ReCALL (2020), 32: 2, 145–161
doi:10.1017/S0958344019000193

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Crowdsourcing evaluation of the quality of automatically generated questions for supporting computer-assisted language teaching

Maria Chinkina
Universität Tübingen, Germany (maria.chinkina@sfs.uni-tuebingen.de)

Simón Ruiz
Universität Tübingen, Germany (simon.ruiz-hernandez@sfs.uni-tuebingen.de)

Detmar Meurers
Universität Tübingen, Germany (detmar.meurers@sfs.uni-tuebingen.de)

Abstract

How can state-of-the-art computational linguistic technology reduce the workload and increase the efficiency of language teachers? To address this question, we combine insights from research in second language acquisition and computational linguistics to automatically generate text-based questions to a given text. The questions are designed to draw the learner's attention to target linguistic forms – phrasal verbs, in this particular case – by requiring them to use the forms or their paraphrases in the answer. Such questions help learners create form-meaning connections and are well suited for both practice and testing. We discuss the generation of a novel type of question combining a *wh-* question with a gapped sentence, and report the results of two crowdsourcing evaluation studies investigating how well automatically generated questions compare to those written by a language teacher. The first study compares our system output to gold standard human-written questions via crowdsourcing rating. An equivalence test shows that automatically generated questions are comparable to human-written ones. The second crowdsourcing study investigates two types of questions (*wh-* questions with and without a gapped sentence), their perceived quality, and the responses they elicit. Finally, we discuss the challenges and limitations of creating and evaluating question-generation systems for language learners.

Keywords: automatic question generation; crowdsourcing; intelligent computer-assisted language learning

1. Introduction

Questions are habitually used by teachers to test comprehension, encourage discussion, and check understanding of learning materials. We argue that in a language-learning classroom particular questions can facilitate the acquisition and practice of different linguistic forms by creating a functional need to notice and process a linguistic form (Robinson, Mackey, Gass & Schmidt, 2012). This idea is supported by a large body of research on input enhancement (Sharwood Smith, 1993; see Simard, 2018, for a recent overview) and processing instruction, particularly research on structured input activities (see VanPatten, 2017, for a recent review).

Cite this article: Chinkina, M., Ruiz, S. & Meurers, D. (2020). Crowdsourcing evaluation of the quality of automatically generated questions for supporting computer-assisted language teaching. *ReCALL* 32(2): 145–161. https://doi.org/10.1017/S0958344019000193

In our work, we combine insights from research in second language acquisition and computational linguistics to automatically generate text-based questions that draw the learner's attention to target linguistic forms by requiring learners to process and use these forms in their answers. In this study we focus on English phrasal verbs. Phrasal verbs are multi-word verbs that function syntactically and semantically as a single unit (e.g. *end up* [finish]). They are the first linguistic form we examine as they represent a considerable teaching and learning load (Garnier & Schmitt, 2016). Phrasal verbs exhibit both lexical and syntactic properties that make them particularly difficult for language learners to master (Larsen-Freeman & Celce-Murcia, 2015). This includes the specifics of their compositionality and the seemingly random nature of some of the particles that are part of phrasal verbs (Side, 1990).

With this in mind, and considering the intersection of second language acquisition, natural language processing (NLP) methods, and computer-assisted language learning (CALL) (see Lu, 2018; Meurers, 2012; Meurers & Dickinson, 2017; Reinders & Stockwell, 2017), we compare questions targeting phrasal verbs written by an English teacher to those automatically generated by an intelligent CALL (ICALL) application in order to assess whether the questions produced by computers are equivalent in terms of quality to those written by humans. As we will show, automatically generated questions are qualitatively comparable to those devised by a language teacher, and therefore we argue that this technology can be integrated into future language instruction via computer-assisted language *teaching* applications.

2. Questions in traditional and computer-assisted language teaching

Form-focused instruction is premised on the idea that mere exposure to input is insufficient for second language acquisition to occur (Long, 1991). Learners need to notice certain features of the language (e.g. grammatical encodings, lexical items) in the input in order for these features to be acquired. In line with Ellis's (2016) remarks about focus on form, it has also been argued that different kinds of attention-drawing activities are needed to facilitate the acquisition and practice of different linguistic forms (Robinson *et al.*, 2012).

Questions offer the possibility to provide form-focused instruction because they can target specific parts in reading materials that contain language constructions that learners need to systematically pay attention to or *notice*, thereby producing a functional demand to process second language input (Ellis, 2016). In the case of phrasal verbs, the linguistic target in this study, questions targeting these structures can be designed to draw the learner's attention to form by focusing on both form and meaning, whereby the only way to answer a given question correctly is by understanding both the lexical and morphological form and the meaning of the targeted phrasal verb (see Appendix for examples).

2.1 From manually written to automatically generated questions in CALL

While input enhancement and language-learning activities are traditionally implemented manually or at times hard-coded in CALL tools, computational linguistic methods can support their automation resulting in ICALL applications (Meurers *et al.*, 2010; Ziegler *et al.*, 2017). By leveraging computational linguistic tools and methods, we have developed a system that automatically generates *wh-* questions and gapped sentences from text, with the primary goal of drawing learners' attention to target linguistic forms. For instance, given the source text (1), a program we developed automatically generated the question (1a) targeting the phrasal verb *tick up*:

(1) *Source text*: [. . .] Cancellations "ticked up slightly and unexpectedly" in early April amid press coverage about the coming increases, the Netflix letter said.

a. *Computer*: According to the Netflix letter, what did cancellations do? Cancellations _____ slightly and unexpectedly in early April amid press coverage about the coming increases.

Our system relies on Stanford CoreNLP, a natural language processing toolkit by Manning, Surdeanu, Bauer, Finkel, Bethard and McClosky (2014). In general terms, the task of NLP is to assign a structure representing syntactic relationships between words in a given sentence. More specifically, we use it for sentence splitting, tokenizing, lemmatizing, constituency and dependency parsing, and to resolve coreferences. Given an analysed sentence, our algorithm generates questions from it as follows. It detects the target linguistic form (e.g. phrasal verbs), identifies grammatical functions in the sentence (e.g. subject, predicate), and turns a declarative sentence into an interrogative one by applying syntactic transformation rules. A sentence with a gap is generated by substituting all parts of a target linguistic construction (e.g. the verb and the particle of a phrasal verb) with a gap. The technical side of the implementation of our system is described in more detail in Chinkina and Meurers (2017). Here, we focus on evaluating the approach and extend it with a novel question type in question-generation research, the combination of a *wh-* question and a gapped sentence.

2.2 Computational linguistic methods for evaluating automatically generated questions

The computational linguistic task of automatic question generation has explored a range of question types, from factual recall questions (Wolfe, 1976) to deeper discussion questions (Adamson, Bhartiya, Gujral, Kedia, Singh & Rosé, 2013). The work at the intersection of computational linguistics and language learning has addressed the generation of *wh-* questions (Heilman, 2011; Mitkov & Ha, 2003) as well as that of cloze exercises (i.e. sentences where the target form is replaced with a gap) (Becker, Basu & Vanderwende, 2012; Brown, Frishkoff & Eskenazi 2005; Mostow *et al.*, 2004). In order to leverage the advances in question generation and apply them in the language-learning context in a focused task, we propose to generate questions consisting of both a *wh-* question and a sentence with a gap. In the following sections, we discuss its advantages over simple open-ended *wh-* questions and compare the two question types in an online study.

As for the performance of question-generation systems, it has been assessed either by using automatic measures, such as BLEU (Papineni, Roukos, Ward & Zhu, 2002), or by collecting human judgments. For instance, Zhang and VanLehn (2016) recruited students to judge the comparability of computer-generated, web-crawled and human-written biology questions based on several 5-point scales (relevance, fluency, ambiguity, pedagogy, depth). Heilman and Smith (2010) conducted a crowdsourcing study to assess the goodness of computer-generated questions using one 5-point scale and used the collected judgments to train a statistical ranker for their question-generation system. Crowdsourcing is an attractive option for evaluating question-generation systems given its time and cost effectiveness along with the similarity of the crowd ratings to expert judgments (Benoit, Conway, Lauderdale, Laver & Mikhaylov, 2016; Snow, O'Connor, Jurafsky & Ng, 2008). Using crowdsourcing to compare computer-generated and human-written questions seems like a logical next step in this line of research.

3. Research questions and hypotheses

The purpose of our study was to compare the perceived quality of automatically generated questions to that of human-written ones. Zhang and VanLehn (2016) conducted a similar kind of evaluation in an offline setting. The researchers showed that university students' ratings to questions generated by a computer to test comprehension of biology texts are comparable to those written by a teacher. Heilman and Smith (2010), on the other hand, turned to crowdsourcing for assessing the quality of their automatically generated factual questions using one "goodness" scale, but did not compare it to the perceived quality of human-written questions. Informed by this research, we opted for a crowdsourcing evaluation and defined two important aspects as the basis for a comparison between automatically generated and manually written questions: well-formedness and answerability. A question is considered well formed if it does not contain

grammar mistakes. The answerability of a question, on the other hand, refers to its semantics. A question is considered answerable if it is formulated in a way that is understandable and an answer to it can be found in the source text. We target these characteristics with our first research question:

> RQ1. Are computer-generated questions comparable to those written by English teachers in well-formedness and answerability?

Although there is no previous research comparing computer-generated and human-written questions via crowdsourcing, based on the aforementioned related work by Zhang and VanLehn (2016) and Heilman and Smith (2010), we expected that the questions produced by the computer and the English teacher would be comparable regarding the two aspects under investigation.

As the combination of a *wh-* question and a gapped sentence that we generate is novel in the question-generation research field (see Heilman, 2011), we were particularly interested in whether this type of question is perceived as better formed and more easily answerable than standard *wh-* questions. Therefore, we formulated the second research question:

> RQ2. Are *wh-* questions followed by a gapped sentence perceived as better with respect to well-formedness and answerability than open-ended *wh-* questions?

For this novel type of question, we predicted that a *wh-* question and a gapped sentence may cancel out each other's potential disadvantages, and thus their combination would be rated higher than a single *wh-* question with respect to both perceived well-formedness and answerability.

Finally, to further explore the differences between the two types of questions (with and without a gapped sentence) in terms of what answers they can elicit, we formulated the third question:

> RQ3. Do *wh-* questions followed by a gapped sentence elicit more correct responses and target phrasal verbs than open-ended *wh-* questions?

We predicted that the addition of a gapped sentence would limit the participants' choice of an answer phrase to the phrasal verb given in the text. Thus, the combination of a *wh-* question and a gapped sentence would increase the likelihood of obtaining a correct response and have a higher probability of containing the target phrasal verbs from the source text as part of the answer than simple open-ended *wh-* questions.

To address these research questions, we conducted two crowdsourcing studies on the Figure Eight platform (https://www.figure-eight.com), discussed in detail in the following sections.

4. Study 1: Quality of automatically generated questions

For questions to be effective in a real-life language classroom, they must be reasonably well formed and answerable. The goal of the first study was to evaluate our system by comparing the quality of computer-generated questions to the gold standard questions written by the English teacher in these two respects.

4.1 Data for Study 1

The data consisted of 138 questions designed to facilitate the acquisition and practice of phrasal verbs. Stanford CoreNLP (Manning *et al.*, 2014) and additional algorithms were used to automatically detect 92 sentences containing unique phrasal verbs in a corpus of 40 English news articles. For these sentences, our question-generation system produced 69 questions, both well and ill

formed, all of which were included in the data set. An English teacher wrote 113 questions targeting the same sentences, so we randomly selected 69 of those to include in the data set. To illustrate, the questions that follow are instances of well-formed questions by a human (2a) and a computer (2b), derived from the same source text. Question (3a) is a well-formed (+) human-written question, whereas the computer-generated question (3b) is ill formed (–).

(2) *Source text*: [. . .] Beijing's drive to make the nation a leader in robotics through its "Made in China 2025" initiative launched last year has set off a rush as municipalities up and down the country vie to become China's robotics center.

a. *Human* (+): What has the "Made in China 2025" initiative done since it was launched last year? It has _____ a rush for municipalities to become China's robotics center.

b. *Computer* (+): According to the article, what has Beijing's drive done? Beijing's drive has _____ a rush as municipalities up and down the country vie to become China's robotics center.

(3) *Source text*: [. . .] Twitter is also working to better define its role in the social media landscape. This week it rolled out a video ad that showed it as the place to go for live news, updates and discussion about current events.

a. *Human* (+): What is Twitter doing to better define its role in the social media landscape? It _____ a video ad this week.

b. *Computer* (–): According to the article, what did this week do? This week _____ a video ad that showed it as the place to go for live news.

4.2 Participants of Study 1

Although the main advantage of crowdsourcing is that it provides access to a large number of people all around the world, it comes with a risk of recruiting unsuitable contributors (see Stewart, Chandler & Paolacci, 2017, for a recent review on the use of crowdsourcing in behavioral research). For this study, we needed judgments that are as close to expert ones (e.g. English teachers) as possible. The following steps helped us achieve this.

First, we used the functionality of the crowdsourcing website to select only English-speaking countries, thereby increasing the probability of the contributors being native speakers of English. However, when we only received one response in the first five hours, we extended the list to include some other European countries where English proficiency is high, which, according to the EF English Proficiency Index (Education First, 2017), are the Netherlands, Denmark, Norway, Sweden, Finland, Germany, and Austria.

We included test questions to further filter out unsuitable contributors. In order to proceed to the main task, each contributor first took a quiz in which they had to correctly rate and answer four out of five test questions. The test questions looked exactly like the questions from the main task, except that some of them were manually edited to either be ungrammatical or unanswerable in order to ensure an even distribution of low-rated and high-rated test questions, as recommended by Figure Eight guidelines. Finally, a small number of test questions looked different and required the participants to specify whether they were in fact proficient speakers of English and whether their answers were reliable. In this way, we made sure that the contributors understood the task at hand, that they were able to distinguish between a well-formed and an ill-formed question, and that their language skills were advanced enough to answer a question given a source text.

In order to perform the main task, participants had to keep their accuracy rate above 70% by correctly answering randomly inserted test questions among the other question items. In total, 364 reliable contributors took part in this study.

4.3 Procedure of Study 1

Participants were presented with a source text (an excerpt from a news article, one to three sentences long) and a question about this text. Each question had to be rated on two separate 5-point Likert scales: one for well-formedness and the other for answerability. To help ensure participants were paying attention, participants were also required to answer the question about the source text. Finally, they were asked to guess whether the presented question was written by either the English teacher or generated by the computer. We collected 10 judgments per question item.

4.4 Results of Study 1

To investigate whether computer-generated questions were rated as high as human-written ones, we first calculated the intra-class correlation (ICC) between the contributors' ratings. The ICC was smaller than .1 (i.e. .08 and .09 for well-formedness and answerability, respectively), meaning that the contributors provided different ratings for different question items, so that we can assume the judgments to be independent.

To test whether the quality of the questions generated by the computer was equivalent to those written by the teacher, we conducted Schuirmann's (1987) two one-sided tests of equivalence (medium effect size $d = 0.5$, alpha level of .05) for each of the two scales. All results were statistically significant on both scales: well-formedness, $t_1(912) = 9.814$, $p_1 < .001$, $t_2(912) = -5.677$, $p_2 < .001$, 90% CI [0.025, 0.220]; answerability, $t_1(944) = 7.322$, $p_1 \leq .001$, $t_2(944) = -8.170$, $p_2 < .001$, 90% CI [−0.134, 0.079]. As the null and the alternative hypothesis are reversed in equivalence testing, statistically significant results indicate that the two samples are indeed equivalent. Thus, the results show that questions generated by the computer are not inferior or superior to those written by the English teacher in well-formedness or answerability, considering medium size effects.

To investigate for differences of smaller effect sizes, we used *t*-tests to compare the questions produced by the computer and the human. The results showed that there is a statistically significant difference between human-written and computer-generated questions with respect to their well-formedness with a small effect size, $t(1{,}316) = 2.48$, $p = .013$, $d = 0.133$.[1] On the answerability scale, there was no such significant difference, $t(1{,}362) = -0.509$, $p = .611$, $d = 0.027$.

Finally, we analysed the contributors' guesses about whether the questions were written by the English teacher or generated by the computer using a mixed-effects model. There was a strong correlation between rating a question high and thinking that it was written by the English teacher on the well-formedness scale, $t(1{,}299) = 17.12$, $p < .001$, $d = 0.806$, and the answerability scale, $t(1{,}307) = 11.71$, $p < .001$, $d = 0.610$. In fact, the top 11% of computer-generated questions (i.e. those having scored the highest on well-formedness) were thought to be written by the English teacher. Overall, participants thought that 74% of human-written and 67% of computer-generated questions were produced by a teacher.

4.5 Discussion of the results of Study 1

The results of the first study imply that the questions automatically generated by our system are comparable to those written by a human with respect to well-formedness and answerability,

[1]The exact numbers differ slightly from those in Chinkina, Ruiz and Meurers (2017), as we excluded two unreliable responses from the original data analysis. However, this did not lead to different levels of statistical significance.

although the questions written by the English teacher were rated as slightly better formed. Interestingly, most of the well-formed and answerable questions were thought to be written by the English teacher, even if they had in fact been generated automatically. This indicates that computers are not expected to be able to produce high-quality output in the sense that automatically generated questions are expected to be more ungrammatical and unnatural.

5. Study 2: Types of questions and the answers they elicit

In the second crowdsourcing study, we wanted to find out (a) whether the addition of a gapped sentence to an otherwise open-ended *wh-* question influences a question rating and (b) whether *wh-* questions followed by a gapped sentence elicit more phrasal verbs than open-ended *wh-* questions. The task and the procedure were the same as in the first study, but the selection criteria for both data and participants differed.

5.1 Data for Study 2

For each source sentence, we generated two types of questions, namely an open-ended *wh-* question and the same *wh-* question followed by a gapped sentence. As we did not intend to evaluate our system in this study, we excluded all ungrammatical and unanswerable computer-generated questions. Given the source text used in Example (2), the following questions were part of the data set in our second study:

(4) *Source text*: [. . .] Beijing's drive to make the nation a leader in robotics through its "Made in China 2025" initiative launched last year has set off a rush as municipalities up and down the country vie to become China's robotics center.

a. *Human*: What has the "Made in China 2025" initiative done since it was launched last year?

b. *Human*: What has the "Made in China 2025" initiative done since it was launched last year? It has _____ a rush for municipalities to become China's robotics center.

c. *Computer*: According to the article, what has Beijing's drive done?

d. *Computer*: According to the article, what has Beijing's drive done? Beijing's drive has _____ a rush as municipalities up and down the country vie to become China's robotics center.

Overall, the data consisted of 96 human-written and 96 computer-generated questions. They were randomized in such a way that the two types of questions (with and without a gapped sentence) for the same source sentence were never shown together on the same page. We collected five judgments per question item.

5.2 Participants in Study 2

For the second study, we selected contributors with a high reliability, as specified in their profile on the crowdsourcing page, but did not limit the participation based on their level of English. To ensure the contributors' suitability, we included a quiz of five test questions, four of which had to be answered and rated correctly in order to proceed to the main task. By assuming that users working on an English-language crowdsourcing website have enough of a language background for this second study, we aimed to mimic a study with English learners of different levels of

proficiency. In this study, we collected judgments from 545 contributors including 68 participants who had already taken part in the first study. However, for the evaluation purposes, we only analysed the data from the 477 new contributors.

5.3 Procedure of Study 2

As in the first study, participants were asked to answer the presented questions and rate them on the two separate 5-point Likert scales in terms of well-formedness and answerability. For this second study, we analysed both the ratings and the responses to the questions. Different from the first study, we did not ask participants to guess whether a question was written by a teacher or generated by a computer.

5.4 Results of Study 2

As participants in the second study were not selected based on their English proficiency level, there was less agreement among subjects when rating questions regarding well-formedness and answerability (ICC = 0.34 and 0.37, respectively). Hence, we used mixed-effect models to account for the dependencies across observations.

The analysis was conducted using the *lme4* package Version 1.1-12 in the *R* environment Version 3.2.1 (R Core Team, 2013). We estimated a model for each of the two continuous dependent variables: the perceived well-formedness and answerability of question items. The models included fixed effects for the source of a question item (human or computer) and the item type (with or without a gapped sentence), as well as crossed random effects for both participants and items (Baayen, 2008). An effect was considered significant if the absolute value of the *t* statistic was greater than or equal to 2.0 (Baayen, 2008; Gelman & Hill, 2006).

First, we found that participants did not rate computer-generated questions significantly lower than human-written questions – well-formedness, $b = 0.024$, $SE = 0.047$, $t = 0.500$; answerability, $b = 0.065$, $SE = 0.060$, $t = 1.080$ – which is in line with the results of our first study with proficient English speakers. As for the addition of a gapped sentence, it did indeed influence the rating of a question item. The results showed that this had an effect on both the perceived well-formedness, $b = 0.158$, $SE = 0.054$, $t = 2.930$, and answerability, $b = 0.127$, $SE = 0.055$, $t = 2.300$. In other words, the addition of a gapped sentence to a simple open-ended *wh-* question improved the perceived well-formedness and answerability of the question.

Finally, we conducted logistic regression analyses (Jaeger, 2008) to investigate which type of questions elicited more correct responses and more phrasal verbs. In the first model, the dependent variable was analysed as a binary outcome: correct versus incorrect. In the second model, the dependent variable was also treated as a binary outcome: presence versus absence of the phrasal verb from the source text. We selected a random sample of 20% of responses and excluded nonsensical (e.g. "*good!*") and non-English (e.g. "*konuşma*") answers from the data. Out of 359 answers, 277 (77.2%) contained exact matches of the phrasal verbs given in the source text. Only 12 (3.3%) contained rephrasings of phrasal verbs, and the remaining 70 (19.5%) answers were marked as incorrect. As expected, the linear regression results showed that, as compared to simple *wh-* questions, questions followed by a gapped sentence had a higher probability of eliciting correct responses, $b = 0.791$, $SE = 0.278$, $p = .004$, as well as of containing the target phrasal verbs from the source text, $b = 2.577$, $SE = 0.484$, $p < .001$.

5.5 Discussion of the results of Study 2

The results of the second study show question ratings in line with those from the proficient English speakers in the first study: computer-generated and human-written questions were rated similarly for both well-formedness and for answerability. This confirms our hypothesis that automatically generated questions are perceived as qualitatively comparable to those written

by humans, in line with findings from previous studies on automatic question generation (e.g. Zhang & VanLehn, 2016).

Going beyond the results of the first study, the second study shows that *wh-* questions followed by a gapped sentence are rated higher than open-ended ones on both well-formedness and answerability scales. Apparently, a gapped sentence providing an answer context for a question can render an otherwise ambiguous question more specific so that it is perceived as better formed and easier to answer. The *wh-* questions followed by a gapped sentence also elicited more correct responses and more phrasal verbs. Therefore, our intuition that such gapped answer sentence can be used to narrow down the reader's focus to the target linguistic form in the source sentence is confirmed.

6. Implications of our work for computer-assisted language teaching

The results of the two studies indicate that participants rated the questions produced by the computer and the English teacher similarly, confirming our initial hypothesis that computer-generated questions are comparable to those produced by humans with respect to well-formedness and answerability. A potential implication of this finding is the possibility that language teachers can use our question-generation system to automatically generate questions from reading materials, which in turn may save them time and effort when preparing their class materials. This becomes particularly relevant when considering individual differences between students, which are particularly substantial in second language learning, so that teachers in principle should offer different reading material to individual or subgroups of students. Question generation here fits naturally with ICALL tools supporting the automated retrieval of reading material in line with the individual learner's zone of proximal development (Chen & Meurers, 2019) and the school curriculum (Chinkina & Meurers, 2016).

As we predicted for our second and third research questions, the form of a question item has proven to be an important factor in judging the quality of a question and eliciting correct responses and target phrasal verbs. The combination of a *wh-* question and a gapped sentence was rated higher in terms of perceived well-formedness and answerability than single open-ended *wh-* questions. The combination of a *wh-* question and a gapped sentence provides a more explicit context for answering a question. From the technical point of view, the generation of short *wh-* questions and verbatim gapped sentences is less prone to errors than that of their longer counterparts. The more specific – and therefore long – a *wh-* question is, the more syntactic elements it contains, thus raising the probability of a question being ungrammatical. At the same time, when the number of syntactic elements is kept to a minimum, there is a risk that a question will be too general or ambiguous. On the other hand, gapped sentences are typically grammatical and unambiguous (Becker *et al.*, 2012), but they do not serve a communicative goal. Therefore, combining a general *wh-* question with a more specific gapped sentence can help avoid the aforementioned pitfalls of the two question types: It maximizes the grammaticality and minimizes the ambiguity of the whole question item while keeping the task communicative.

7. Linguistic and technical limitations, challenges, and considerations

Importantly, the perceived similarity between computer-generated and human-written questions does not only provide evidence for the generally good quality of questions that can be generated but also reveals the limitations of both approaches. In particular, in addition to occasional grammar mistakes, both automatically and manually produced questions can be too vague, overly specific, or include superfluous information. This is illustrated by the following examples:

(5) *Source text*: [. . .] "It is a mirror of how sensitive the issue is and that people don't want to talk about it," Hage told the Thomson Reuters Foundation. "The number one reason they are

not speaking up is because of the social stigma and the victims are afraid to be blamed, so there is a deafening silence around the issue," she added.

a. *Computer* (wrong grammar): According to the article, what are people not doing? People are not _____ *is because* of the social stigma and the victims are afraid to be blamed.

(6) *Source text*: [. . .] Bowing out on Wednesday, Cameron said: "Nothing is really impossible if you put your mind to it. After all, as I once said: 'I was the future once.'"

a. *Human* (too vague): What did Cameron do on Wednesday? He _____.

(7) *Source text*: [. . .] The oldest of the field of candidates, he has just taken up a position at Yale University although a source familiar with his plans indicated he was reluctant to take on the post.

a. *Computer* (superfluous information in the question): According to the article, what has the oldest of the field of candidates done? The oldest of the field of candidates has _____ a position at Yale University *although a source familiar with his plans indicated he was reluctant to take on the post*.

Although leveraging and fine-tuning of computational linguistic tools can help improve the quality of automatically generated questions, there are linguistic and technical considerations that need to be taken into account when creating questions and evaluating their quality. We discuss them in detail in this section.

7.1 Inclusion of non-restrictive phrases and clauses

The computer-generated questions that received the highest scores in our studies were concise, which showcases the importance of considering the syntactic structure of a sentence. For instance, removing non-restrictive clauses (usually separated by commas or other punctuation), but keeping restrictive types, usually led to well-formed questions, such as the one that follows that received the highest score on both the well-formedness and answerability scales:

(8) *Source text*: [. . .] Meanwhile, LeEco has spun out sports and cloud units, bringing in private equity capital from conglomerate HNA Group, Alibaba boss Jack Ma's Yunfeng Capital, and others.

a. *Computer*: According to the article, what has LeEco done? LeEco has _____ sports and cloud units.

Interestingly, this seemed to be the case even when not enough information was provided in the question in order to answer it correctly. For example, the following question does not specify the conditions under which Jia might be forced to put up more collateral. Nevertheless, the question also received the highest scores on both scales:

(9) *Source text*: [. . .] Such share pledges can be risky: if Leshi Internet stock fell sharply, Jia might be forced to put up more collateral or sell down his stake.

a. *Computer*: According to the article, what might Jia be forced to do? Jia might be forced to _____ his stake.

We only removed non-restrictive clauses from a gapped sentence when they were separated by commas, which was the case for 33% of computer-generated questions. We never removed prepositional phrases when they were in the same clause with the target form. Subordinate and coordinate clauses were not removed when they followed the main clause and were not separated by a comma, as in example (8).

To obtain some quantitative evidence regarding our intuition about the superiority of the questions with removed non-restrictive clauses, we conducted the following pilot analyses. First, we filtered out obviously ungrammatical computer-generated questions, where the errors were caused by the parser or the coreference resolution module. We then annotated the remaining 60 computer-generated questions ($M_{\text{well-formedness}} = 4.59$, $M_{\text{answerability}} = 4.62$) and conducted Welch's *t*-tests. The results showed that the perceived well-formedness of the computer-generated questions with removed non-restrictive clauses ($M = 4.73$, $SD = 0.23$) was higher than that of the ones where no part of the sentence following the target form was removed ($M = 4.52$, $SD = 0.49$), and the difference was significant, $t(58) = 2.30$, $p = .02$, 95% CI [4.73, 4.52]. For answerability, on the other hand, the questions with removed clauses ($M = 4.59$, $SD = 0.79$) were rated as more difficult to answer than the ones that did not undergo the modification ($M = 4.64$, $SD = 0.54$). However, the difference was non-significant, $t(28) = -0.23$, $p = .82$, 95% CI [0.45, 0.36]. The results confirm that the removal of non-restrictive clauses in general leads to better-formed questions, but more data would be relevant to explore when they remain easy to answer.

Although the heuristics of splitting the sentence into clauses separated by commas seems to be working well, excluding conditional clauses may lead to unanswerable questions, especially in a richer context. This leads us to the next subsection, where we discuss the limitations of the task of automatic question generation and its evaluation.

7.2 Limitations of natural language processing tools and algorithms

The quality of automatically generated questions relies on the accuracy of the natural language processing tools that our question-generation system is built on. In fact, the main causes of ill-formed questions were erroneous coreference resolution (43%) and incorrect parses (28%) of the source sentences, with ill-formedness being operationalized as an average rating below 3 on a 5-point scale. Other factors influencing question quality include:

i. The question item may not present enough information to answer it correctly (e.g. a missing restrictive clause) or be too specific compared to a more general context of the paragraph:

(10) *Source text*: [. . .] Chinese retailers have also cut staff and seen inventories pile up, luxury sector growth has dried up, and fast-food giants such as KFC-parent Yum Brands Inc and McDonald's Corp are grappling for growth.

a. *Computer* (–): According to the article, what do inventories do? Inventories _____.

ii. The question item may have superfluous information (e.g. a non-restrictive phrase or a clause) making it too long and potentially unnatural:

(11) *Source text*: [. . .] Musk on Wednesday sketched out his vision for an integrated carbon-free energy enterprise offering products and services beyond electric cars and batteries.

a. *Computer* (–): According to the article, what did Musk on Wednesday do? Musk on Wednesday _____ his vision for an integrated carbon-free energy enterprise offering products and services beyond electric cars and batteries.

iii. According to feedback from the participants of the studies, questions may be perceived as less well formed if the subject in the gapped sentence repeats the subject in the *wh-* question. Although the question item as a whole could sound more natural if the subject in the gapped sentence were substituted with a pronoun, it poses a computational challenge because of the aforementioned suboptimal performance of coreference resolution tools. Given the alternative of generating a wrong pronoun (e.g. *he* instead of *she*), we opted for the safe, albeit slightly less natural option of keeping the subject in both the *wh-* question and the gapped sentence. As a result, all computer-generated examples in this paper demonstrate this limitation.

7.3 Evaluation of question-generation systems

First and foremost, it should be noted that any kind of human evaluation is subjective. In our studies, this issue became particularly salient when raters encountered the test questions in the first crowdsourcing experiment. The Figure Eight guidelines recommend an even distribution of answers for test questions (i.e. testing both good and bad questions in our case). Although ill-formed or unanswerable test questions were not difficult to write and did not receive criticism from the participants, even by those who rated these questions incorrectly (we accepted any rating below 4 for such questions), the rating of good test questions proved to be more challenging and subjective. As an alternative, one could test participants only on ill-formed questions, possibly also giving only a binary choice (*Is this question grammatical or ungrammatical?*) instead of the 5-point scale (*How well formed is this question?*).

Malicious activities (e.g. randomly clicking through the task, copy-pasting answers) are another limitation of a crowdsourcing experiment design – or any web-based design, for that matter (Gadiraj, Demartini, Kawase & Dietze, 2015). In our second study, where the quality control mechanism was not as strict as in the first, participants used the exact wording from the source text in 97% of their answers. When designing a similar study in the future, one could block the copy-paste functionality in order to prevent participants from directly copying answers from the text.

8. Conclusion and outlook

To conclude, answering questions is an integral part of facilitating and practicing vocabulary and grammar in a language-learning classroom. In the two studies presented in this paper, we found evidence that automatically generated and human-written questions can be comparable with respect to both well-formedness and answerability. The findings are in line with previous research involving expert judges evaluating the quality of computer-generated and human-written questions (e.g. Zhang & VanLehn, 2016) – although our discussion also identified clear room for improvement in question generation.

We found that the addition of a gapped sentence to a *wh-* question significantly improves its well-formedness and answerability. Moreover, the responses elicited by *wh-* questions followed by a gapped sentence contain significantly more correct answers and phrasal verbs than those elicited by open-ended *wh-* questions. From the computational linguistic perspective, these findings imply that question-generation systems can benefit from leveraging and combining different types of questions.

Although we focused on phrasal verbs as the target linguistic form in this study, our system is able to generate questions to any verb phrase, and in principle any automatically identifiable dependent. In future studies, we plan to assess the quality of computer-generated questions targeting different linguistic forms appearing in texts of different genres to empirically test question-generation effectiveness. For this purpose, a large-scale randomized controlled field study with intermediate English language learners is currently being planned as part of a grant proposal. The study is designed to provide an evidence-based assessment of the effectiveness of question-generation technology in a real-life educational setting and compare it to more traditional approaches.

Interestingly, proficient speakers of English thought that most of the questions were written by an English teacher, although the proportion of computer-generated and human-written questions in the study was the same. This finding shows that people are often unaware of the state of the art in computational linguistics and how it can or could connect to the needs of real-life teaching and learning. We believe that computer-assisted language *teaching*, that is, the use of technology by not only language learners but also primarily by language teachers, can play an important role in supporting teachers in facing current challenges. Automated approaches arguably will become particularly important for the class-internal differentiation that is increasingly required to adaptively support different subgroups of learners, for which automatically generated materials are ideally suited.

Acknowledgements. This research was supported by the LEAD Graduate School & Research Network [GSC1028], a project of the Excellence Initiative of the German federal and state governments. Maria Chinkina and Simón Ruiz were doctoral students in the LEAD Graduate School & Research Network at the time of the submission of this manuscript. We would like to thank our LEAD colleagues Michael Grosz and Johann Jacoby for sharing their expertise and insights in the field of statistical analysis. Special thanks to the three anonymous reviewers, whose comments helped us improve the paper.

Ethical statement. All respondents participated in the study voluntarily and provided informed consent prior to commencement of the study. The confidentiality and anonymity of the research respondents was maintained throughout the study and the consequent analysis.

References

Adamson, D., Bhartiya, D., Gujral, B., Kedia, R., Singh, A. & Rosé, C. P. (2013) Automatically generating discussion questions. In Lane, H. C., Yacef, K., Mostow, J. & Pavlik, P. (eds.), *Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 2013*. Heidelberg: Springer-Verlag Berlin Heidelberg, 81–90. https://doi.org/10.1007/978-3-642-39112-5_9

Baayen, R. H. (2008) *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511801686

Becker, L., Basu, S. & Vanderwende, L. (2012) Mind the gap: learning to choose gaps for question generation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Stroudsburg: Association for Computational Linguistics, 742–751.

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M. & Mikhaylov, S. (2016) Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, 110(2): 278–295. https://doi.org/10.1017/s0003055416000058

Brown, J. C., Frishkoff, G. A. & Eskenazi, M. (2005) Automatic question generation for vocabulary assessment. In *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing: Proceedings of the Conference*. Stroudsburg: Association for Computational Linguistics, 819–826. https://doi.org/10.3115/1220575.1220678

Chen, X. & Meurers, D. (2019) Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer Assisted Language Learning*, 32(4): 418–447. https://doi.org/10.1080/09588221.2018.1527358

Chinkina, M. & Meurers, D. (2016) Linguistically aware information retrieval: Providing input enrichment for second language learners. In Tetreault, J., Burstein, J., Leacock, C. & Yannakoudakis, H. (eds.), *The 11th Workshop on Innovative Use of NLP for Building Educational Applications: Proceedings of the Workshop*. Stroudsburg: Association for Computational Linguistics, 188–198. https://doi.org/10.18653/v1/w16-0521

Chinkina, M. & Meurers, D. (2017) Question generation for language learning: From ensuring texts are read to supporting learning. In Tetreault, J., Burstein, J., Kochmar, E., Leacock, C. & Yannakoudakis, H. (eds.), *The 12th Workshop on Innovative Use of NLP for Building Educational Applications: Proceedings of the Workshop*. Stroudsburg: Association for Computational Linguistics, 334–344. https://doi.org/10.18653/v1/w17-5038

Chinkina, M., Ruiz, S. & Meurers, D. (2017) Automatically generating questions to support the acquisition of particle verbs: Evaluating via crowdsourcing. In Borthwick, K., Bradley, L. & Thouësny, S. (eds.), *CALL in a climate of change: Adapting to turbulent global conditions – short papers from EUROCALL 2017*. Voillans: Research-publishing.net, 73–78. https://doi.org/10.14705/rpnet.2017.eurocall2017.692

Education First. (2017) EF English Proficiency Index. https://www.ef.com/epi/

Ellis, R. (2016) Focus on form: A critical review. *Language Teaching Research*, 20(3): 405–428. https://doi.org/10.1177/1362168816628627

Gadiraju, U., Demartini, G., Kawase, R. & Dietze, S. (2015) Human beyond the machine: Challenges and opportunities of microtask crowdsourcing. *IEEE Intelligent Systems*, 30(4): 81–85. https://doi.org/10.1109/mis.2015.66

Garnier, M. & Schmitt, N. (2016) *Picking up* polysemous phrasal verbs: How many do learners know and what facilitates this knowledge? *System*, 59: 29–44. https://doi.org/10.1016/j.system.2016.04.004

Gelman, A. & Hill, J. (2006) *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511790942

Heilman, M. (2011) *Automatic factual question generation from text*. Carnegie Mellon University, PhD.

Heilman, M. & Smith, N. A. (2010) Rating computer-generated questions with Mechanical Turk. In Callison-Burch, C. & Dredze, M. (eds.), *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Stroudsburg: Association for Computational Linguistics, 35–40.

Jaeger, T. F. (2008) Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4): 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Larsen-Freeman, D. & Celce-Murcia, M. (2015) *The grammar book: Form, meaning, and use for English language teachers* (3rd ed.). Boston: National Geographic Learning/Heinle Cengage Learning.

Long, M. H. (1991) Focus on form: A design feature in language teaching methodology. *Foreign Language Research in Cross-Cultural Perspective*, 2(1): 39–52. https://doi.org/10.1075/sibil.2.07lon

Lu, X. (2018) Natural language processing and intelligent computer-assisted language learning (ICALL). In Liontas, J. I. & DelliCarpini, M. (eds.), *The TESOL encyclopedia of English language teaching*. Hoboken: Wiley-Blackwell. https://doi.org/10.1002/9781118784235.eelt0422

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J. & McClosky, D. (2014) The Stanford CoreNLP natural language processing toolkit. In Bontcheva, K. & Zhu, J. (eds.), *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Stroudsburg: Association for Computational Linguistics, 55–60. https://doi.org/10.3115/v1/p14-5010

Meurers, D. (2012) Natural language processing and language learning. In Chapelle, C. A. (ed.), *Encyclopedia of applied linguistics*. Hoboken: John Wiley.

Meurers, D. & Dickinson, M. (2017) Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning*, 67(S1): 66–95. https://doi.org/10.1111/lang.12233

Meurers, D., Ziai, R., Amaral, L., Boyd, A., Dimitrov, A., Metcalf, V. & Ott, N. (2010) Enhancing authentic web pages for language learners. In Tetreault, J., Burstein, J. & Leacock, C. (eds.), *Fifth Workshop on Innovative Use of NLP for Building Educational Applications: Proceedings of the Workshop*. Stroudsburg: Association for Computational Linguistics, 10–18.

Mitkov, R. & Ha, L. A. (2003) Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing – Volume 2*. Stroudsburg: Association for Computational Linguistics, 17–22.

Mostow, J., Beck, J., Bey, J., Cuneo, A., Sison, J., Tobin, B. & Valeri, J. (2004) Using automated questions to assess reading comprehension, vocabulary, and effects of tutorial interventions. *Technology, Instruction, Cognition, and Learning*, 2: 103–140.

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2002) BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 311–318*. https://doi.org/10.3115/1073083.1073135

R Core Team (2013) *R: A language and environment for statistical computing*. Vienna: The R Foundation for Statistical Computing. http://www.R-project.org/

Reinders, H. & Stockwell, G. (2017) Computer-assisted SLA. In Loewen, S. & Sato, M. (eds.), *The Routledge handbook of instructed second language acquisition*. New York: Routledge, 361–375.

Robinson, P., Mackey, A., Gass, S. M. & Schmidt, R. (2012) Attention and awareness in second language acquisition. In Gass, S. M. & Mackey, A. (eds.), *The Routledge handbook of second language acquisition*. New York: Routledge, 247–267.

Schuirmann, D. J. (1987) A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6): 657–680. https://doi.org/10.1007/bf01068419

Sharwood Smith, M. (1993) Input enhancement in instructed SLA: Theoretical bases. *Studies in Second Language Acquisition*, 15(2): 165–179. https://doi.org/10.1017/s0272263100011943

Side, R. (1990) Phrasal verbs: Sorting them out. *ELT Journal*, 44(2): 144–152. https://doi.org/10.1093/elt/44.2.144

Simard, D. (2018) Input enhancement. In Liontas, J. I. & DelliCarpini, M. (eds.), *The TESOL encyclopedia of English language teaching*. Hoboken: Wiley-Blackwell. https://doi.org/10.1002/9781118784235.eelt0072

Snow, R., O'Connor, B., Jurafsky, D. & Ng, A. Y. (2008) Cheap and fast — but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: Association for Computational Linguistics, 254–263. https://doi.org/10.3115/1613715.1613751

Stewart, N., Chandler, J. & Paolacci, G. (2017) Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences*, 21(10): 736–748. https://doi.org/10.1016/j.tics.2017.06.007

VanPatten, B. (2017) Processing instruction. In Loewen, S. & Sato, M. (eds.), *The Routledge handbook of instructed second language acquisition*. New York: Routledge, 166–180.

Wolfe, J. H. (1976) Automatic question generation from text: An aid to independent study. *ACM SIGCSE Bulletin*, 8(1): 104–112.

Zhang, L. & VanLehn, K. (2016) How do machine-generated questions compare to human-generated questions? *Research and Practice in Technology Enhanced Learning*, 11(1): 1–28. https://doi.org/10.1186/s41039-016-0031-7

Ziegler, N., Meurers, D., Rebuschat, P., Ruiz, S., Moreno-Vega, J. L., Chinkina, M., Li, W. & Grey, S. (2017) Interdisciplinary research at the intersection of CALL, NLP, and SLA: Methodological implications from an input enhancement project. *Language Learning*, 67(S1): 209–231. https://doi.org/10.1111/lang.12227

Appendix

The appendix illustrates the ratings of well-formedness and answerability for the computer-generated and human-written questions discussed in the paper. Five-point scales were used for both variables, with 1 being the lowest and 5 the highest score:

1. How well formed is this question? (1 – very ill formed … 5 – very well formed)
2. Can this question be answered by the source text? (1 – no, not at all … 5 – yes, easily)

Source text, URL	Question item	Well formed	Answerable
	Computer generated		
[. . .] Cancellations "ticked up slightly and unexpectedly" in early April amid press coverage about the coming increases, the Netflix letter said. http://www.reuters.com/article/us-netflix-results-idUSKCN0ZY2H4	According to the Netflix letter, what did cancellations do? Cancellations _____ slightly and unexpectedly in early April amid press coverage about the coming increases.	5	5
[. . .] Beijing's drive to make the nation a leader in robotics through its "Made in China 2025" initiative launched last year has set off a rush as municipalities up and down the country vie to become China's robotics center. http://www.reuters.com/article/us-china-debt-robotics-insight-idUSKCN10E0EV	According to the article, what has Beijing's drive done? Beijing's drive has _____ a rush as municipalities up and down the country vie to become China's robotics center.	4.9	4.9
[. . .] Twitter is also working to better define its role in the social media landscape. This week it rolled out a video ad that showed it as the place to go for live news, updates and discussion about current events. http://www.reuters.com/article/us-twitter-results-idUSKCN1062JW	According to the article, what did this week do? This week _____ a video ad that showed it as the place to go for live news.	1.9	3.3
[. . .] "It is a mirror of how sensitive the issue is and that people don't want to talk about it," Hage told the Thomson Reuters Foundation. "The number one reason they are not speaking up is because of the social stigma and the victims are afraid to be blamed, so there is a deafening silence around the issue," she added. http://news.trust.org/item/20160725143529-udc98/	According to the article, what are people not doing? People are not _____ is because of the social stigma and the victims are afraid to be blamed.	3.3	3.9
[. . .] The oldest of the field of candidates, he has just taken up a position at Yale University although a source familiar with his plans indicated he was reluctant to take on the post. http://in.reuters.com/article/india-rbi-candidates-idINKCN10615K	According to the article, what has the oldest of the field of candidates done? The oldest of the field of candidates has _____ a position at Yale University although a source familiar with his plans indicated he was reluctant to take on the post.	4.8	4.8

[. . .] Meanwhile, LeEco has spun out sports and cloud units, bringing in private equity capital from conglomerate HNA Group, Alibaba boss Jack Ma's Yunfeng Capital, and others. http://blogs.reuters.com/breakingviews/2016/08/16/chinas-netflix-to-tesla-is-spread-painfully-thin/	According to the article, what has LeEco done? LeEco has _____ sports and cloud units.	5	5
[. . .] Such share pledges can be risky: if Leshi Internet stock fell sharply, Jia might be forced to put up more collateral or sell down his stake. http://blogs.reuters.com/breakingviews/2016/08/16/chinas-netflix-to-tesla-is-spread-painfully-thin/	According to the article, what might Jia be forced to do? Jia might be forced to _____ his stake.	5	5
[. . .] Chinese retailers have also cut staff and seen inventories pile up, luxury sector growth has dried up, and fast-food giants such as KFC-parent Yum Brands Inc and McDonald's Corp are grappling for growth. http://www.reuters.com/article/us-china-box-office-idUSKCN10L2PR	According to the article, what do inventories do? Inventories _____.	4.3	4.3
[. . .] Musk on Wednesday sketched out his vision for an integrated carbon-free energy enterprise offering products and services beyond electric cars and batteries. http://www.reuters.com/article/us-tesla-masterplan-research-idUSKCN1011H5	According to the article, what did Musk on Wednesday do? Musk on Wednesday _____ his vision for an integrated carbon-free energy enterprise offering products and services beyond electric cars and batteries.	3.6	4.7
	Human written		
[. . .] Beijing's drive to make the nation a leader in robotics through its "Made in China 2025" initiative launched last year has set off a rush as municipalities up and down the country vie to become China's robotics center. http://www.reuters.com/article/us-china-debt-robotics-insight-idUSKCN10E0EV	What has the "Made in China 2025" initiative done since it was launched last year? It has _____ a rush for municipalities to become China's robotics center.	5	4.6
[. . .] Twitter is also working to better define its role in the social media landscape. This week it rolled out a video ad that showed it as the place to go for live news, updates and discussion about current events. http://www.reuters.com/article/us-twitter-results-idUSKCN1062JW	What is Twitter doing to better define its role in the social media landscape? It _____ a video ad this week.	4.8	5
[. . .] Bowing out on Wednesday, Cameron said: "Nothing is really impossible if you put your mind to it. After all, as I once said: 'I was the future once.'" http://www.reuters.com/article/us-britain-eu-cameron-pmqs-idUSKCN0ZT1MF	What did Cameron do on Wednesday? He _____.	4.7	4.5

About the authors

Maria Chinkina is a doctoral candidate at the LEAD Graduate School & Research Network and the University of Tübingen, Germany. In her thesis, she explores and implements computational linguistic techniques, such as information retrieval, input enrichment, and question generation, that help language learners to create a richer grammatical intake from the given text

input. Her research focus lies at the intersection of computational linguistics, second language acquisition and computer-assisted language learning.

Simón Ruiz is a post-doctoral researcher at the English department of the University of Tübingen, Germany, from where he also obtained his PhD. His research focuses on individual differences in second language acquisition, second language teaching and learning, implicit and explicit learning in second language acquisition, and intelligent computer-assisted language learning.

Detmar Meurers is professor of computational linguistics at the University of Tübingen, Germany, and on the steering board of the LEAD Graduate School & Research Network in empirical educational science there. As head of the ICALL-Research.com group, his work focuses on intelligent computer-assisted language learning and computational linguistic methods in second language acquisition research and language teaching. He has published on automatic short-answer assessment, the analysis of learner corpora, linguistic complexity analysis, tutoring systems, and input enrichment and enhancement applications.

Author ORCiD. Maria Chinkina, https://orcid.org/0000-0002-6566-5332
Author ORCiD. Simón Ruiz, https://orcid.org/0000-0003-2501-0486
Author ORCiD. Detmar Meurers, https://orcid.org/0000-0002-9740-7442