', we fitted linear, quadratic, and cubic growth models to the data using orthogonal polynomials of 'number of interactions' as our 'Time' variables. All models were fitted using the BOBYQA algorithm for optimization as implemented in the package lme4 (version 1.1-21, Bates et al., 2014) for the R language and environment for statistical computing and graphics (R Core Team, 2018). Model comparisons using Akaike's Information Criterion (AIC) revealed that the cubic unconditional growth model best represented the empirical data. The plot of the cubic model also best mirrored the plot of the empirical data (see Figure 6: center panel). On average, the cubic model indicates that users have an estimated cumulative vocabulary of approximately 15 word types, with an average increase of about 2 words types per 100 presented items. The rightmost plot in Figure 6 shows the predicted vocabulary growth at the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles.

Next we explored the relationship between vocabulary growth and each of our 17 L2 experience-related, personality-related and cognitive individual differences measures. All IDs variables were dichotomized based on median splits (high vs. low). The best fitting (minimal adequate) model was identified using a forward model selection procedure based on likelihood ratio tests, i.e. we started with the cubic unconditional growth model and evaluated the added value of each IDs predictor. We subsequently included the most significant predictor, re-estimated the model and repeated the procedure until no significant term was left to include. In all models we used the maximal random effects structure justified by our design, which included by-subject random intercepts and slopes on all time terms. The results of the models are summarized in Table 3. Preliminary analyses (data not shown) indicated that – when considered on their own – 9 out of the 17 IDs variables were significant predictors of growth trajectories (L2 experience-related: ART (sig. quadratic

[7] https://www.psychopy.org/

Table 1: Descriptive statistics of the AISLE metrics

	mean	sd	obs. range
Total number of responses	828.67	391.34	271 – 1980
Number of responses per minute	3.84	1.92	1.69 – 10.64
Average number of repetitions per item	4.67	1.97	2.24 – 10.54
Number of items already known	58.84	38.99	17 – 172
Number of new items presented	160.80	26.30	111 – 216
Number of items learned	19.18	17.90	4 – 84
Average number of presentations until word learned	2.36	0.99	1.00 – 5.52
Mean time until items learned (in min)	23.00	6.94	10.63 – 49.38
Number of items learned per hour	5.16	4.36	0.89 – 23.03

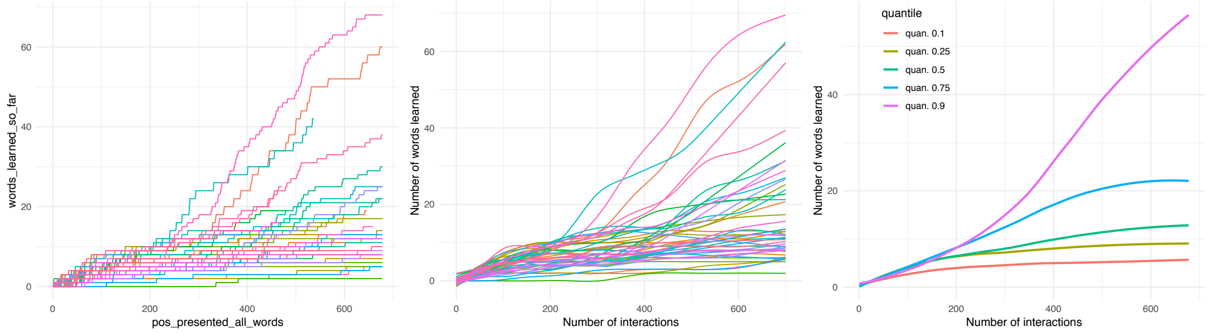


Figure 6: Plots of empirical growth trajectories (left), predicted growth trajectories from cubic model (center), and average predicted vocabulary growth at the 10th, 25th, 50th, 75th, and 90th percentiles.

and cubic change), LexTALE (sig. quadratic change), NFC (sig. linear, quadratic and cubic change); VLT (sig. quadratic change); Cognitive: Ericsen-Flanker (sig. linear, quadratic and cubic change); Personality-related: Openness (sig. linear, quadratic and cubic change), Extraversion (sig. quadratic and cubic change), Agreeableness (sig. quadratic and cubic change), Neuroticsm (sig. linear, quadratic and cubic change). No effects were found for the conscientiousness personality trait and the cognitive predictors AVA, AVN, VNA, pSRT, RSPAN, Simon, and Stroop. The NFC score was the strongest single predictor of linear, quadratic and cubic growth (all $p < .01$, see Table 3), indicating that participants with higher NFC scores exhibited significantly faster rates of increase, relative to participants with lower NFC scores. The best fitting (minimal adequate) model contained the participants scores on the Need for Cognition (NFC) scale as well as scores on the Openness to Experience personality trait. This model indicated that learning rates were significantly associated with the openness personality trait even after controlling for the effects of L2 experience, such that individuals with high openness scores showed faster learning rates. These effects are visualized in Figure 7, which shows that the trajectories of vocabulary growth began to separate early (around 200 presented items) based on whether or not the participant has a high or low NFC score. The effect of openness became apparent after 400 presentations, where individuals with lower scores level-off while the culmulative vocabulary of individuals with higher scores kept increasing.

4 Discussion and Future Work

It is widely recognized that vocabulary skills play a critical role in people's lives and future prospects as they are shown to be strongly related to individuals' overall educational success and academic achievement (Hart and Risley, 1995; Townsend et al., 2012). As a consequence, research on vocabulary growth has emphasized the importance of understanding not only the causes of individual differences in vocabulary growth rates but also the consequences of acquiring vocabulary at different rates (Rowe et al., 2012; Duff et al., 2015). Much of the cognitive developmental research in the area of vocabulary growth has utilized cross-

Table 2: Descriptive statistics for all cognitive, personality-related and L2 experience-related individual difference measures investigated in this study.

Task	Dependent measure	Mean (SD)	Obs. range
Statistical Learning			
pSRT	Mean reaction time (RT) difference between improbable and probable trials (in sec) ($\Delta RT_{improbable} - RT_{probable}$)	0.04 (0.03)	-0.02 – 0.1
VNA	Percent correct (out of 32 2-alternative forced choice trails)	49.03 (10.01)	31.25 – 71.88
AVA	Percent correct (out of 36 2-alternative forced choice trails)	51.36 (12.77)	16.67 – 86.11
AVN	Percent correct (out of 36 2-alternative forced choice trails)	52.04 (11.94)	33.33 – 80.56
Verbal Working Memory			
RSPAN	Percentage of responses (out of 60) that were accurate*	68.16 (20.13)	11.11 – 96.67
Cognitive Control			
Ericsen-Flanker	Mean reaction time (RT) difference between congruent and incongruent items ($\Delta RT_{incongruent} - RT_{congruent}$)	0.07 (0.1)	-0.14 – 0.42
Simon	$\Delta RT_{incongruent} - RT_{congruent}$	0.07 (0.06)	-0.08 – 0.21
Stroop	$\Delta RT_{incongruent} - RT_{congruent}$	0.18 (0.21)	-0.19 – 1.13
Personality Traits			
Openness	For all 5 indicators: person-centered z-scores adjusted for differences in acquiescent response styles (‘yea-saying’ vs. ‘nay-saying’)	0.00 (0.68)	-1.53 – 1.05
Conscientiousnness		0.35 (0.45)	-0.65 – 1.28
Extraversion		-0.36 (0.59)	-1.28 – 1.20
Agreeablenees		0.07 (0.56)	-1.34 – 1.18
Neuroticism		0.28 (0.68)	-0.85 – 3.67
L2 Experience			
LexTALE$_{\text{English}}$	Average % correct	73.75 (10.16)	53.75 – 93.75
VLT	Num. correct (out of 150 items)	121.25 (20.28)	21.00 – 142.00
ART	Num. correctly identified authors minus foils marked (out of 81)	11.22 (6.59)	0.00 – 26.00
NFC	Avg. of responses to all items (out of 18) with negative polarity items reverse scored	3.60 (0.57)	2.06 – 4.72

**NOTE:* Responses on the RSPAN task were coded as accurate if participants recalled the final word and judged the sentence in which it had occurred correctly.

Table 3: Results of growth curve analysis - Estimates of fixed effects and Goodness of Fit for the unconditional cubic growth model (left) and models including the L2 experience predictor Need for Cognition (NFC; middle) and Openness to Experience personality trait (right). The variable 'Time' refers to the number of interactions with the AISLE system.

	Dependent variable: Number of words learned		
	Unconditional GCM	added NFC (L2 exp.)	added Openness (Pers.) (best-fitting model)
Constant	12.065*** (1.272)	17.657*** (2.073)	19.443*** (2.359)
Linear change	68.295*** (14.492)	193.456*** (34.185)	216.504*** (37.897)
Quadratic change	6.356 (5.780)	81.212*** (12.188)	91.081*** (11.791)
Cubic change	9.733* (5.119)	34.998*** (6.115)	38.739*** (6.912)
NFC		-5.501^{*} (2.997)	-3.815 (3.012)
NFC x Time		-99.283^{**} (46.577)	-59.411 (50.181)
NFC x Time^2		-62.305^{***} (13.738)	-38.427^{**} (18.165)
NFC x Time^3		-27.749^{***} (8.498)	-20.236^{**} (9.203)
Openness			-5.267^{*} (2.941)
Openness x Time			-85.196^{*} (47.702)
Openness x Time^2			-41.335^{***} (15.750)
Openness x Time^3			-13.417 (8.794)
Log Likelihood	$-97{,}698.190$	$-97{,}692.140$	$-97{,}688.800$
Note:			*p<0.1; **p<0.05; ***p<0.01

sectional methodologies to capture snapshots of children's competence at different stages. While cross-sectional studies are useful to describe vocabulary growth in the general population over time, only longitudinal studies can shed light on the pace and pattern of vocabulary development, i.e. estimate rates of growth. It is, thus, unfortunate that the bulk of discussions within the field of (both child and second) language acquisition favors a cross-sectional view of vocabulary development and, as a consequence, discussions about longitudinal research are scarce.

In the present paper we introduced an adaptive language learning system (AISLE) designed to track and accelerate academic vocabulary growth in university students. The extraction pipeline relied on NLP techniques to arrive at statistically relevant items ('optimal language input'). The learning algorithm was designed to adapt in real-time during learning sessions to match the student's progress and memory patterns ('optimal repetition intervals').

In a second step, we showcased how the dense, longitudinal data generated by the system can be utilized to understand the dynamics of individual vocabulary growth trajectories. To this end, we presented first results of a study on a group who engaged with the AISLE system in a laboratory setting for several hours across three sessions over a period of three weeks. The goal of the study was twofold. First, we aimed to make use of our dense longitudinal data to examine the pace and shape of vocabulary growth trajectories. Second, we aimed to understand the role that experience-related, cognitive and affective factors play in explaining variation in students' vocabulary acquisition. We began by fitting the best longitudinal model to our dense observational data of vocabulary growth. We found that the empirical data were best represented by a cubic growth curve model. This result is consistent with the results reported in previous studies on children's vocabulary growth (e.g. Ganger and Brent, 2004; Rowe et al., 2012), suggesting that the vocabulary growth trajectories exhibit similar shapes across different learning contexts. The cubic model indicates that, on average, users increased their vocabulary size by approximately two words per 100 presented vocabulary items and increased their vocabulary by about 15 words in the course of a three-hour period of engagement with the system. There was, however, substantial variation in vocabulary growth with in-

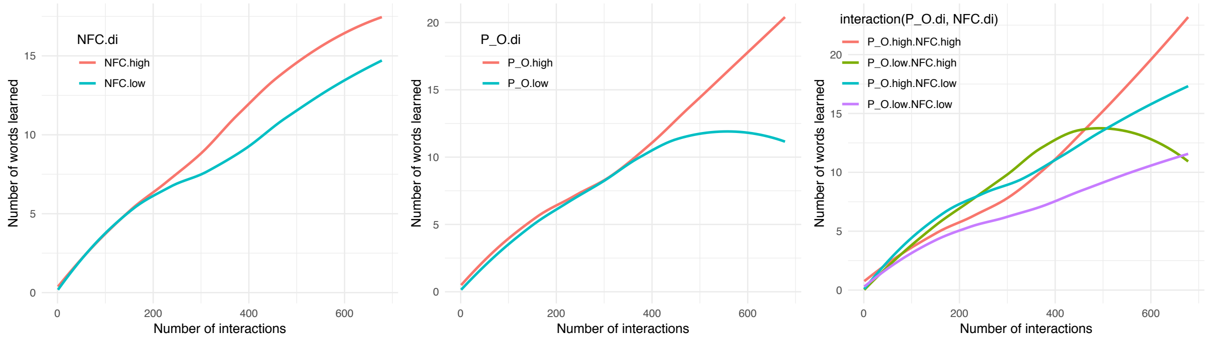


Figure 7: Predicted growth trajectories for participants with higher or lower NFC scores (left) and higher and lower scores on the openness to experience personality dimension (center). The plot on the right displays the results of the final model containing the effects of both NFC and Openness.

dividuals above the 90^{th} percentile reaching an estimated vocabulary growth of about 50 words after 600 presented vocabulary items, whereas individuals below the the 10^{th} percentile acquired about 5 words overall. Considerable between-subject variation was also observed for all other engagement and performance indicators collected by the system (see Table 1). To achieve our second goal, we next incorporated a total of 17 experience-related, cognitive and affective predictors measured into this growth model to examine whether and to what extent they affected the velocity (linear change) and acceleration (quadratic change) of learners' vocabulary growth. We found that - when considered on their own - 9 out of 17 IDs factors (four experience-related, four affective, and one cognitive factor) were significantly associated with vocabulary development. The best-fitting (minimal adequate) model assessing the joint effects of the IDs factors indicated (1) that participants with higher scores on the NFC experience proxy measure exhibited significantly faster rates of increase, relative to participants with lower NFC scores, and (2) that individuals with higher scores on the openness personality scale show faster learning rates – relative to those with lower scores on that scale. These results contribute to and expand the existing literature on the role of individual differences in second language acquisition (Dörnyei and Skehan, 2008; Ellis, 2004; Dewaele, 2009).

In conclusion, advancing our understanding of the dynamics of vocabulary growth is of central importance. There is a growing awareness in the cognitive sciences that an adequate theoretical model of language acquisition should be first and foremost constrained by empirical demonstrations of IDs as well as predict and account for the complex interrelationships between variation in the quantity and quality of language input, cognitive and affective factors in language development and attainment (for a recent review, see Kidd et al., 2017). The data obtained from an adaptive language learning system such as AISLE have the potential to transform our current understanding of vocabulary growth and to provide a new window into the mechanisms and principles underlying language development in general.

References

John AE Anderson, Lorinda Mak, Aram Keyvani Chahi, and Ellen Bialystok. 2018. The language and social background questionnaire: Assessing degree of bilingualism in a diverse population. *Behavior Research Methods*, 50(1):250–263.

Inbal Arnon and Morten H Christiansen. 2017. The role of multiword building blocks in explaining l1–l2 differences. *Topics in Cognitive Science*, 9(3):621–636.

Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, et al. 2014. lme4: Linear mixed-effects models using eigen and s4. *R package version*, 1(7):1–23.

Kimberly Becker and Phuong Nguyen. 2017. Review of technology-enhanced language learning for specialized domains: Practical applications and mobility. In Elorza I. Martin-Monje, E. and B. G. Riaza, editors, *Technology-enhanced language learning for specialized domains: Practical applications and mobility*, volume 21, page 6771.

Andrew Biemiller. 1999. *Language and reading success*, volume 5. Brookline Books.

John T Cacioppo, Richard E Petty, and Chuan Feng Kao. 1984. The efficient assessment of need for cognition. *Journal of Personality Assessment*, 48(3):306–307.

Kate Cain and Jane Oakhill. 2011. Matthew effects in young readers: Reading comprehension and reading experience aid vocabulary development. *Journal of Learning Disabilities*, 44(5):431–443.

Averil Coxhead. 2000. A new academic word list. *TESOL quarterly*, 34(2):213–238.

Jean-Marc Dewaele. 2009. Individual differences in second language acquisition. In Ritchie W. C. and Bhatia T.K., editors, *The new handbook of second language acquisition*, pages 623–646. Emerald Insight Bingley, England.

Stephen J Dollinger, Anna M Matyja, and Jamie L Huber. 2008. Which factors best account for academic success: Those which college students can control or those they cannot? *Journal of Research in Personality*, 42(4):872–885.

Zoltán Dörnyei and Peter Skehan. 2008. Individual differences in second language learning. In *The handbook of second language acquisition*, chapter 18, pages 589–630. Wiley-Blackwell.

Fiona J Duff, Gurpreet Reen, Kim Plunkett, and Kate Nation. 2015. Do infant vocabulary skills predict school-age language and literacy outcomes? *Journal of Child Psychology and Psychiatry*, 56(8):848–856.

Rod Ellis. 2004. *Individual differences in second language learning*. Blackwell Publishing.

Thomas A Farmer, Alex B Fine, Jennifer B Misyak, and Morten H Christiansen. 2017. Reading span task performance, linguistic experience, and the processing of unexpected syntactic events. *The Quarterly Journal of Experimental Psychology*, 70(3):413–433.

Jennifer Ganger and Michael R Brent. 2004. Reexamining the vocabulary spurt. *Developmental Psychology*, 40(4):621.

Dee Gardner and Mark Davies. 2013. A new academic vocabulary list. *Applied Linguistics*, 35(3):305–327.

Judith C Goodman and Elizabeth Bates. 2013. On the emergence of grammar from the lexicon. In *The emergence of language*, pages 47–98. Psychology Press.

Betty Hart and Todd R Risley. 1995. *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing.

Jacy Ippolito, Jennifer L Steele, and Jennifer F Samson. 2008. Introduction: Why adolescent literacy matters now. *Harvard Educational Review*, 78(1):1–6.

Heba M Ismail, Saad Harous, and Boumediene Belkhouche. 2016. Review of personalized language learning systems. In *2016 12th International Conference on Innovations in Information Technology (IIT)*, pages 1–6. IEEE.

Vickki Jacobs. 2008. Adolescent literacy: Putting the crisis in context. *Harvard Educational Review*, 78(1):7–39.

Oliver P John, Laura P Naumann, and Christopher J Soto. 2008. Paradigm shift to the integrative big five trait taxonomy. *Handbook of Personality: Theory and research*, 3(2):114–158.

Scott Barry Kaufman, Colin G DeYoung, Jeremy R Gray, Luis Jiménez, Jamie Brown, and Nicholas Mackintosh. 2010. Implicit learning as an ability. *Cognition*, 116(3):321–340.

Evan Kidd, Seamus Donnelly, and Morten H Christiansen. 2017. Individual differences in language acquisition and processing. *Trends in Cognitive Sciences*, pages 154–169.

Kristin Lemhöfer and Mirjam Broersma. 2012. Introducing lextale: A quick and valid lexical test for advanced learners of english. *Behavior Research Methods*, 44(2):325–343.

Clas Linnman, Per Carlbring, Åsa Åhman, Håkan Andersson, and Gerhard Andersson. 2006. The Stroop effect on the internet. *Computers in Human Behavior*, 22(3):448–455.

Viorica Marian, Henrike K Blumenfeld, and Margarita Kaushanskaya. 2007. The language experience and proficiency questionnaire (leap-q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research.*

James Milton. 2013. Measuring the contribution of vocabulary knowledge to proficiency in the four skills. *C. Bardel, C. Lindqvist, & B. Laufer (Eds.) L*, 2:57–78.

Daniel Mirman. 2017. *Growth curve analysis and visualization using R*. Chapman and Hall/CRC.

Valerie Muter, Charles Hulme, Margaret J Snowling, and Jim Stevenson. 2004. Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: evidence from a longitudinal study. *Developmental Psychology*, 40(5):665.

William Nagy and Dianna Townsend. 2012. Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, 47(1):91–108.

ISP Nation. 1993. Vocabulary size, growth, and use. *The bilingual lexicon*, pages 115–134.

Jonathan Peirce, Jeremy R Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. 2019. Psychopy2: Experiments in behavior made easy. *Behavior Research Methods*, pages 1–9.

Ana Pellicer-Sánchez. 2018. Examining second language vocabulary growth: Replications of schmitt (1998) and webb & chang (2012). *Language Teaching*, pages 1–12.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Treena Eileen Rohde and Lee Anne Thompson. 2007. Predicting academic achievement with cognitive ability. *Intelligence*, 35(1):83–92.

Meredith L Rowe, Stephen W Raudenbush, and Susan Goldin-Meadow. 2012. The pace of vocabulary growth helps predict later vocabulary skill. *Child Development*, 83(2):508–525.

Norbert Schmitt, Diane Schmitt, and Caroline Clapham. 2001. Developing and exploring the behaviour of two new versions of the vocabulary levels test. *Language Testing*, 18(1):55–88.

Noam Siegelman and Ram Frost. 2015. Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, 81:105–120.

Kendra R Tannenbaum, Joseph K Torgesen, and Richard K Wagner. 2006. Relationships between word knowledge and reading comprehension in third-grade children. *Scientific Studies of Reading*, 10(4):381–398.

Dianna Townsend, Alexis Filippini, Penelope Collins, and Gina Biancarosa. 2012. Evidence for the importance of academic word knowledge for the academic achievement of diverse middle school students. *The Elementary School Journal*, 112(3):497–518.

Mehrnoosh Vahdat, Luca Oneto, Davide Anguita, Mathias Funk, and Matthias Rauterberg. 2016. Can machine learning explain human learning? *Neurocomputing*, 192:14–28.

Ludo Verhoeven, Jan van Leeuwe, and Anne Vermeer. 2011. Vocabulary growth and reading development across the elementary school years. *Scientific Studies of Reading*, 15(1):8–25.

Ludo Verhoeven and Jan Van Leeuwe. 2008. Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 22(3):407–423.

Richard F West, Keith E Stanovich, and Harold R Mitchell. 1993. Reading in the real world and its correlates. *Reading Research Quarterly*, pages 35–50.

Oliver Wilhelm, Andrea Hildebrandt Hildebrandt, and Klaus Oberauer. 2013. What is working memory capacity, and how can we measure it? *Frontiers in psychology*, 4:433.