Assessing Language Learners' Free Productive Vocabulary with Hidden-task-oriented Dialogue Systems

Dolça Tellols
Tokyo Institute of Technology
Tokyo, Japan
tellols.d.aa@m.titech.ac.jp

Takenobu Tokunaga
Tokyo Institute of Technology
Tokyo, Japan
take@c.titech.ac.jp

Hilofumi Yamamoto
Tokyo Institute of Technology
Tokyo, Japan
yamagen@ila.titech.ac.jp

ABSTRACT

This paper proposes a new task to assess language learners' free productive vocabulary, which is related to being able to articulate certain words without getting explicit hints about them. To perform the task, we propose the use of a new kind of dialogue systems which induce learners to use specific words during a natural conversation to assess if they are part of their free productive vocabulary. Though systems have a task, it is hidden from the users. Consequently, these may consider systems task-less. Because these systems do not fall into the existing categories for dialogue systems (task-oriented and non-task-oriented), we named them as hidden-task-oriented dialogue systems. To study the feasibility of our approach, we conducted three experiments. The Question Answering experiment evaluated how easily learners could recall a target word from its dictionary gloss. Through the Wizard of Oz experiment, we confirmed that the proposed task is hard, but humans can achieve it to some extent. Finally, the Context Continuation experiment showed that a simple corpus-retrieval approach might not work to implement the proposed dialogue systems. In this work, we analyse the experiments results in detail and discuss the implementation of dialogue systems capable of performing the proposed task.

CCS CONCEPTS

• **Computing methodologies** → **Intelligent agents**; • **Applied computing** → **Education**.

KEYWORDS

Computer Aided Language Learning, Dialogue Systems, Productive Vocabulary

ACM Reference Format:

Dolça Tellols, Takenobu Tokunaga, and Hilofumi Yamamoto. 2020. Assessing Language Learners' Free Productive Vocabulary with Hidden-task-oriented Dialogue Systems. In *IUI '20 Workshops, March 17, 2020, Cagliari, Italy*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Second language (L2) learning has attracted much attention in recent years since revitalised Artificial Intelligence (AI) research opened the door to the possibility of more sophisticated Intelligent Computer Assisted Language Learning (ICALL) [18]. Among others, vocabulary assessment by computers has been an active research area with studies focusing on the automatic generation of vocabulary evaluation questions [3] [8] or the measurement of vocabulary size through computerised adaptive testing (CAT) [27]. However, these studies concerned the assessment of *receptive vocabulary*, which is used to comprehend texts or utterances. In contrast, there is a lack of studies on the computerised assessment of *productive vocabulary*, which is used to speak and write [29]. From the viewpoint of linguistic proficiency, receptive vocabulary is related to language understanding and productive vocabulary to language production. It is said that there is a gap between understanding the meaning of a particular word (passive or receptive vocabulary) and being able to articulate it (active or productive vocabulary) [12].

Although there exist many approaches to evaluate receptive vocabulary, studies that focus on the assessment of productive vocabulary are scarce. Meara et al. [17] and Laufer et al. [13], who propose the *Lex30* task and the *LFP* (Lexical Frequency Profile) measure respectively, are two exceptions. Lex30 is a word association task where learners have to provide words given another word stimulus. LFP measures vocabulary size based on the proportion of words in different vocabulary-frequency levels that learners use in their writing.

It is considered that productive ability may comprise different degrees of knowledge. We refer to the ability to use a word at one's free will as *free productive ability*, while *controlled productive ability* refers to the ability to use a word when driven to do so [14]. Fill-in-the-blank tasks evaluate controlled productive ability and, though the Lex30 task wants to asses free productive ability, stimulus words make it controlled to some extent. We can use the Lexical Frequency Profile to measure free productive vocabulary size, but it is unable to determine if learners are capable of freely using specific words.

We may ideally assess free productive ability in conversational contexts but this complicates, even more, the design of tasks for this purpose. Speaking tests used in language certification exams are one option to overcome this deficiency, but they require human resources for the evaluation and hardly specify words to test if learners can use them. Suendermann-Oeft et al. [25] tried to solve the human resource problem by replacing the evaluators with a multi-modal dialogue system, but they do not provide solutions to the latter, the evaluation of specific words.

Against this backdrop, the present work proposes a new task for dialogue systems to evaluate free productive vocabulary by inducing learners to naturally use the words to assess during a conversation without providing explicit hints about them. Our hypothesis for the assessment is that a certain set of words forms

part of people's free productive vocabulary if they can naturally use those words in a conversation without having been asked explicitly to do so.

Dialogue systems are usually divided into two categories: task-oriented and non-task-oriented. Systems capable of performing the proposed task can be considered non-task-oriented from the user point of view and task-oriented from the system point of view (though the task is hidden from the user). Given the asymmetrical nature of the proposed systems, it is hard to fit them into one of the available categories. Consequently, we propose a new one named *hidden-task-oriented* dialogue systems. We will further explain this new category in section 4.

In our previous work, we briefly presented the proposed task and investigated some of the difficulties that its implementation may have to deal with [26]. In this work, we review the experiments and expand them. Additionally, we analyse the requirements for the design of dialogue systems capable of performing the proposed task and discuss the techniques that we may use for their implementation, which we leave as future work.

2 RELATED WORK

Recent studies on vocabulary assessment concern various aspects, e.g. asking words with or without a context, and different forms of questions, e.g. multiple-choice or fill-in-the-blank questions [23, 24]. Others also point out the importance of domain when assessing lexical knowledge [21]. We focus on the distinction between receptive and productive vocabulary and, more specifically, propose a new method for assessing language learners' free productive vocabulary through dialogue systems.

Laufer and Nation [14] proposed evaluating controlled productive vocabulary by using sentence completion tasks where they gave some initial letters of the target word. However, this technique is controversial because it may assess receptive vocabulary instead of productive vocabulary as they provided a hint to guess the target words [19]. Others used translation tasks that ask learners to translate L1 (mother tongue) expressions into L2 (language being learned) [29]. The problem of this approach is that they need to adapt tests according to the L1 language of the learners. Moreover, target words need to be chosen carefully to ensure that learners use the expected target word and not a synonym. In our proposal, we do not plan on giving any explicit hints for the target words and neither need adaptation according to the L1, since dialogues will be directly in the L2.

Regarding computer-assisted vocabulary assessment, Brown at al. [3] and Heilman and Eskenazi [8] studied the automatic generation of vocabulary assessment questions and Tseng [27] focused on the measurement of English learners' vocabulary size. Allen and McNamara [1] utilised Natural Language Processing (NLP) tools to analyse the lexical sophistication of learners' essays to estimate their vocabulary size. They also pointed out the importance of providing personalised instructions to each learner. We take this aspect into account by controlling dialogue topics according to the learner's interests and the words being assessed.

Fryer and Carpenter [6] discuss the possibility of utilising dialogue systems in language education. Nowadays, many language learning commercial applications provide conversations with chatbots, e.g. Duolingo Bots[1], Andy[2], Mondly[3] and Eggbun Education[4]. However, most of them base their interactions on predefined answers or have a rigidly guided task-oriented dialogue. Research level systems are more versatile than commercial ones. As an example, Genie tutor [10] is a dialogue-based language learning system that is designed for native Korean speakers to learn English. It accepts free text input in a given scenario and can respond by retrieving utterances in a dialogue corpus based on their context similarity. Höhn [9] introduces an Artificial Intelligence Markup Language (AIML)-based chat-bot for conversational practice, which recognises repair initiations and generates repair carry-outs. And Wilske [30] also examines how NLP, particularly dialogue systems, can contribute to language learning. In her dialogue system, learners can receive feedback on their utterances.

Research on automated language proficiency evaluation through dialogue is scarce. Some studies include the assessment of the verbal skill of English learners through task-oriented dialogues [15] or through simulated conversations [5]. There is also an already mentioned proposal of a multimodal dialogue system for the evaluation of English learners' speech capabilities [25]. Our contribution is proposing a new free productive vocabulary assessment methodology in the form of a new task for dialogue systems. Because our dialogue systems do not fall into any of the existing categories (task-oriented and non-task-oriented), we propose a new one named *hidden-task-oriented* dialogue systems.

3 PROPOSED TASK TO ASSESS FREE PRODUCTIVE VOCABULARY

3.1 Hypotheses

This work takes base on the following hypothesis: "If a person can naturally use a certain word during a conversation, we can assume that it belongs to their free productive vocabulary".

3.2 Task goal

Taking into consideration this hypothesis, we propose a new task for dialogue systems (DS) that will be used to evaluate free productive vocabulary. The goal of the task is inducing learners to naturally use certain target words (TWs) during a conversation by generating an appropriate dialogue context. Directly asking the words or providing explicit hints about them is prohibited. Figure 1 illustrates appropriate and inappropriate examples of the DS behaviour.

To motivate this task goal, we took inspiration from a theory about second language acquisition called the *Natural Approach* [11]. This theory states that conversation is the base of language learning. As our proposal is a task for dialogue systems, it follows its main principle.

There is also a technique some teachers use, named dialogue journals, which also relates to our proposal. Peyton [22] describes dialogue journals as written conversations between a teacher and

[1] http://bots.duolingo.com
[2] https://andychatbot.com
[3] https://www.mondly.com
[4] https://web.eggbun.net

Appropriate	Inappropriate
S: I think I want to travel somewhere. What would you recommend to me?	S: How do you call a railway vehicle that is self-propelled on a track carrying people and luggage thanks to electricity?
L: You could go to London.	
S: Nice idea! How could I get there?	
L: I think nowadays you can go by plane or *train*.	L: A *train*.

S: system, L: learner

Figure 1: Appropriate and inappropriate dialogue examples of the proposed task (TW: "train")

a student, where the teacher avoids acting as an evaluator. Baudrand [2] researched the impact of using this technique in a foreign language class where students had to communicate through the diaries in the target language. While journals are closer to exchanging letters without a clear evaluation purpose, we propose the use of real-time written conversations aiming at the assessment of specific terms.

4 HIDDEN-TASK-ORIENTED DIALOGUE SYSTEMS

Though there is a huge variety of dialogue systems deployed, they are usually classified into one of the two categories: task-oriented and non-task-oriented.

Task-oriented dialogue systems are usually topic-constrained and their goal is to help the user achieve a certain task. Into this category fall reservation, shopping or personal assistant systems like Apple's Siri[5] or Google Assistant[6].

On the other hand, non-task-oriented (or conversational) systems are commonly chit-chat dialogue systems whose only purpose is to keep the conversation with the user ongoing. Conversations are usually not restrained to a certain topic; they are considered open-domain or free. Consequently, if systems want to provide informative responses, large amounts of data are necessary for their implementation. However, if that is not the case, conversations can easily keep going by giving generic answers that may make the user assume the system understanding. Some examples of this kind of systems include Microsoft's Japanese chatbot Rinna[7] or ALICE, a chatbot implemented using AIML (Artificial Intelligence Markup Language) [28].

To achieve the task proposed in section 3, we need dialogue systems such that:

- From the *user point of view*, since we are aiming for free topic chit-chat conversation, they look like a non-task-oriented dialogue system.
- From the *system point of view*, as the system has the goal of making the user use a certain target word during the dialogue, they are task-oriented dialogue systems. Their peculiarity is that the task is hidden from the user.

[5]https://www.apple.com/siri/
[6]https://assistant.google.com/
[7]https://www.rinna.jp/profile

Recently, storytelling dialogue systems are emerging [20]. They usually interact with the user to reach the end of a story plot, but dialogue can diverge during the process by getting questions or ideas from the user. Though they can be considered as a hybridisation of task-oriented and non-task-oriented systems and may resemble our proposed dialogue systems, there is a clear difference between them. During the flow of the dialogue, storytelling dialogue systems change between task-oriented and non-task-oriented interactions. However, our proposed systems always have the same kind of interaction, but they look different depending on the dialogue participant roles: the user vs. the system. Additionally, if we consider our systems in general, they have a clear task, with the peculiarity that this task is *hidden* from the user. Consequently, we do not consider that the term *hybrid* is appropriate enough and named our proposed systems *hidden-task-oriented* dialogue systems.

Note that Yoshida [31] also used the word 'hidden task' to describe the dialogue journals task referenced in section 3. Because the teacher responds naturally while keeping in mind the student's language ability and interests, what the teacher does can be considered a 'hidden task' from the user's point of view.

5 EXPERIMENTS AND RESULTS

5.1 Experimental design

To study the feasibility of the task and to analyse ideas for the implementation of hidden-task-oriented dialogue systems capable of achieving the proposed task, we conducted three different kinds of experiments.

The Question Answering (QA) experiment asks a word by providing learners with its definition, taken from a dictionary and turned into a question, as shown in the inappropriate example in Figure 1. This experiment is not assessing free productive but it serves us as a reference and shows how easily learners can recall a specific target word from their definition. Additionally, it can also help us detect if there are certain words harder to assess.

In the Wizard of Oz (WOZ) experiment, one of a pair plays the system role and tries to make their counterpart, playing the learner role, use the target word in their utterances. System role participants must not reveal their intention nor use the target word in their utterances. Learner role participants believe they are doing goal-less chatting. The dialogue, for which we did not set a time limit, can be terminated by anyone at any time and is performed through a text chat interface. The aim of this experiment is showing the difficulty of the proposed task for humans and gathering data that may serve to implement the proposed dialogue systems.

The Context Continuation (CC) experiment asks learners to estimate the next utterance given a dialogue context. We made the context by extracting a sequence of utterances from a human-human dialogue corpus so that the next utterance of the sequence (not shown in the experiment) includes the TW (see example in Figure 2). This experiment shows if such a corpus-retrieval approach might work for the implementation of the dialogue systems.

In all the experiments, tasks succeed if learners use the TWs.

B:	Aren't 3 books a little bit expensive?
A:	I don't think so.
B:	But it is quite a lot, right?
A:	*(utterance in the original corpus)* Well, but if the number of words increases, it makes sense that the *price* also increases.
A:	(*success*) I think their *price* is quite appropriate.
A:	(*failure*) I don't think so, but if you do, don't buy them.

upper: context, middle: corpus continuation, bottom: answer examples

Figure 2: CC experiment example (TW: "price")

5.2 Material

Language. Our target language is Japanese, but the methodology can apply to any language.

Target words. We decided six nouns as the TWs by the following criteria. Since we wanted to implement the CC experiment, we selected words that frequently appear in the Nagoya University Conversation Corpus [7], which consists of 129 transcripted dialogues by 161 persons with an approximate total duration of 100 hours. We chose words appearing in utterances with more than two and less than eleven preceding utterances, not counting the ones with less than four words if they did not contain a noun. We filtered out words categorised into N1 (the hardest) and N5 (the easiest) levels in terms of the Japanese Language Proficiency Test (JLPT), and further filtered out those having a one-word gloss as their definition in the employed dictionary [16]. We picked up these six words from the remaining ones: "*kao* (face)", "*syôgakkô* (primary school)", "*rokuon* (audio recording)", "*konpyûtâ* (computer)", "*tîzu* (cheese)" and "*fun'iki* (atmosphere)".

Participants. We recruited ten native Japanese speakers and divided them into two groups: *S* and *L*. Group *S* performed the QA experiment first and then played the system role in the WOZ experiment, while group *L* played the learner role in WOZ, and then, performed the CC experiment. Each pair performed six dialogues (one per target word). After every WOZ dialogue, group *L* evaluated the dialogue naturalness.

Group *S* answered six questions in the QA experiment (one per target word). We explicitly informed participants they should only rely on their knowledge and do not check any other external information source when providing the answers.

Group *L* continued eighteen contexts (three per target word) in the CC experiment.

Assuming that native speakers have large enough vocabulary, we can assess the feasibility of our approach itself.

Platform. We designed a system that consists of a Unity[8] application communicating with a Django[9] Python server to perform the experiments and gather the data. Participants accessed the system with a given username and password and the application automatically lead them to the appropriate experiment instructions screen. Figure 3 illustrates how dialogue took place in the WOZ experiment.

5.3 Results

Table 1: Results of the QA experiment

Target word	Success rate
"face"	5/5
"primary school"	5/5
"audio recording"	4/5
"computer"	1/5
"cheese"	1/5
"atmosphere"	3/5
Total	19/30

QA *experiment.* Table 1 shows the results of the QA experiment. The success rate ($19/30 = 63.3\%$) is rather low considering that participants are native speakers, i.e. they should know the target words. In addition, we can observe how the success rate differs across individual words. The gloss we used to ask the target word is written originally to explain the headword and not vice versa. This directionality may explain this low success rate. For instance, the gloss of "cheese" can be similar to that of other dairy products like yogurt and butter, which are examples of wrong answers given by the participants. From these, we can deduce, due to the same reason, how the gloss is not specific enough to identify the headword.

Table 2: Results of the WOZ experiment

	Success rate	Dialogue length (min)	Number of utterances	Naturalness (1–5)
"face"	1/5	16.4	35.0	3.6
"primary school"	3/5	14.4	41.2	4.0
"audio recording"	1/5	15.9	38.4	4.8
"computer"	0/4*	13.9	32.3	4.0
"cheese"	2/4*	14.9	22.3	3.4
"atmosphere"	3/5	13.0	26.4	4.4
Pair 1	1/5*	18.8	47.0	3.4
Pair 2	1/6	11.4	18.0	4.7
Pair 3	1/5*	21.1	64.0	3.8
Pair 4	4/6	7.3	18.8	4.7
Pair 5	3/6	13.7	19.8	3.3
Success		10.3	24.7	4.0
Failure		16.1	33.4	4.0
Total	10/28	14.1	30.8	4.0

Dialogue length, number of utterances and naturalness indicate the average value across dialogues.
Participants accidentally skipped two dialogues (*).

[8] https://unity.com/
[9] https://www.djangoproject.com/

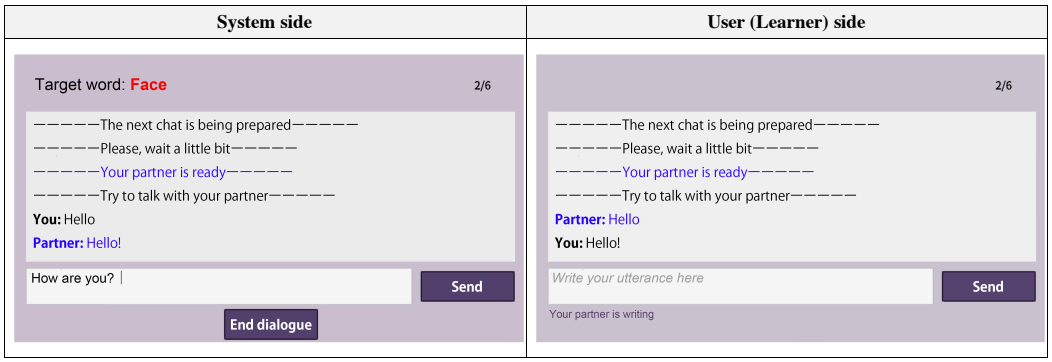

Figure 3: Screenshots of the application used to perform the WOZ experiment (translated from Japanese to English)

WOZ *experiment.* Table 2 shows the target word-wise (upper section), pair-wise (middle section) and success/failure-wise (bottom section) statistics of the WOZ experiment. The overall success rate ($10/28 = 35.7\%$) is lower than that of the QA experiment. This suggests that it is harder to make learners think about a specific word within a dialogue. The success rate across words is diverse, but it is not directly related to the word difficulty level. It is rather related to the abundance of synonyms. For instance, learner role participants used words like "PC" instead of "computer". Since we strictly required using the exact same word, such synonyms did not lead to success. When assessing learners' productive vocabulary, we need to decide what ability we evaluate, i.e. an ability to express a concept or that to use an exact word.

The middle section indicates the difference in performance among the pairs. Pair 4 and 5 performed better than Pair 1, 2 and 3. In particular, Pair 4 performed the best in terms of both dialogue length and dialogue naturalness. We should aim at realising a dialogue system that performs at least as well as Pair 4.

The bottom section indicates that there is no big difference in naturalness between successful and failed dialogues but failed dialogues tend to be longer. Note that we did not set a time limit for a dialogue in the present experiments and this sometimes leads to quite long conversations. The average failed dialogue length would be a good reference for the time limit in future experiments.

CC *experiment.* Lastly, there was no success case among 90 in the CC experiment. In terms of linguistic quality of utterances, the retrieval-based approach has an advantage, but it is hard to retrieve an appropriate context from a corpus of this size.

6 DISCUSSION

Reflections about the proposed task. The results of the WOZ experiment lead to reflections regarding the number of target words and the knowledge about the user. Concerning the number of target words, the current experiment systems (Wizards) focus on a single target word at a time. As we can see from the results, it is quite hard for systems to succeed in this scenario. One of the reasons is that having just a single target word constrains the freedom of the dialogue, i.e. restricts the choice of topics and the flow of the dialogue. Thus, it becomes difficult to induce the user to use the target word. For instance, when the system failed to induce the TW with a certain utterance, it should stick to the TW and try a different utterance (strategy) even though the current context may have varied and may be more related to different (potential target) words. We should redefine the proposed task such that the systems consider a pool of target words simultaneously during the dialogue. This pool could contain words from different difficulty levels and be updated dynamically according to the current conversation topic, word difficulty in user utterances and the already-achieved TWs. Based on the achieved TWs and their difficulty level, we may be able to assess the user's free productive vocabulary automatically. As for user profiles, they facilitate choosing appropriate dialogue topics. For example, given "graduation" as TW, knowing that the user has just graduated from a school makes it easier to bring a related topic into the conversation. Consequently, we should consider introducing user modelling into the proposed dialogue systems.

Gathering dialogue data. The results of the "Context Continuation (CC) experiment" suggest that the amount of available data is so limited that it is difficult to implement the proposed systems using a simple retrieval-based approach. We expected that the WOZ experiment would also serve to gather dialogue data which would be more appropriate to implement dialogue systems capable of performing the proposed task. During the arrangement of the WOZ experiment, however, we had difficulties in finding participants and matching them for the dialogue. There were also some problems during the data gathering process due to internet connection problems and platform instability. We plan on developing a simpler and more accessible system to avoid the manual search of participants. To cope with these problems in data gathering, we plan to implement and launch a gamified platform in which players (dialogue participants) will be automatically matched and try to compete to make their counterparts use the target words. In this gamified setting, each player takes both the learner and the system role.

Implementing dialogue systems with limited amounts of data. In our case, as users will be language learners, system utterances should be grammatically correct. Retrieval-based approaches are advantageous in this respect. As we did in the CC experiment, we can retrieve contexts from the dialogue corpora that are similar to the current context and precede an utterance that includes the target word. Then, we can use the previous utterance to the utterance that includes the target word as a system utterance. However,

insufficient dialogue data might prevent us from retrieving the contexts in the first place. We need to use query expansion techniques by considering synonyms and similar words of the target word to cope with this problem. The contexts retrieved by query expansion, however, might provide system utterances irrelevant to the current context at the lexical level. One possibility to solve this inappropriateness would be adopting the skeleton-to-response method [4], which replaces not-context-related words in the utterance with open slots (skeleton generation) and applies a generative model to fill the slots with appropriate words.

If we also consider implementing the pool of target words as mentioned above, we could retrieve a set of contexts for each target word in the pool in parallel. We then construct the system utterance from all contexts across the different target words. This method would increase the task success rate because we can choose the most appropriately-contextualised target word in the pool.

7 CONCLUSIONS AND FUTURE WORK

This paper proposed a novel task to assess language learners' free productive vocabulary. The task goal is making learners use a certain word in their utterances during a natural dialogue. It aims to verify if the word is in the vocabulary learners *use* (productive) rather than in the one they *understand* (receptive). To perform this task, we proposed a new category of dialogue systems, namely hidden-task-oriented dialogue systems. To study the feasibility of our proposal, we conducted three experiments, including one employing the WOZ approach. The experiments showed that the proposed task is more difficult than a simple QA task to answer the target word but can be achieved by humans to some extent. The results made us reflect on the proposed task and gave us hints for redesigning the task. Because we noticed how insufficient dialogue data causes problems in the implementation of the systems, particularly when adopting the retrieval-based approach, we proposed two possible solutions. One option is gathering additional dialogue data through a gamified data gathering platform. The other one is enhancing retrieval-based approaches with techniques like query expansion and template-filling

Our future work includes the implementation and evaluation of the proposed dialogue systems. We would also like to develop and deploy a gamified approach to gather more dialogue data. Finally, we also need to investigate how to appropriately create a pool of target words for the systems and implement the mechanism that will adjust them dynamically during the conversations.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP19H04167.

REFERENCES

[1] Laura K Allen and Danielle S McNamara. 2015. You Are Your Words: Modeling Students' Vocabulary Knowledge with Natural Language Processing Tools. *International Educational Data Mining Society* (2015).

[2] Lynn Patricia Baudrand-aertker. 1992. Dialogue Journal Writing in a Foreign Language Classroom: Assessing Communicative Competence and Proficiency. (1992).

[3] Jonathan C Brown, Gwen A Frishkoff, and Maxine Eskenazi. 2005. Automatic question generation for vocabulary assessment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 819–826.

[4] Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2019. Skeleton-to-response: Dialogue generation guided by retrieval memory. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 1219–1228.

[5] Keelan Evanini, Sandeep Singh, Anastassia Loukina, Xinhao Wang, and Chong Min Lee. 2015. Content-based automated assessment of non-native spoken language proficiency in a simulated conversation. In *NIPS Workshop on Machine Learning for Spoken Language Understanding and Interaction*.

[6] Luke Fryer and Rollo Carpenter. 2006. Emerging Technologies. *Language Learning & Technology* 10, 3 (2006), 8–14.

[7] Itsuko Fujimura, Shoju Chiba, and Mieko Ohso. 2012. Lexical and grammatical features of spoken and written Japanese in contrast: Exploring a lexical profiling approach to comparing spoken and written corpora. In *Proceedings of the VIIth GSCP International Conference. Speech and Corpora*. 393–398.

[8] Michael Heilman and Maxine Eskenazi. 2007. Application of automatic thesaurus extraction for computer generation of vocabulary questions. In *Workshop on Speech and Language Technology in Education*.

[9] Sviatlana Höhn. 2017. A data-driven model of explanations for a chatbot that helps to practice conversation in a foreign language. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. 395–405.

[10] Jin-Xia Huang, Kyung-Soon Lee, Oh-Woog Kwon, and Young-Kil Kim. 2017. A chatbot for a dialogue-based second language learning system. *CALL in a climate of change: adapting to turbulent global conditions* (2017), 151.

[11] Stephen D Krashen and Tracy D Terrell. 1983. *The natural approach: Language acquisition in the classroom*. Alemany Press.

[12] Batia Laufer and Zahava Goldstein. 2004. Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language learning* 54, 3 (2004), 399–436.

[13] Batia Laufer and Paul Nation. 1995. Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics* 16, 3 (1995), 307–322.

[14] Batia Laufer and Paul Nation. 1999. A vocabulary-size test of controlled productive ability. *Language testing* 16, 1 (1999), 33–51.

[15] Diane Litman, Steve Young, Mark Gales, Kate Knill, Karen Ottewell, Rogier van Dalen, and David Vandyke. 2016. Towards using conversations with spoken dialogue systems in the automated assessment of non-native speakers of english. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 270–275.

[16] Akira Matsumura. 2010, 2013. *Super Daijirin Japanese Dictionary*. Sanseido Co., Ltd.

[17] Paul Meara and Tess Fitzpatrick. 2000. Lex30: An improved method of assessing productive vocabulary in an L2. *System* 28, 1 (2000), 19–30.

[18] Detmar Meurers and Markus Dickinson. 2017. Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning* 67, S1 (2017), 66–95.

[19] John Morton. 1979. Word recognition. *Psycholinguistics: Series 2. Structures and processes* (1979), 107–156.

[20] Leire Ozaeta and Manuel Graña. 2018. A View of the State of the Art of Dialogue Systems. In *International Conference on Hybrid Artificial Intelligence Systems*. Springer, 706–715.

[21] P David Pearson, Elfrieda H Hiebert, and Michael L Kamil. 2007. Vocabulary assessment: What we know and what we need to learn. *Reading research quarterly* 42, 2 (2007), 282–296.

[22] Joy Kreeft Peyton. 1997. Dialogue journals: Interactive writing to develop language and literacy. *Teacher Librarian* 24, 5 (1997), 46.

[23] John Read. 2007. Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies* 7, 2 (2007), 105–126.

[24] Katherine A Dougherty Stahl and Marco A Bravo. 2010. Contemporary classroom vocabulary assessment for content areas. *The Reading Teacher* 63, 7 (2010), 566–578.

[25] David Suendermann-Oeft, Vikram Ramanarayanan, Zhou Yu, Yao Qian, Keelan Evanini, Patrick Lange, Xinhao Wang, and Klaus Zechner. 2017. A Multimodal Dialog System for Language Assessment: Current State and Future Directions. *ETS Research Report Series* 2017, 1 (2017), 1–7.

[26] Dolça Tellols, Hitoshi Nishikawa, and Takenobu Tokunaga. 2019. Dialogue Systems for the Assessment of Language Learners' Productive Vocabulary. In *Proceedings of the 7th International Conference on Human-Agent Interaction*. ACM, 223–225.

[27] Wen-Ta Tseng. 2016. Measuring English vocabulary size via computerized adaptive testing. *Computers & Education* 97 (2016), 69–85.

[28] Richard S Wallace. 2009. The Anatomy of A.L.I.C.E. In *Parsing the Turing Test*. Springer, 181–210.

[29] Stuart Webb. 2008. Receptive and productive vocabulary sizes of L2 learners. *Studies in Second language acquisition* 30, 1 (2008), 79–95.

[30] Sabrina Wilske. 2015. *Form and meaning in dialog-based computer-assisted language learning*. Ph.D. Dissertation. Universität des Saarlandes.

[31] Kayo Yoshida et al. 2012. Genre-based Tasks and Process Approach in Foreign Language Writing. *Language and Culture: The Journal of the Institute for Language and Culture* 16 (2012), 89–96.