Question Generation for Language Learning: From ensuring texts are read to supporting learning

Maria Chinkina **Detmar Meurers**
LEAD Graduate School & Research Network
Department of Linguistics
Eberhard Karls Universität Tübingen
{maria.chinkina,detmar.meurers}@uni-tuebingen.de

Abstract

In Foreign Language Teaching and Learning (FLTL), questions are systematically used to assess the learner's understanding of a text. Computational linguistic (CL) approaches have been developed to generate such questions automatically given a text (e.g., Heilman, 2011). In this paper, we want to broaden the perspective on the different functions questions can play in FLTL and discuss how automatic question generation can support the different uses.

Complementing the focus on meaning and comprehension, we want to highlight the fact that questions can also be used to make learners notice form aspects of the linguistic system and their interpretation. Automatically generating questions that target linguistic forms and grammatical categories in a text in essence supports incidental focus-on-form (Loewen, 2005) in a meaning-focused reading task. We discuss two types of questions serving this purpose, how they can be generated automatically; and we report on a crowdsourcing evaluation comparing automatically generated to manually written questions targeting particle verbs, a challenging linguistic form for learners of English.

1 Introduction

"Learning is goal-oriented ... Teaching therefore becomes an active thinking and decision-making process in which the teacher is constantly assessing what students already know, what they need to know, and how to provide for successful learning." (O'Malley and Chamot, 1990)

One of the most common ways to find out what students do and do not know is to ask questions. In communicative and task-based language teaching, where the meaning and function of language drives the pedagogy, questions are asked to support the task at hand. Relatedly, when dealing with written language material, recall or comprehension questions can spell out typical goals for reading a text: searching for specific information or more comprehensively integrating the information provided in the text into the reader's background knowledge to draw inferences on that basis.

An increasing body of CL research supports the automatic generation of questions in order to assist teachers in constructing practice exercises and tests. For example, Heilman (2011) is a prominent approach for the generation of factual, low-level questions suitable for beginner or intermediate students. His goal is to assess the reader's knowledge of the information in the text, which is relevant for both content and language teaching.

At the same time, Second Language Acquisition (SLA) research since the 90s has emphasized that language input and meaning-based tasks alone are not sufficient to ensure successful language acquisition. Learners must also notice linguistic forms and grammatical categories (Schmidt, 1990) and teaching can facilitate such noticing through so-called focus on form (Doughty and Williams, 1998). Focus on form is designed to draw the learner's attention to relevant linguistic features of the language as they arise, while keeping the overriding focus on meaning (Long, 1991, pp. 45f). For written language, input enhancement (Sharwood Smith, 1993) has been proposed to make relevant forms more salient in the input, e.g., by coloring or font choices. Such visual input enhancement has also been automated using CL methods (Meurers et al., 2010), as part of a system also generating in-text exercises.

One problem with *form-based visual input enhancement* is that coloring a form or otherwise

making it visually more salient neither ensures that it is noticed and cognitively processed more thoroughly nor do we know which aspect of that form the reader will notice and how it is interpreted. For example, coloring the form *has been raining* in a text may draw the reader's attention to any aspect of those forms (e.g., number or length of the words, or the *-ing* suffix of the last word), and noticing the form does not necessarily map it to its present perfect continuous interpretation.

In this paper, we propose another option for providing input enhancement, *functionally-driven input enhancement*. Concretely, we propose to generate two types of questions creating a functional need to process the targeted linguistic features. The first type of questions we generate are content questions about the clause containing the targeted form. So these questions are like Heilman's factual questions, but they are targeting sentences containing particular linguistic features to be acquired. To answer such questions, the learner must process form and meaning of the clause, ensuring increased activation of the targeted form. The goal of these questions is to ensure more exposure to the forms, so we will refer to them as *form exposure questions*.

The second type of functionally-driven input enhancement is designed to also ensure interpretation of the targeted form. For this, the nature of the question that is generated must be changed from asking about the content of the text to asking about the interpretation of the form being targeted. In the spirit of the concept questions of Workman (2008), we will refer to such questions as *grammar concept questions*.

The goal of this paper is to combine insights from SLA research with CL techniques to explore new options for question generation in support of language learning. In section 2, we first characterize the overall spectrum of questions we consider to be of relevance to FLTL, from supporting communication via ensuring texts are read to supporting learning of linguistic forms and their function. Section 3.1 then surveys the computational linguistic work on automatic question generation, which has focused on the content-side of the spectrum. Section 3.2 spells out the SLA background needed to motivate our research on question generation targeting linguistic forms and their interpretation. In section 4 we then present the question generation approach we developed, mostly concentrating on the two new types of questions designed to provide functionally-driven input enhancement. For such questions to be effective, they must be reasonably well-formed and answerable, so in section 5 we present the results from a crowd sourcing experiment we conducted to evaluate whether the automatically generated form exposure questions are comparable to manually written questions in those two respects. Finally, section 6 provides a conclusion and outlook.

2 A spectrum of questions for FLTL

In an FLTL context, questions can be asked to serve a broad range of different goals:

1. We can ask about the learner's experience or general knowledge (e.g., "What do you know about Japan?"), which can serve a communicative goal.
2. Comprehension or recall questions can be asked to check whether the learner has understood a text or read it at all.
3. Questions can also be asked with the goal of eliciting a linguistic form from the learner (e.g., the question "What would you do if you won in a lottery?" requires the learner to produce conditionals.)
4. As introduced in the previous section, we can use questions to provide functionally-driven input enhancement drawing the learner's attention to the linguistic forms used in a given text. *Form exposure questions* ensure that the sentence containing the targeted forms are read and generally understood. Answering *grammar-concept questions* in addition requires an understanding of the interpretation of the targeted form.
5. Finally, there also are meta-linguistic questions checking the learner's explicit knowledge of the language system (e.g., "From which verb is the noun *decision* derived?" or "What is the synonym of *staff*?").

The aforementioned goals are presented in a particular order, from more communicative to more formal ones. In the work presented in this paper, we primarily focus on the idea of functionally-driven input enhancement captured by the fourth type: questions drawing the learner's attention to particular linguistic forms in the reading material and their interpretation. To contex-

tualize our approach, we first provide some background on automatic question generation and the SLA concepts grounding our proposal.

3 Background

3.1 Automatic Question Generation

A typical text-based Question Generation (QG) system consists of three components: target selection (sentences and words), generation of questions (and answers), and the generation of distractors, which is applicable for a multiple choice answer setup. Most of work on *target selection* follows a top-down perspective on the text: First, a set of suitable sentences is selected based on different criteria (e.g., Pino et al., 2008; Pilán et al., 2013). Then the target words or linguistic forms are selected within the set of suitable sentences (e.g., Becker et al., 2012). Given our focus on input enhancement for language learning, we instead pursue a bottom-up approach: Given one or more target linguistic forms (e.g., the passive voice, or the present perfect tense), we automatically select all the candidate sentences in a text containing the target forms, apply basic constraints to filter out unsuitable sentences (such as those containing unresolvable pronouns), and then generate questions to the remaining ones.

Once the target sentence has been selected, it can be used to *generate questions* targeting particular linguistic forms contained in the sentence. Heilman (2011) discusses the generation of factual, low-level questions suitable for beginner or intermediate students and gives a comprehensive overview of QG methods. Among the most prominent ones are: replacing the target form with a gap (Agarwal et al., 2011; Becker et al., 2012), applying transformation rules (Mitkov et al., 2006), filling templates (Curto et al., 2012), and generating all possible questions to a sentence and ranking them afterwards using a supervised learning algorithm (Heilman and Smith, 2009). Finally, QG is not an exception to the wave of neural networks, and Du et al. (2017) have recently approached automatic generation of reading comprehension questions on that basis. All of the mentioned QG systems either assess vocabulary or target reading comprehension, which contrasts with the focus of our work on functionally supporting focus on form in language learning.

Distractor generation is a separate complex task that has received some attention in the QG community. It supports the provision of answers in a multiple-choice setup, and the choice of distractors is closely tied to what is intended to be assessed by the question. Traditionally, distractors are selected among words that are semantically related to the correct answer (Mitkov et al., 2006; Araki et al., 2016). Brown et al. (2005) select the distractors among the most frequent words that have the same part of speech as the correct answer. Pino and Eskenazi (2009) inform the distractor generation component by the wrong answers provided by the users of their system. Given that we do not focus on the multiple choice answer format here, distractor generation is not discussed further in this paper.

3.2 Relevant SLA concepts

Attention, *input*, and *form-meaning mapping* are key SLA concepts that are directly related to our work. We already saw in our introduction in section 1 that both meaning and form play important roles in SLA. Pushing this discussion one step further, work in the Input Processing paradigm (VanPatten and Cadierno, 1993), based on Krashen's (1977) input hypothesis, provides several relevant studies showing that "learners process input for meaning before they process it for form" (VanPatten, 1990; Wong, 2001). However, Norris and Ortega (2000) argued that simultaneously directing the learner's attention to form and meaning in the input does not hinder L2 development or reading comprehension. Leow et al. (2008) came to the same conclusion after revisiting the methodology used in the replication studies mentioned above and conducting a new study. Their results did not show any statistically significant differences in comprehension between different intervention groups. Finally, a study by Morgan-Short et al. (2012) demonstrated that learners who attended to and processed linguistic forms while reading for meaning scored higher on comprehension than those only reading for meaning.

In line with the Noticing Hypothesis (Schmidt, 1990), the most straightforward way to draw the learner's attention to particular linguistic forms in a text is to increase their salience. As the meta-analysis of Lee and Huang (2008) shows, results on the isolated effect of visual input enhancement on L2 development has been mixed. One option for pushing this research further is to investigate other types of input enhancement and the combi-

nation of visual input enhancement with other input activities.

The Input Processing approach to SLA has given rise to a pedagogical intervention called *processing instruction* (VanPatten, 2004). Its goal is to ensure that learners make form-meaning connections during reading. This goes beyond textual enhancement, which only ensures noticing (Benati, 2016). One of the components of processing instruction, structured input practice, has been identified as particularly effective in fostering L2 development (VanPatten and Oikkenon, 1996; Benati, 2004; Wong, 2004). Structured input is "input that is manipulated in particular ways to push learners to become dependent on form and structure to get meaning" (Lee and VanPatten, 1995).

Structured input activities can be seen as an umbrella term for a wide range of language teaching techniques. They provide the learners with enriched input and prompt them to process and eventually produce the target linguistic forms. While in the original approach, the input enrichment and development of structured input activities is done manually, CL methods can support this process. We have developed a system for automatic input enrichment, FLAIR (Chinkina and Meurers, 2016), which supports retrieval of documents containing targeted linguistic features. The linguistic features covered by the system include the full set of grammatical constructions spelled out in the official English language curriculum of schools in Baden-Württemberg (Germany). On this enriched input basis, automating the generation of questions as structured input activities is the logical next step. In the next section, we spell out the different types of questions that we are able to generate automatically and discuss the algorithms and challenges behind their generation.

4 Generating questions for FLTL

As mentioned in section 3.1, most of the work on QG has dealt with vocabulary (Brown et al., 2005) and comprehension questions (Mostow et al., 2004), not on linguistic form and grammar. For approaches automatically generating exercises that facilitate grammar acquisition and practice, cloze sentences are the most ubiquitous type. They are generated by substituting the target linguistic form with a gap, and the challenge usually lies in the selection of good sentences and gaps (Becker et al., 2012; Niraula and Rus, 2015).

(1) The advisory group had _____ a list of all the different territorial arrangements in the EU. (draw up)

Metalinguistic questions, which are designed to test learners' explicit knowledge of the language system, have not received much attention in the QG community. The reason probably lies in the fact that they require the use of a limited number of templates and only a minimal amount of NLP. Their frequent use by teachers is also widely criticized by educators and researchers alike, mainly because they do not serve a communicative goal. For example, in order to generate question (2), one would only need a POS-tagger and the WordNet database (Miller, 1995).

(2) From which verb is the noun *generation* derived?

To cover the whole spectrum of exercises facilitating the acquisition and practice of grammar, we also generate cloze and metalinguistic questions. However, the focus of the paper is on questions providing functionally-driven input enhancement, so we limit the discussion to those two types for space reasons.

4.1 Form Exposure Questions

Form exposure questions focus on a particular linguistic form, which can either be part of the question or be expected in the learner's production. They can take the form of a *wh*-, yes/no, or an alternative question. For example, when asking a question about the source text (3), one can think of different linguistic targets: relative clauses, past forms of irregular and regular verbs, etc. Question (3a) is asked about the subject and targets the particle verb *brought in*.

(3) Indeed, Semel and the media executives he brought in by all accounts turned a scrappy young internet startup into a highly profitable company that brought old-line advertising to a new medium.

   a. Who turned a scrappy young internet startup into a highly profitable company? Semel and the media executives he _____.

Generation We generate form exposure questions to subjects, objects, and predicates. The main linguistic form we focus on is the grammatical tense, so our form exposure questions target verbs and verb phrases.

We use the Java implementation of Stanford CoreNLP 3.7.0 for part-of-speech tagging, parsing, and resolving coreferences (Manning et al., 2014). After extracting a sentence or a clause containing the target form, we perform the following steps: adjust and normalize the auxiliaries, resolve pronouns and other referential expressions, and detect quotation sources, if any. Then the algorithm proceeds to detect specific syntactic components of the sentence, to modify them if necessary, and finally *transformation rules* are used to turn a sentence into a question. Let us inspect the algorithm for generating questions to predicates.

A. Active
(e.g., *What have Chinese retailers done?*)

1) Insert the question word "What" at the beginning of the sentence.
2) Identify or generate an auxiliary verb.
   - If there is an auxiliary verb modifying the main verb, identify it.
   - Otherwise, identify the grammatical tense of the main verb and generate an appropriate auxiliary verb.
3) Move the auxiliary verb to right after "What".
4) Identify the grammatical form of the main verb and replace the rest of the sentence with the same form of the verb *do*.

B. Passive
(e.g., *What happened to the staff?*)

1) Insert the question word "What" at the beginning of the sentence.
2) Identify the grammatical tense of the main verb and replace the whole predicate with the same form of the verb *happen* (including the auxiliary verb, if any).
3) Insert the preposition *to* left of the subject.
4) Remove the rest of the sentence.

In addition to generating questions, we also generate gap sentences (e.g., for particle verbs, *Chinese retailers have _____ staff.*). These question items can be used as fill-in-the-blanks or multiple choice exercises. In the latter case, one can ensure a deeper level of processing of the target linguistic form by having its synonym as the solution and semantically related words as distractors.

Challenges There is a two-stage process identifying the main syntactic components, POS- and dependency-based, and both of these are obligatory for the system to be able to generate a question. If there is an error, a syntactic component may not be detected. For instance, in example (4), *Skype* was identified as a verb by the statistical parser. Consequently, no subject was detected, and it was not possible to generate a question.

(4) Skype was snapped up by eBay Inc.

The most challenging case that results in generating ungrammatical questions is when the parser incorrectly identifies secondary parts of speech, which does not prevent the system from generating a question. Given the source text (5) below, the question (5a) was generated. The parse tree of the source includes the noun phrase *(NP (VBG meaning) (NNS fans))*, which was then identified as the subject of the sentence.

(5) Internet access in the Communist-ruled island is restricted, meaning fans can not easily look up series and mangas on the web.
   a. What can *meaning* fans not do? *Meaning* fans can not _____ series and mangas on the web.

Another type of error occurs when the coreference resolution component maps a referring expression to the wrong noun phrase. Given the source sentence (6), the program generated the question in (6a). *The manager* is resolved incorrectly as Dean Saunders instead of Chris Coleman.

(6) Former Wales striker Dean Saunders says his country will struggle to hang on to Chris Coleman after their startling run to the Euro 2016 semi-finals and believes the manager could be tempted away soon.
   a. According to the article, what could happen to *former Wales striker Dean Saunders*? Former Wales striker Dean Saunders could be _____ soon.

In questions to subjects and objects, coreference resolution was originally used to determine the question word, *Who* or *What*. However, the error rate was high for rare names that occasionally occur in news articles at the beginning of sentences. Thus, we now combine the two question words in one question phrase *Who or what*. The English teachers we consulted preferred this solution over

erroneously generated question words. To further minimize the effect of errors caused by coreference resolution, we do not substitute the subject of a gap sentence with a pronoun, which often leads to repetition of subject noun phrases.

4.2 Grammar-concept Questions

When it comes to grammar, questions can either focus the reader's attention on the form or the meaning of linguistic forms. In addition to testing the learner's understanding of text, meaning-driven questions also help raise the learner's (meta-)linguistic awareness and read and learn the language in a focused way. Rephrasing and form manipulation is one example of such meaning-driven grammar questions. The passive voice, for instance, is normally substituted with the active voice (or vice versa) to make the learner make inferences based on its semantics.

Similarly, grammar-concept questions make the learner infer information by isolating defining semantic characteristics of linguistic forms. Once the grammatical concept of a linguistic form is broken down into a series of semantic statements, yes/no or alternative questions can be asked about each of these statements. Consider the following example by Workman (2008):

Sentence: He *used to* play football.
Concept: *Used to* expresses a discontinued past habit. It highlights the fact that the person does not do this anymore in the present.
Concept questions:

1. Does he play football now? (No)
2. Did he play football in the past? (Yes)
3. Did he play once or often? (Often)

One important application of grammar-concept questions is scaffolding feedback. The questions can incrementally guide the learner towards task completion by scaffolding the use of correct forms. Grammar-concept questions then not only make the learners aware of the form but also guide them towards production.

Generation Depending on the linguistic form, we use different templates to generate the grammar-concept questions, and we transform the target verb into the appropriate tense form.[1]

Let us take a closer look at the case of the present perfect tense. Its two key characteristics are (i) the finished state of the action and (ii) the irrelevance of the exact time in the past when the action took place. The templates (7) and (8) are used for generating grammar-concept questions about these aspects.

(7) `Be-form subject still verb`ing `(particle) (dir-obj) (indir-obj)`?

e.g., *Are Chinese retailers still cutting staff?*

(8) Is it more important when exactly `subject verb-past (particle) (dir-obj) (indir-obj)` or that `verb`ing `(dir-obj) (indir-obj)` took place at all?

e.g., *Is it more important when exactly Chinese retailers cut staff or that cutting staff took place at all?*

Since the correct answers are known for each template, they can be hard-coded there. As the templates show, a target sentence should always contain a subject and a verb. The particle element is there for the case of particle verbs, and the object elements are optional.

Challenges One limitation of the current implementation of grammar-concept questions is that without identifying the specific interpretation of a grammatical tense, we can only specify rather general templates, one or two per grammatical tense. The task of tense sense disambiguation (Reichart and Rappoport, 2010) is very relevant to our work and can facilitate the creation of more fine-grained templates. For example, in case of the past simple tense, one could also ask about the repetitive versus single occurrence of an action in the past; in case of the present perfect continuous tense, a question about the (in)completeness of an action would be plausible.

5 Comparing computer-generated and human-written questions

For questions to be effective in real-life FLTL, they must be reasonably well-formed and answerable. We therefore conducted a crowdsourcing study[2] to determine how automatically generated questions and manually written questions are perceived in those two respects.

We started with a corpus of 40 news articles and 96 questions written by Simón Ruiz, an English

[1] For this step, we make use of the Java library `https://github.com/simplenlg/simplenlg`.

[2] For this study, we used the CrowdFlower platform: `https://crowdflower.com`

teacher and SLA researcher, to test the learner's knowledge of particle verbs. We used the question generation approach introduced in section 4 and generated 69 form exposure questions to particle verbs. To obtain an equal number of questions for the experiment, we randomly selected 69 questions from the manually created ones.

The crowd workers were selected among proficient speakers of English. This requirement was enforced by a website functionality restricting participating countries, three so-called test questions asking the participants about their level of English and self-perceived reliability of their judgements, and other test questions assessing their proficiency in English, which we now turn to.

In a crowdsourcing experiment, test questions are crucial because they limit the set of workers to those satisfying the requirements and make it possible to verify they are paying attention and follow the instructions. To create test questions assessing the workers' proficiency in English, we first created eight ungrammatical or unanswerable question items as follows: We edited four out of the 27 human-written questions not used in the study and four automatically generated questions to make them either ungrammatical or unanswerable. To obtain test questions on the clearly grammatical and answerable side of the spectrum, we ran a pilot study and selected sentences rated high with a high agreement among the contributors. Four human-written and four computer-generated ones were chosen as good examples of well-formed and answerable test questions. In order for the crowd workers to be eligible to start judging non-test questions, they had to pass through the so-called quiz mode and achieve 70% accuracy on five randomly selected test question items.

We investigated whether computer-generated questions are on a par with human-written ones based on two criteria, well-formedness and answerability. In other words, whether the question is written in acceptable English and whether it can be answered given the information in the source text. In addition, we asked the crowd workers whether they thought the question was written by an English teacher or generated automatically by a computer. Concretely, each task presented to the crowd workers consisted of an excerpt from the source news text and the human-written or automatically-generated question. The workers were asked to answer four questions:

1. How well-formed is this question item? Is it written in good English? (5-point Likert scale)
2. Can this question item be completed with the information from the source text? (5-point Likert scale)
3. Please, answer this question – in your words, in as few words as possible – based on the information from the source text. (free input)
4. Do you think this question was written by an English teacher or generated by a computer? (binary choice)

There also was an optional comment field.

Below you can find an example for a news excerpt (9) and the questions which were written manually (9a) and automatically-generated (9b).

(9) "Scotland is a part of the UK," a spokesman for the European Commission said. "All parts of the UK should sort out what they want to do," he added, calling the options "speculation".

   a. What did a spokesman for the European Commission say about the UK? He said that all parts of the UK should _____ what they want to do.
   b. According to a spokesman for the European Commission, what should all parts of the UK do? All parts of the UK should _____ what they want to do.

We received 1,384 judgements by 364 crowd workers classified as reliable, who identified as proficient English speakers and passed the quiz mode with the test questions. On the well-formedness scale, the means were 4.53 for human-written and 4.40 for computer-generated questions. On the answerability scale, the means were 4.44 and 4.47, respectively. We calculated the intra-class correlation (ICC) for the contributors and got 0.08 and 0.09 for well-formedness and answerability, respectively. The low contributor ICC ($< .1$) implies that the contributors provided different ratings for different question items, so we can ignore the dependencies among the observations and did not need a multi-level analysis.

To find out whether the difference in ratings between computer-generated and human-written questions is statistically significant, we ran Welch's t-test. On the well-formedness scale, the results turned out to be statistically significant, but

the effect size was small: $t(913) = 2.06$, $p = .03$, Cohen's $d = 0.13$. On the answerability scale, the results were non-significant: $t(944) = -0.42$, $p \geq .1$, Cohen's $d = 0.02$.

However, the absence of evidence does not imply the evidence of absence. To test whether the computer-generated and human-written questions are equivalent in quality (well-formedness and answerability), we used Schuirmann's (1987) two one-sided test (TOST). The TOST is commonly used in medical research to determine if one treatment is as effective as another one. To prove our alternative hypothesis that computer-generated and human-written questions are comparable in quality, we needed to reject two parts of the null hypothesis:

$H0_1$: Computer-generated questions are inferior in quality to human-written ones.

$H0_2$: Computer-generated questions are superior in quality to human-written ones.

In statistical terms, the null hypothesis is that there is a true effect larger than a Smallest Effect Size of Interest (SESOS) between the two samples (Lakens, 2014). For this task, we opted for an SESOS of 0.5, a medium effect size according to Cohen (1977), and an alpha level of .05 (Lakens, 2017). We used the R package TOSTER[3] to conduct TOST testing for equivalence of the samples. All results were statistically significant on both scales ($p \leq .001$), so we could reject the null hypothesis (for more details, see Table 1).

Scale	t_1	t_2	p_1 and p_2	90% CI
well-formed	9.81	-5.68	$\leq$.001	[0.02;0.22]
answerable	7.32	-8.17	$\leq$.001	[-0.13;0.08]

Table 1: Results of Schuirmann's TOST for equivalence of computer-generated and human-written questions. Effect size d = 0.5; alpha = 0.05.

The results indicate that any difference in the ratings for well-formedness and answerability of the human-written and computer-generated questions is of an effect size smaller than the SESOS. In line with this finding, the contributors' answers guessing whether a question was written by an English teacher or generated by a computer were similar for both question classes: 74% of human-written and 67% of computer-generated questions were thought to be written by an English teacher. Our goal at this stage was to identify whether the questions as generated can effectively be used on a par with manually written questions – which indeed seems to be the case.

[3] https://cran.r-project.org/web/packages/TOSTER/

6 Conclusion and Outlook

We discussed question generation for FLTL and proposed that, in addition to the typical focus of such work on meaning and understanding, questions can also play an important role for functionally-driven input enhancement. In line with the focus-on-form perspective in Second Language Acquisition research and the notion of structured input activities, such questions help the learner in processing relevant forms and draw form-meaning connections while engaging in a meaning-based activity.

We proposed two types of questions designed to provide functionally-driven input enhancement of a text. *Form exposure questions* serve to engage a learner in more thoroughly processing a sentence containing a targeted form. *Grammar-concept questions* require the learner to interpret the targeted form in addition to processing it. We discussed the transformation- and template-based question generation approach we implemented for this purpose and exemplified the approach for particular tenses and verb classes. To evaluate whether the automatically generated form exposure questions are up to real-life use, we compared the well-formedness and answerability of automatically generated questions targeting particle verbs to human-written questions of the same type. The crowd sourcing results suggest that the automatic question generation can meaningfully be put to real-life use in a system, thereby paving the way for an external evaluation in terms of the learning outcomes that can be achieved by a functionally-driven input enhancement approach.

Using NLP technology integrated in web-based tools to support the intervention, a large-scale randomized controlled field study can be set up and run over an entire semester or school year, which is significantly longer than typical interventions, but it is the time span in which real-life foreign language learning takes place. Crucially, such a setup can also include collection of measures of individ-

ual differences and other relevant factors. For example, grammar-concept questions may be particularly valuable when the learner's first language does not have a particular linguistic form, as suggested by Workman (2008). The data from such an NLP-supported intervention study will stand to showcase the synergy that can result at the intersection of SLA and CL research (Meurers and Dickinson, 2017). In addition to empirically testing and advancing SLA hypotheses, the insights could further improve CL applications by integrating a learner model to parametrize the generation of questions for those target forms that are particularly relevant for a given user.

From the CL perspective, the task of generating such questions is feasible yet challenging and is interestingly intertwined with other NLP tasks. For instance, the tasks of named entity recognition and coreference resolution can be used to make questions more precise. However, there often is a trade-off between allowing for somewhat general phrases ("who or what" as a question phrase) and using a coreference resolution component with a suboptimal accuracy. We intend to explore this trade-off further in the future. In a similar vein, we also intend to develop filters to further reduce the number of generated questions that are suboptimal in terms of well-formedness, typically resulting from errors in parsing the sentence to be questioned.

In terms of conceptual outlook, there also are some issues we intend to pursue. When grammar-concept questions are asked, they may or may not draw the reader's attention to the target linguistic form, especially if semantic redundancy is present. The issue is exemplified by (10).

(10) John *used to* play football, but since moving back to Tuvalu doesn't do so anymore.

a. Does John still play football?

As the semantics of *used to* implies a discontinued past habit, the grammar-concept question shown in (10a) could be generated. However, the clause *doesn't do so anymore* has exactly the same implication, which can interfere with the learner noticing and processing the target linguistic form *used to*. This issue is reminiscent of VanPatten's Preference for Non-redundancy Principle (VanPatten, 2004). Short of changing the text as such, one option for ensure noticing of the relevant target is to combine the function-driven input enhancement with visual input enhancement. In practice, automatic question generation here can be combined with automatic visual input enhancement (Meurers et al., 2010) by both asking a question about the semantics of a targeted linguistic form and highlighting it. Arguably, both types of input enhancement should be preceded by a text selection step that ensures a rich representation of the form to be targeted in the text. A linguistically-aware search engine, such as FLAIR (Chinkina and Meurers, 2016), can provide automatic input enrichment to support teachers and learners in text selection.

In terms of practical plans, we plan to integrate automatic visual and function-driven input enhancement into the FLAIR system. Going further towards activity generation, it could also be attractive to provide an interface from input enrichment and enhancement tools to applications supporting activity generation, such as the Language Muse Activity Palette (Burstein et al., 2017).

Acknowledgments

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 is a doctoral student at the LEAD Graduate School & Research Network.

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.

The crowd sourcing experiment was made possible by the manual gold-standard questions targeting English particle verbs, which are a part of the PhD research of Simón Ruiz. We are grateful to him for providing us with the questions and allowing us to use them in the experiment.

Finally, we would like to thank the anonymous reviewers and the organizers whose feedback helped us improve the paper.

References

Manish Agarwal, Rakshit Shah, and Prashanth Mannem. 2011. Automatic question generation using discourse cues. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*. Portland, OR, pages 1–9. http://aclweb.org/anthology/W11-1401.

Jun Araki, Dheeraj Rajagopal, Sreecharan Sankaranarayanan, Susan Holm, Yukari Yamakawa, and

Teruko Mitamura. 2016. Generating questions and multiple-choice answers using semantic analysis of texts. In *COLING*. pages 1125–1136. http://aclweb.org/anthology/C16-1107.

Lee Becker, Sumit Basu, and Lucy Vanderwende. 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*. Association for Computational Linguistics, pages 742–751. http://aclweb.org/anthology/N12-1092.

Alessandro Benati. 2004. The effects of structured input activities and explicit information on the acquisition of the italian future tense. *Processing instruction: Theory, research, and commentary* pages 207–225.

Alessandro Benati. 2016. Input manipulation, enhancement and processing: Theoretical views and empirical research. *Studies in Second Language Learning and Teaching* 6(1):65–88. https://doi.org/10.14746/ssllt.2016.6.1.4.

Jonathan Brown, Gwen Frishkoff, and Maxine Eskenazi. 2005. Automatic question generation for vocabulary assessment. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada, pages 819–826. http://aclweb.org/anthology/H05-1103.

Jill Burstein, Nitin Madnani, John Sabatini, Dan McCaffrey, Kietha Biggers, and Kelsey Dreier. 2017. Generating language activities in real-time for english learners using language muse. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*. ACM, New York, NY, USA, L@S '17, pages 213–215. https://doi.org/10.1145/3051457.3053988.

Maria Chinkina and Detmar Meurers. 2016. Linguistically-aware information retrieval: Providing input enrichment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. San Diego, CA, pages 188–198. http://aclweb.org/anthology/W16-0521.pdf.

Jacob Cohen. 1977. *Statistical power analysis for the behavioral sciences*. Academic Press, New York.

Sérgio Curto, Ana Cristina Mendes, and Luísa Coheur. 2012. Question generation based on lexico-syntactic patterns learned from the web. *Dialogue & Discourse* 3(2):147–175. https://doi.org/10.5087/dad.2012.207.

Catherine Doughty and John Williams, editors. 1998. *Focus on form in classroom second language acquisition*. Cambridge University Press, Cambridge.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. *arXiv preprint* https://arxiv.org/pdf/1705.00106.

Michael Heilman. 2011. *Automatic factual question generation from text*. Ph.D. thesis, Carnegie Mellon University.

Michael Heilman and Noah A. Smith. 2009. Question generation via overgenerating transformations and ranking. Technical report, DTIC Document.

Stephen Krashen. 1977. Some issues relating to the monitor model. *On Tesol* 77(144-158).

Daniël Lakens. 2014. Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology* 44(7):701–710.

Daniel Lakens. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science* .

James F. Lee and Bill VanPatten. 1995. *Making communicative language teaching happen*, volume 1: Directions for Language Learning and Teaching. ERIC. https://doi.org/10.2307/328644.

Sang-Ki Lee and Hung-Tzu Huang. 2008. Visual Input Enhancement and Grammar Learning: A meta-analytic review. *Studies in Second Language Acquisition* 30:307–331. https://doi.org/10.1017/s0272263108080479.

Ronald P. Leow, Hui-Chen Hsieh, and Nina Moreno. 2008. Attention to form and meaning revisited. *Language Learning* 58(3):665–695. https://doi.org/10.1111/j.1467-9922.2008.00453.x.

Shawn Loewen. 2005. Incidental focus on form and second language learning. *Studies in Second Language Acquisition* 27(3):361–386. https://doi.org/10.1017/S0272263105050163.

Michael H. Long. 1991. Focus on form: A design feature in language teaching methodology. In K. De Bot, C. Kramsch, and R. Ginsberg, editors, *Foreign language research in cross-cultural perspective*, John Benjamins, Amsterdam, pages 39–52.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland, pages 55–60. https://doi.org/10.3115/v1/p14-5010.

Detmar Meurers and Markus Dickinson. 2017. Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning* S1(2). http://dx.doi.org/10.1111/lang.12233.

Detmar Meurers, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf, and Niels Ott. 2010. Enhancing authentic web pages for language learners. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications (BEA-5) at NAACL-HLT 2010*. Los Angeles, pages 10–18. http://aclweb.org/anthology/W10-1002.pdf.

George Miller. 1995. Wordnet: a lexical database for English. *Communications of the ACM* 38(11):39–41. http://aclweb.org/anthology/H94-1111.

Ruslan Mitkov, Le An Ha, and Nikiforos Karamanis. 2006. A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering* 12(2):177–194. https://doi.org/10.1017/S1351324906004177.

Kara Morgan-Short, Jeanne Heil, Andrea Botero-Moriarty, and Shane Ebert. 2012. Allocation of attention to second language form and meaning. *Studies in Second Language Acquisition* 34(04):659–685. https://doi.org/10.1017/s027226311200037x.

J. Mostow, J. E. Beck, J. Bey, A. Cuneo, J. Sison, B. Tobin, and J. Valeri. 2004. Using automated questions to assess readingcomprehension, vocabulary, and effects of tutorial interventions. *Technology, Instruction, Cognition and Learning* 2(1-2):97–134.

Bikram Nobal Niraula and Vasile Rus. 2015. *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, Association for Computational Linguistics, chapter Judging the Quality of Automatically Generated Gap-fill Question using Active Learning, pages 196–206. https://doi.org/10.3115/v1/W15-0623.

John M. Norris and Lourdes Ortega. 2000. Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning* 50(3):417–528. https://doi.org/10.1111/0023-8333.00136.

Michael O'Malley and Anna Chamot. 1990. *Learning Strategies in Second Language Acquisition*. Cambridge University Press, New York.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2013. Automatic selection of suitable sentences for language learning exercises. In L. Bradley and S. Thouësny, editors, *20 Years of EUROCALL: Learning from the Past, Looking to the Future. Proceedings of the 2013 EUROCALL Conference*. pages 218–225. https://doi.org/10.14705/rpnet.2013.000164.

Juan Pino and Maxine Eskenazi. 2009. Semi-automatic generation of cloze question distractors effect of students' l1. In *SLaTE*. pages 65–68.

Juan Pino, Michael Heilman, and Maxine Eskenazi. 2008. A selection strategy to improve cloze question quality. In *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains. 9th International Conference on Intelligent Tutoring Systems*. Montreal, Canada, pages 22–34.

Roi Reichart and Ari Rappoport. 2010. Tense sense disambiguation: a new syntactic polysemy task. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 325–334. http://aclweb.org/anthology/D10-1032.

R. Schmidt. 1990. The role of consciousness in second language learning. *Applied Linguistics* 11:206–226.

Donald J Schuirmann. 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 Pharmacodynamics* 15(6):657–680.

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

Bill VanPatten. 1990. Attending to form and content in the input. *Studies in second language acquisition* 12(03):287–301. https://doi.org/10.1017/s0272263100009177.

Bill VanPatten. 2004. *Processing instruction: Theory, research, and commentary*. Routledge. https://doi.org/10.4324/9781410610195.

Bill VanPatten and Teresa Cadierno. 1993. Explicit instruction and input processing. *Studies in Second Language Acquisition* 15(02):225–243. https://doi.org/10.1017/S0272263100011979.

Bill VanPatten and Soile Oikkenon. 1996. Explanation versus structured input in processing instruction. *Studies in Second Language Acquisition* 18(04):495–510. https://doi.org/10.1017/s0272263100015394.

Wynne Wong. 2001. Modality and attention to meaning and form in the input. *Studies in Second Language Acquisition* 23(03):345–368. https://doi.org/10.1017/s0272263101003023.

Wynne Wong. 2004. Processing instruction in French: The roles of explicit information and structured input. *Processing instruction: Theory, research, and commentary* pages 187–205. https://doi.org/10.4324/9781410610195.

Graham Workman. 2008. *Concept questions and time lines*. Gem Publishing.