A CALL System for Learning Preposition Usage

John Lee
Department of
Linguistics and Translation
City University of Hong Kong
jsylee@cityu.edu.hk

Donald Sturgeon
Fairbank Center
for Chinese Studies
Harvard University
djs@dsturgeon.net

Mengqi Luo
Department of
Linguistics and Translation
City University of Hong Kong
mengqluo@cityu.edu.hk

Abstract

Fill-in-the-blank items are commonly featured in computer-assisted language learning (CALL) systems. An item displays a sentence with a blank, and often proposes a number of choices for filling it. These choices should include one correct answer and several plausible distractors. We describe a system that, given an English corpus, automatically generates distractors to produce items for preposition usage.

We report a comprehensive evaluation on this system, involving both experts and learners. First, we analyze the difficulty levels of machine-generated carrier sentences and distractors, comparing several methods that exploit learner error and learner revision patterns. We show that the quality of machine-generated items approaches that of human-crafted ones. Further, we investigate the extent to which mismatched L1 between the user and the learner corpora affects the quality of distractors. Finally, we measure the system's impact on the user's language proficiency in both the short and the long term.

1 Introduction

Fill-in-the-blank items, also known as gap-fill or cloze items, are a common form of exercise in computer-assisted language learning (CALL) applications. Table 1 shows an example item designed for teaching English preposition usage. It contains a sentence, "The objective is to kick the ball into the opponent's goal", with the preposition "into" blanked out; this sentence serves as the *stem* (or *carrier sentence*). It is followed by four choices for the blank, one of which is the *key* (i.e., the correct answer), and the other three are distractors. These choices enable the CALL application to provide immediate and objective feedback to the learner.

A high-quality item must meet multiple requirements. It should have a stem that is fluent and matches the reading ability of the learner; a blank that is appropriate for the intended pedagogical goal; exactly one correct answer among the choices offered; and finally, a number of distractors that seem plausible to the learner, and yet would each yield an incorrect sentence. Relying on language teachers to author these items is time consuming. Automatic generation of these items would not only expedite item authoring, but also potentially provide personalized items to suit the needs of individual learners. This paper addresses two research topics:

- How do machine-generated items compare with human-crafted items in terms of their quality?
- Do these items help improve the users' language proficiency?

For the first question, we focus on automatic generation of preposition distractors, comparing three different methods for distractor generation. One is based on word co-occurrence in standard

The objective is to kick the ball _____ the opponent's goal. (A) in (B) into (C) to (D) with

Table 1: An automatically generated fill-in-the-blank item, where "into" is the key, and the other three choices are distractors.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 984–993,
Berlin, Germany, August 7-12, 2016.

corpora; a second leverages error annotations in learner corpora; the third, a novel method, exploits learners' revision behavior. Further, we investigate the effect of tailoring distractors to the user's native language (L1). For the second question, we measure users' performance in the short and in the long term, through an experiment involving ten subjects, in multiple sessions tailored to their proficiency and areas of weakness.

Although a previous study has shown that learner error statistics can produce competitive items for prepositions on a narrow domain (Lee and Seneff, 2007), a number of research questions still await further investigation. Through both expert and learner evaluation, we will compare the quality of carrier sentences and the plausibility of automatically generated distractors against human-crafted ones. Further, we will measure the effect of mismatched L1 between the user and the learner corpora, and the short- and long-term impact on the user's preposition proficiency. To the best of our knowledge, this paper offers the most detailed evaluation to-date covering all these aspects.

The rest of the paper is organized as follows. Section 2 reviews previous work. Section 3 outlines the algorithms for generating the fill-in-the-blank items. Section 4 gives details about the experimental setup and evaluation procedures. Section 5 analyzes the results. Section 6 concludes the paper.

2 Previous Work

2.1 Distractor generation

Most research effort on automatic generation of fill-in-the-blank items has focused on vocabulary learning. In these items, the key is typically from an open-class part-of-speech (POS), e.g., nouns, verbs, or adjectives.

To ensure that the distractor results in an incorrect sentence, the distractor must rarely, or never, collocate with other words in the carrier sentence (Liu et al., 2005). To ensure the plausibility of the distractor, most approaches require it to be semantically close to the key, as determined by a thesaurus (Sumita et al., 2005; Smith et al., 2010), an ontology (Karamanis et al., 2006), rules handcrafted by experts (Chen et al., 2006), or context-sensitive inference rules (Zesch and Melamud, 2014); or to have similar word frequency (Shei, 2001; Brown et al., 2005). Sakaguchi et al. (2013) applied machine learning methods to select verb distractors, and showed that they resulted in items that can better predict the user's English proficiency level.

Less attention has been paid to items for closed-class POS, such as articles, conjunctions and prepositions, which learners also often find difficult (Dahlmeier et al., 2013). For these POS, the standard algorithms based on semantic relatedness for open-class POS are not applicable. Lee and Seneff (2007) reported the only previous study on using learner corpora to generate items for a closed-class POS. They harvested the most frequent preposition errors in a corpus of Japanese learners of English (Izumi et al., 2003), but performed an empirical evaluation with native Chinese speakers on a narrow domain.

We expand on this study in several dimensions. First, carrier sentences, selected from the general domain rather than a specific one, will be analyzed in terms of their difficulty level. Second, distractor quality will be evaluated not only by learners but also by experts, who give scores based on their plausibility; in contrast to most previous studies, their quality will be compared with the human gold standard. Thirdly, the effect of mismatched L1 will also be measured.

2.2 Learner error correction

There has been much recent research on automatic correction of grammatical errors. Correction of preposition usage errors, in particular, has received much attention. Our task can be viewed as the inverse of error correction — ensuring that the distractor yields an incorrect sentence — with the additional requirement on the plausibility of the distractor.

Most approaches in automatic grammar correction can be classified as one of three types, according to the kind of statistics on which the system is trained. Some systems are trained on examples of correct usage (Tetreault and Chodorow, 2008; Felice and Pulman, 2009). Others are trained on examples of pairs of correct and incorrect usage, either retrieved from error-annotated learner corpora (Han et al., 2010; Dahlmeier et al., 2013) or simulated (Lee and Seneff, 2008; Foster and Andersen, 2009). More recently, a system has been trained on revision statistics from Wikipedia (Cahill et al., 2013). We build on all three paradigms, using standard English cor-

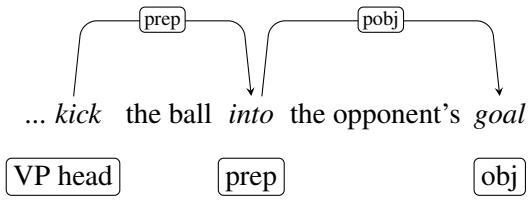

Figure 1: Parse tree for the carrier sentence in Table 1. Distractors are generated on the basis of the prepositional object ("obj") and the NP/VP head to which the prepositional phrase is attached (Section 3).

pora (Section 3.1), error-annotated learner corpora (Section 3.2) and learner revision corpora (Section 3.3) as resources to predict the most plausible distractors.

3 Item generation

The system assumes as input a set of English sentences, which are to serve as candidates for carrier sentences. In each candidate sentence, the system scans for prepositions, and extracts two features from the linguistic context of each preposition:

- The **prepositional object**. In Figure 1, for example, the word "goal" is the prepositional object of the key, "into".
- The head of the noun phrase or verb phrase (**NP/VP head**) to which the prepositional phrase (PP) is attached. In Figure 1, the PP "into the opponent's goal" is attached to the VP head "kick".

The system passes these two features to the following methods to generate distractors.[1] If all three methods are able to return a distractor, the preposition qualifies to serve as the key. If more than one key is found, the system randomly chooses one of them.

In the rest of this paper, we will sometimes abbreviate these three methods as the "Co-occur" (Section 3.1), "Error" (Section 3.2), and "Revision" (Section 3.3) methods, respectively.

3.1 Co-occurrence method

Proposed by Lee and Seneff (2007), this method requires co-occurrence statistics from a large corpus of well-formed English sentences.

[1]We do not consider errors where a preposition should be inserted or deleted.

Co-occurrence method ("Co-occur")
... *kicked* the chair *with* ...
... *kicked* the can *with* ...
... *with* the *goal* of ...
Learner error method ("Error")
... *kicked* it <error>*in*</error> the *goal*.
... *kick* the ball <error>*in*</error> the other team's *goal*.
Learner revision method ("Revision")
... *kick* the ball ~~to~~ into his own *goal*.
... *kick* the ball ~~to~~ towards his own *goal*.

Table 2: The Co-occurrence Method (Section 3.1) generates "with" as the distractor for the carrier sentence in Figure 1; the Learner Error Method (Section 3.2) generates "in"; the Learner Revision Method (Section 3.3) generates "to".

This method first retrieves all prepositions that co-occur with *both* the prepositional object and the NP/VP head in the carrier sentence. These prepositions are removed from consideration as distractors, since they would likely yield a correct sentence. The remaining candidates are those that co-occur with *either* the prepositional object *or* the NP/VP head, but not both. The more frequently the candidate co-occurs with either of these words, the more plausible it is expected to appear to a learner. Thus, the candidate with the highest co-occurrence frequency is chosen as the distractor. As shown in Table 2, this method generates the distractor "with" for the carrier sentence in Figure 1, since many instances of "kick ... with" and "with ... goal" are attested.

3.2 Learner error method

This method requires examples of English sentences from an error-annotated learner corpus. The corpus must mark wrong preposition usage, but does not need to provide corrections for the errors.

This method first retrieves all PPs that have the given prepositional object and are attached to the given NP/VP head. It then computes the frequency of prepositions that head these PPs and are marked as wrong. The one that is most frequently marked as wrong is chosen as the distractor. As shown in Table 2, this method generates the distractor "in" for the carrier sentence in Figure 1, since it is often marked as an error.

3.3 Learner revision method

It is expensive and time consuming to annotate learner errors. As an alternative, we exploit the revision behavior of learners in their English writing. This method requires draft versions of texts written by learners. In order to compute statistics on how often a preposition in an earlier draft ("draft n") is replaced with another one in the later draft ("draft $n + 1$"), the sentences in successive drafts must be sentence- and word-aligned.

This method scans for PPs that have the given prepositional object and are attached to the given NP/VP head. For all learner sentences in draft n that contain these PPs, it consults the sentences in draft $n+1$ to which they are aligned; it retains only those sentences whose prepositional object and the NP/VP head remain unchanged, but whose preposition has been replaced by another one. Among these sentences, the method selects the preposition that is most frequently edited between two drafts. Our assumption is that frequent editing implies a degree of uncertainty on the part of the learner as to which of these prepositions is in fact correct, thus suggesting that they may be effective distractors. As shown in Table 2, this method generates the distractor "to" for the carrier sentence in Figure 1, since it is most often edited in the given linguistic context. This study is the first to exploit a corpus of learner revision history for item generation.[2]

4 Experimental setup

In this section, we first describe our datasets (Section 4.1) and the procedure for item generation (Section 4.2). We then give details on the expert evaluation (Section 4.3) and the learner evaluation (Section 4.4).

4.1 Data

Carrier sentences. We used sentences in the English portion of the Wikicorpus (Reese et al., 2010) as carrier sentences. To avoid selecting stems with overly difficult vocabulary, we ranked the sentences in terms of their most difficult word. We measured the difficulty level of a word firstly with the graded English vocabulary lists compiled by the Hong Kong Education Bureau (EDB, 2012); and secondly, for words not occurring in any of these lists, with frequency counts derived from the Google Web Trillion Word Corpus.[3] In order to retrieve the prepositional object and the NP/VP head (cf. Section 3), we parsed the Wikicorpus, as well as the corpora mentioned below, with the Stanford parser (Manning et al., 2014).

Co-occurrence method ("Co-occur"). The statistics for the Co-occurrence method were also based on the English portion of Wikicorpus.

Learner Revision method ("Revision"). We used an 8-million-word corpus of essay drafts written by Chinese learners of English (Lee et al., 2015). This corpus contains over 4,000 essays, with an average of 2.7 drafts per essay. The sentences and words between successive drafts have been automatically aligned.

Learner Error method ("Error"). In addition to the corpus of essay drafts mentioned above, we used two other error-annotated learner corpora. The NUS Corpus of Learner English (NUCLE) contains one million words of academic writing by students at the National University of Singapore (Dahlmeier et al., 2013). The EF-Cambridge Open Language Database (EFCAMDAT) contains over 70 million words from 1.2 million assignments written by learners from a variety of linguistic background (Geertzen et al., 2013). A subset of the database has been error-annotated. We made use of the writings in this subset that were produced by students from China and Russia.

Human items ("Textbook"). To provide a comparison with human-authored items, we used the practise tests for preposition usage offered in an English exercise book designed for intermediate and advanced learners (Watcyn-Jones and Allsop, 2000). From the 50 tests in a variety of formats, we harvested 56 multiple-choice items, all of which had one key and three distractors.

4.2 Item generation procedure

We gathered three sets of 400 carrier sentences, for use in three evaluation sessions (see Section 4.4). Each sentence in Set 1 has one counterpart in Set 2 and one counterpart in Set 3 that have the same key, NP/VP head and prepositional object. We will refer to the items created from these counterpart carrier sentences as "similar" items. We will use these "similar" items to measure the learning impact on the subjects.

Each item has one key and distractors generated

[2] A similar approach, using revision statistics in Wikipedia, has been used for the purpose of correcting preposition errors (Cahill et al., 2013).

[3] http://norvig.com/ngrams/

by each of the three methods. For about half of the items, the three methods complemented one another to offer three distinct distractors. In the other half, two of the methods yielded the same distractor, resulting in only two distractors for those items. In Set 1, for control purposes, 56 of the items were replaced with the human items.

4.3 Expert evaluation procedure

Two professional English teachers (henceforth, the "experts") examined each of the 400 items in Set 1. They annotated each item, and each choice in the item, as follows.

For each item, the experts labeled its difficulty level in terms of the preposition usage being tested in the carrier sentence. They did not know whether the item was human-authored or machine-generated. Based on their experience in teaching English to native speakers of Chinese, they labeled each item as suitable for those in "Grades 1-3", "Grades 4-6", "Grades 7-9", "Grades 10-12", or ">Grade 12". We mapped these five categories to integers — 2, 5, 8, 11 and 13, respectively — for the purpose of calculating difficulty scores.

For each choice in the item, the experts judged whether it is correct or incorrect. They did not know whether each choice was the key or a distractor. They may judge one, multiple, or none of the choices as correct. For an incorrect choice, they further assessed its plausibility as a distractor, again from their experience in teaching English to native speakers of Chinese. They may label it as "Plausible", "Somewhat plausible", or "Obviously wrong".

4.4 Learner evaluation procedure

Ten university students (henceforth, the "learners") took part in the evaluation. They were all native Chinese speakers who did not major in English. The evaluation consisted of three one-hour sessions held on different days. At each session, the learner attempted 80 items on a browser-based application (Figure 2). The items were distributed in these sessions as follows.

Session 1. The 400 items in Set 1 were divided into 5 groups of 80 items, with 11 to 12 human items in each group. The items in each group had comparable difficulty levels as determined by the experts, with average scores ranging from 7.9 to 8.1. Each group was independently attempted by two learners. The system recorded the items to which the learner gave wrong answers; these will be referred to as the "wrong items". Among the items to which the learner gave correct answers, the system randomly set aside 10 items; these will be referred to as "control items".

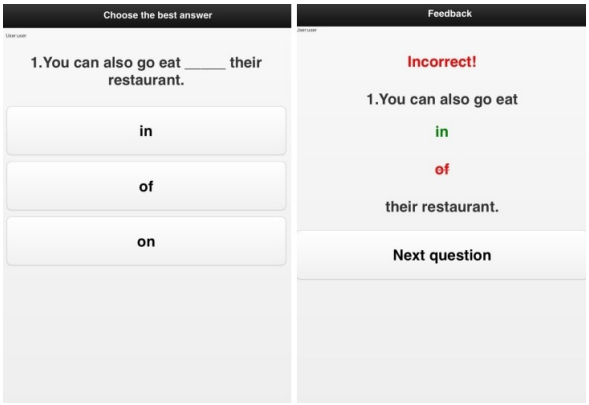

Figure 2: Interface for the learner evaluation. On the left, the learner selects a choice by tapping on it; on the right, the learner receives feedback.

Session 2. To measure the short-term impact, Session 2 was held on the day following Session 1. Each learner attempted 80 items, drawn from Set 2. These items were personalized according to the "wrong items" of the individual learner. For example, if a learner had 15 "wrong items" from Session 1, he or she then received 15 similar items[4] from Set 2. In addition, he or she also received ten items that were similar to the "control items" from Session 1. The remaining items were drawn randomly from Set 2. As in Session 1, the system noted the "wrong items" and set aside ten "control items".

Session 3. To test the long-term effect of these exercises, Session 3 was held two weeks after Session 2. Each learner attempted another 80 items, drawn from Set 3. These 80 items were chosen in the same manner as in Session 2.

5 Results

We first report inter-annotator agreement between the two experts on the difficulty levels of the carrier sentences and the distractors (Section 5.1). We then compare the difficulty levels of the human- and machine-generated items (Section 5.2). Next, we analyze the reliability and difficulty[5] of the

[4] See definition of "similar" in Section 4.2.

[5] Another metric, "validity", measures the ability of the distractor to discriminate between students of different proficiency levels. This metric is relevant for items intended for

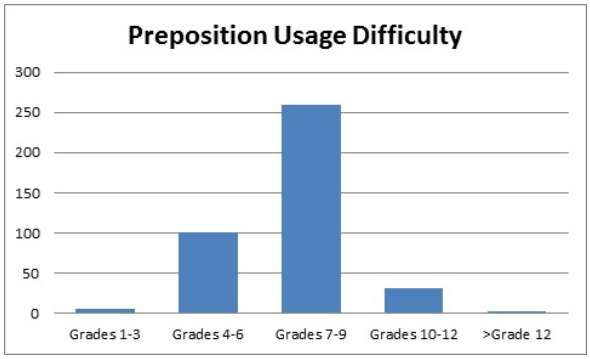

Figure 3: The difficulty level of the items in Set 1, as annotated by the experts.

automatically generated distractors (Sections 5.3 and 5.4), and the role of the native language (Section 5.5). Finally, we measure the impact on the learners' preposition proficiency (Section 5.6).

5.1 Inter-annotator agreement

For estimating the difficulty level of the preposition usage in the carrier sentences, the experts reached "substantial" agreement with kappa at 0.765 (Landis and Koch, 1977). In deciding whether a choice is correct or incorrect, the experts reached "almost perfect" agreement with kappa at 0.977. On the plausibility of the distractors, they reached "moderate" agreement with kappa at 0.537. The main confusion was between the categories "Obviously wrong" and "Somewhat plausible".

On the whole, expert judgment tended to correlate with actual behavior of the learners. For distractors considered "Plausible" by both experts, 63.6% were selected by the learners. In contrast, for those considered "Obviously wrong" by both experts, only 11.8% attracted any learner.

5.2 Carrier sentence difficulty

Figure 3 shows the distribution of difficulty level scores for the preposition usage in carrier sentences. Most items were rated as "Grades 7-9", with "Grades 4-6" being the second largest group.

A common concern over machine-generated items is whether the machine can create or select the kind of carrier sentences that illustrate challenging or advanced preposition usage, compared to those crafted by humans. In our system, the preposition errors and revisions in the learner corpora — as captured by the NP/VP head and the prepositional object — effectively served as the filter for selecting carrier sentences. Some of these errors and revisions may well be careless or trivial mistakes, and may not necessarily lead to the selection of appropriate carrier sentences.

To answer this question, we compared the difficulty levels of preposition usage in the machine-generated and human-crafted items. The average difficulty score for the human items was 8.7, meaning they were suitable for those in Grade 8. The average for the machine-generated items were lower, at 7.2. This result suggests that our system can select carrier sentences that illustrate challenging preposition usage, at a level that is only about 1.5 grade points below those designed by humans.

5.3 Distractor reliability

A second common concern over machine-generated items is whether their distractors might yield correct sentences. When taken out of context, a carrier sentence often admits multiple possible answers (Tetreault and Chodorow, 2008; Lee et al., 2009). In this section, we compare the performance of the automatic distractor generation methods against humans.

A distractor is called "reliable" if it yields an incorrect sentence. The Learner Revision method generated the most reliable distractors[6]; on average, 97.4% of the distractors were judged incorrect by both experts (Table 3). The Co-occurrence method ranked second at 96.1%, slightly better than those from the Learner Error method. Many distractors from the Learner Error method indeed led to incorrect sentences in their original contexts, but became acceptable when their carrier sentences were read in isolation. Items with unreliable distractors were excluded from the learner evaluation.

Surprisingly, both the Learner Revision and Co-occurrence methods outperformed the humans. Distractors in some of the human items did indeed yield sentences that were technically correct, and were therefore deemed "unreliable" by the experts. In many cases, however, these distractors were accompanied with keys that provided more natural choices. These items, therefore, remained valid.

assessment purposes (Brown et al., 2005; Sakaguchi et al., 2013) rather than self-learning.

[6]The difference with the Co-occurrence method is not statistically significant, in part due to the small sample size.

Method	Reliable distractor
Co-occur	96.1%
Error	95.6%
Revision	**97.4%**
Textbook	95.8%

Table 3: Distractors judged reliable by both experts.

5.4 Distractor difficulty

In the context of language learning, an item can be considered more useful if one of its distractors elicits a wrong choice from the learner, who would then receive corrective feedback. In this section, we compare the "difficulty" of the distractor generated by the various methods, in terms of their ability to attract the learners.

Expert evaluation. The two methods based on learner statistics produced the highest-quality distractors (Table 4). The Learner Error method had the highest rate of plausible distractors (51.2%) and the lowest rate of obviously wrong ones (22.0%). In terms of the number of distractors considered "Plausible", this method significantly outperformed the Learner Revision method.[7]

According to Table 4, all three automatic methods outperformed the humans in terms of the number of distractors rated "Plausible". This comparison, however, is not entirely fair, since the human items always supplied three distractors, whereas about half of the machine-generated items supplied only two, when two of the methods returned the same distractor.

An alternate metric is to compute the average number of distractors rated "Plausible" *per item*. On average, the human items had 0.91 plausible distractors; in comparison, the machine-generated items had 1.27. This result suggests that automatic generation of preposition distractors can perform at the human level.

Learner evaluation. The most direct way to evaluate the difficulty of a distractor is to measure how often a learner chose it. The contrast is less clear cut in this evaluation. Overall, the learners correctly answered 76.2% of the machine-generated items, and 75.5% of the human items, suggesting that the human distractors were more challenging. One must also take into account, however, the fact that the carrier sentences are more difficult in the human items than in the machine-generated ones. Broadly speaking, the machine-generated distractors were almost as successful as those authored by humans.

[7] $p < 0.05$ by McNemar's test, for both expert annotators.

Method	Plausible	Somewhat plausible	Obviously wrong
Co-occur	34.6%	31.5%	33.9%
Error	**51.2%**	**26.8%**	**22.0%**
Revision	45.4%	28.5%	26.1%
Textbook	31.4%	34.2%	34.5%

Table 4: Plausibility judgment of distractors by experts.

Consistent with the experts' opinion (Table 4), the Learner Error method was most successful among the three automatic methods (Table 5). The learner selection rate of its distractors was 13.5%, which was significantly higher[8] than its closest competitor, the Learner Revision method, at 9.5%. The Co-occurrence method ranked last, at 9.2%. It is unfortunately difficult to directly compare these rates with that of the human distractors, which they were offered in different carrier sentences.

5.5 Impact of L1

We now turn our attention to the relation between the native language (L1) of the user, and that of the learner corpora used for training the system. Specifically, we wish to measure the gain, if any, in matching the L1 of the user with the L1 of the learner corpora. To this end, for the Learner Error method, we generated distractors from the EF-Cambridge corpus with two sets of statistics: one harvested from the portion of the corpus with writings by Chinese students, the others from the portion by Russian students.

Expert evaluation. Table 6 contrasts the experts' plausibility judgment on distractors generated from these two sets. Chinese distractors were

[8] $p < 0.05$ by McNemar's test.

Method	Learner selection rate
Co-occur	9.2%
Error	**13.5%**
Revision	9.5%

Table 5: Percentage of distractors selected by learners.

Method	Plausible	Somewhat plausible	Obviously wrong
Chinese	**57.7%**	**24.0%**	**18.3%**
Russian	55.3%	22.0%	22.7%

Table 6: Plausibility judgment of distractors generated from the Chinese and Russian portions of the EF-Cambridge corpus, by experts.

Difficulty level	Retention rate
Below 6	74.0%
6-8	71.3%
9-11	60.0%
12 or above	25%

Table 7: Retention rate for items at different levels of difficulty.

slightly more likely to be rated "plausible" than the Russian ones, and less likely to be rated "obviously wrong".[9] The gap between the two sets of distractors was smaller than may be expected.

Learner evaluation. The difference was somewhat more pronounced in terms of the learners' behavior. The learners selected Chinese distractors, which matched their L1, 29.9% of the time over the three sessions. In contrast, they fell for the Russian distractors, which did not match their L1, only 25.1% of the time. This result confirms the intuition that matching L1 improves the plausibility of the distractors, but the difference was nonetheless relatively small. This result suggets that it might be worth paying the price for mismatched L1s, in return for a much larger pool of learner statistics.

5.6 Impact on learners

In this section, we consider the impact of these exercises on the learners. The performance of the learners was rather stable across all sessions; their average scores in the three sessions were 73.0%, 73.6% and 69.9%, respectively. It is difficult, however, to judge from these scores whether the learners benefited from the exercises, since the composition of the items differed for each session.

Instead, we measured how often the learners retain the system feedback. More specifically, if the learner chose a distractor and received feedback (cf. Figure 2), how likely would he or she succeed in choosing the key in a "similar"[10] item in a subsequent session.

We compared the learners' responses between Sessions 1 and 2 to measure the short-term impact, and between Sessions 2 and 3 to measure the long-term impact. In Session 2, when the learners attempted items that were "similar" to their "wrong items" from Session 1, they succeeded in choosing the key in 72.4% of the cases.[11] We refer to this figure as the "retention rate", in this case over the one-day period between the two sessions. The retention rate deteriorated over a longer term. In Session 3, when the learners attempted items that were "similar" to their "wrong items" from Session 2, which took place two weeks before, they succeeded only in 61.5% of the cases.[12]

Further, we analyzed whether the difficulty level of the items affected their retention rate. Statistics in Table 7 show that the rate varied widely according to the difficulty level of the "wrong items". Difficult items, at Grade 12 or beyond, proved hardest to learn, with a retention rate of only 25%. At the other end of the spectrum, those below Grade 6 were retained 74% of the time. This points to the need for the system to reinforce difficult items more frequently.

6 Conclusions

We have presented a computer-assisted language learning (CALL) system that automatically creates fill-in-the-blank items for prepositions. We found that the preposition usage tested in automatically selected carrier sentences were only slightly less challenging than those crafted by humans. We compared the performance of three methods for distractor generation, including a novel method that exploits learner revision statistics. The method based on learner error statistics yielded the most plausible distractors, followed by the one based on learner revision statistics. The items produced jointly by these automatic methods, in both expert and learner evaluations, rivalled the quality of human-authored items. Further, we evaluated the extent to which mismatched

[9] Data sparseness prevented us from generating both Chinese and Russian distractors for the same carrier sentences for evaluation. These statistics are therefore not controlled with regard to the difficulty level of the sentences.

[10] See definition of "similar" in Section 4.2.

[11] As a control, the retention rate for correctly answered items in Session 1 was 80% in Session 2.

[12] As a control, the retention rate for correctly answered items in Session 2 was 69.0% in Session 3.

native language (L1) affects distractor plausibility. Finally, in a study on the short- and long-term impact on the learners, we showed that difficult items had lower retention rate. In future work, we plan to conduct larger-scale evaluations to further validate these results, and to apply these methods on other common learner errors.

Acknowledgments

We thank NetDragon Websoft Holding Limited for their assistance with system evaluation, and the reviewers for their very helpful comments. This work was partially supported by an Applied Research Grant (Project no. 9667115) from City University of Hong Kong.

References

Jonathan C. Brown, Gwen A. Frishkoff, and Maxine Eskenazi. 2005. Automatic Question Generation for Vocabulary Assessment. In *Proc. HLT-EMNLP*.

Aoife Cahill, Nitin Madnani, Joel Tetreault, and Diane Napolitano. 2013. Robust Systems for Preposition Error Correction using Wikipedia Revisions. In *Proc. NAACL-HLT*.

Chia-Yin Chen, Hsien-Chin Liou, and Jason S. Chang. 2006. FAST: An Automatic Generation System for Grammar Tests. In *Proc. COLING/ACL Interactive Presentation Sessions*.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English. In *Proc. 8th Workshop on Innovative Use of NLP for Building Educational Applications*.

EDB. 2012. *Enhancing English Vocabulary Learning and Teaching at Secondary Level*. http://www.edb.gov.hk/vocab_learning_sec.

Rachele De Felice and Stephen Pulman. 2009. Automatic Detection of Preposition Errors in Learner Writing. *CALICO Journal*, 26(3):512–528.

Jennifer Foster and Øistein E. Andersen. 2009. GenERRate: Generating Errors for Use in Grammatical Error Detection. In *Proc. 4th Workshop on Innovative Use of NLP for Building Educational Applications*.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proc. 31st Second Language Research Forum (SLRF)*.

Na-Rae Han, Joel Tetreault, Soo-Hwa Lee, and Jin-Young Ha. 2010. Using Error-annotated ESL Data to Develop an ESL Error Correction System. In *Proc. LREC*.

Emi Izumi, Kiyotaka Uchimoto, Toyomi Saiga, Thepchai Supnithi, and Hitoshi Isahara. 2003. Automatic Error Detection in the Japanese Learners' English Spoken Data. In *Proc. ACL*.

Nikiforos Karamanis, Le An Ha, and Ruslan Mitkov. 2006. Generating Multiple-Choice Test Items from Medical Text: A Pilot Study. In *Proc. 4th International Natural Language Generation Conference*.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33:159–174.

John Lee and Stephanie Seneff. 2007. Automatic generation of cloze items for prepositions. In *Proc. Interspeech*.

John Lee and Stephanie Seneff. 2008. Correcting Misuse of Verb Forms. In *Proc. ACL*.

John Lee, Joel Tetreault, and Martin Chodorow. 2009. Human Evaluation of Article and Noun Number Usage: Influences of Context and Construction Variability. In *Proc. Linguistic Annotation Workshop*.

John Lee, Chak Yan Yeung, Amir Zeldes, Marc Reznicek, Anke Lüdeling, and Jonathan Webster. 2015. CityU Corpus of Essay Drafts of English Language Learners: a Corpus of Textual Revision in Second Language Writing. *Language Resources and Evaluation*, 49(3):659–683.

Chao-Lin Liu, Chun-Hung Wang, Zhao-Ming Gao, and Shang-Ming Huang. 2005. Applications of Lexical Information for Algorithmically Composing Multiple-Choice Cloze Items. In *Proc. 2nd Workshop on Building Educational Applications Using NLP*, pages 1–8.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proc. ACL System Demonstrations*, pages 55–60.

Samuel Reese, Gemma Boleda, Montse Cuadros, Lluís Padró, and German Rigau. 2010. Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus. In *Proc. LREC*.

Keisuke Sakaguchi, Yuki Arase, and Mamoru Komachi. 2013. Discriminative Approach to Fill-in-the-Blank Quiz Generation for Language Learners. In *Proc. ACL*.

Chi-Chiang Shei. 2001. FollowYou!: An Automatic Language Lesson Generation System. *Computer Assisted Language Learning*, 14(2):129–144.

Simon Smith, P. V. S. Avinesh, and Adam Kilgarriff. 2010. Gap-fill Tests for Language Learners: Corpus-Driven Item Generation. In *Proc. 8th International Conference on Natural Language Processing (ICON)*.

Eiichiro Sumita, Fumiaki Sugaya, and Seiichi Yamamoto. 2005. Measuring Non-native Speakers Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions. In *Proc. 2nd Workshop on Building Educational Applications using NLP*.

Joel Tetreault and Martin Chodorow. 2008. The Ups and Downs of Preposition Error Detection in ESL Writing. In *Proc. COLING*.

Peter Watcyn-Jones and Jake Allsop. 2000. *Test Your Prepositions*. Penguin Books Ltd.

Torsten Zesch and Oren Melamud. 2014. Automatic Generation of Challenging Distractors Using Context-Sensitive Inference Rules. In *Proc. Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.