Using Learning Analytics for Adaptive Exercise Generation

Tanja Heck
Universität Tübingen / Germany
tanja.heck@
uni-tuebingen.de

Detmar Meurers
Universität Tübingen / Germany
detmar.meurers@
uni-tuebingen.de

Abstract

Single Choice exercises constitute a central exercise type for language learning in a learner's progression from mere implicit exposure through input enhancement to productive language use in open exercises. Distractors that support learning in the individual zone of proximal development should not be derived from static analyses of learner corpora, but rely on dynamic learning analytics based on half-open exercises. We demonstrate how a system's error diagnosis module can be re-used for automatic and dynamic generation and adaptation of distractors, as well as to inform exercise generation in terms of relevant learning goals and reasonable chunking in Jumbled Sentences exercises.

1 Introduction

Supporting language learners to progress in their zone of proximal development requires exercises of different complexities (Shabani et al., 2010). While input enhancement for implicit exposure to linguistic constructions can foster receptive skills at the lower end of the complexity range (Meurers et al., 2010), open exercises that elicit production of linguistic constructions and the entire sentence context constitute the other extreme (Becker and Roos, 2016). In order to advance from one to the other, learners need to acquire the constructions relevant for language production in a controlled way. To this purpose, half-open exercises require learners to produce only the target form whereas closed exercise types provide a range of answer alternatives to choose from (Spada and Tomita, 2010). The closed-type Single Choice (SC) exercises require special attention as they expose learners to incorrect linguistic material in the form of distractor options. While distractors should cover developmental misconceptions in order to be sufficiently challenging and thus relevant to learning, they should not expose learners to any misconceptions they would not have come up with on their own (Yamada, 2019).

Given these considerations, it is not surprising that distractor generation is seen as the most challenging aspect of generating SC exercises (Mitkov et al., 2006). In order to determine pedagogically valid and plausible distractors, human judgement is often deemed best (Susanti et al., 2018), yet even manually created distractors do often not meet these requirements (Haladyna and Downing, 1993; Patil et al., 2016). In order to automate distractor generation and at the same time increase plausibility and validity, data-driven approaches base distractors on common misconceptions of learners (Lee et al., 2016). This in addition allows a more learner-centered adaptation of distractors by dynamically selecting those distractors for each learner from a pool of options that target their individual misconceptions.

However, abstracting learner errors into patterns that facilitate generating distractors for arbitrary target answers is not a trivial task. On the other hand, many Intelligent Language Tutoring Systems (ILTS) incorporate error diagnosis mechanisms. Approaches anticipating the most common correct and incorrect learner answers, henceforth referred to as answer hypotheses, and matching them to error diagnoses (Meurers, 2012), are particularly interesting for distractor generation. An example that successfully pursues this approach constitutes the ILTS *FeedBook* (Rudzewitz et al., 2018). The process shows strong similarities to distractor generation: The most frequent learner errors constitute the most plausible distractors whereas alternative, correct answers represent unreliable distractors that need to be avoided. Systems generating answer hypotheses for error diagnoses therefore inherently have the means to automatically generate distractors. This is especially valuable if SC exercises are used for remedial practice as it opens the possibility to directly associate SC exercises with learner errors and select exercises that best target the learner's misconceptions. The parallels of error

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 44–56

analysis based on answer hypotheses and distractor generation are striking, yet these two subfields of Natural Language Processing (NLP) have never been approached in tandem.

Although previous approaches to exercise generation have used learner errors solely for distractor generation, they can similarly inform chunking of Jumbled Sentences for word order practice, and determination of required exercise material. Grammatical constructions that are not challenging for learners do not need excessive practice. On the other hand, constructions where learners make many errors should be practiced in a variety of exercises focusing on remedying these misconceptions.

In order to fill the gap, we show the feasibility of using a system's error diagnosis mechanism for distractor generation, as well as for sentence chunking and learning goal definition, at the example of real learner data collected in the Interact4School (I4S) study (Parrisius et al., 2022a,b).

The rest of the paper is structured as follows: Section 2 presents related work on distractor generation. After outlining the research questions and the approach to answer them in section 3, section 4 introduces the data on which the approach was piloted. Section 5 describes the pilot analyses and presents their results before section 6 concludes with a summary.

2 Related work

Distractor generation usually consists of candidate generation and candidate filtering and/or ranking, although they are sometimes executed in a single step. Many approaches combine a number of different filtering and re-ranking approaches.

For question answering and vocabulary-focused gap exercises, approaches differ in the source from which the pool of distractor candidates is compiled, as well as in the filtering and ranking strategies. The candidates are either extracted from unstructured data such as text corpora (Quan et al., 2018; Gates, 2011), from structured data such as databases (Karamanis et al., 2006; Smith et al., 2009) or word lists (Coniam, 1997; Shei, 2001), or else generated based on machine learning (Liang et al., 2017; Sakaguchi et al., 2013) or on transformation rules (Žitko et al., 2009). The candidate pool then comprises either a subset (Sumita et al., 2005; Stasaski and Hearst, 2017) or all entries (Smith et al., 2010; Pérez and Cuadros, 2017) of the resource, or transformations thereof (Maritxalar et al., 2011; Quan et al., 2018) or of the target answer (Zesch and Melamud, 2014). Filtering and ranking depend on the intended distractor type such as ungrammatical, nonsensical and plausible distractors (Mostow and Jang, 2012), which determines for example the usefulness of grammaticality checks (Pino et al., 2008; Moser et al., 2012). For plausible distractors, the desired similarity of the distractors with the target answer constitutes an additional factor. This is on the one hand influenced by the task setup as for example synonyms may be context-inappropriate and therefore useful distractors for contextualized exercises (Knoop and Wilske, 2013), yet would constitute unreliable distractors if they can correctly replace the target answer (Hill and Simha, 2016). In addition, since exercise difficulty increases with distractor plausibility, target similarity can be adjusted according to the learner's proficiency (Alsubait et al., 2015; Chen et al., 2015; Correia et al., 2012). Similarity can target the surface form (Jiang and Lee, 2017), linguistic complexity (Lee and Seneff, 2007; Susanti et al., 2018), phonetics (Mitkov et al., 2009), morphology (Goto et al., 2010), syntax (Guo et al., 2016), or semantics (Susanti et al., 2015) and be based on NLP tools including part-of-speech taggers (Liu et al., 2005), latent semantic analysis (Aldabe and Maritxalar, 2014) and word embedding models (Kumar et al., 2015; Yeung et al., 2019), on external resources such as ontologies (Papasalouros et al., 2008), WordNet (Mitkov et al., 2006; Brown et al., 2005) or FrameNet (Pilán and Volodina, 2014), or else on statistical methods including classification (Welbl et al., 2017; Gao et al., 2020), regression (Liu et al., 2017) and deep learning (Liang et al., 2018). If the final candidate selection is not based on the ranking, it may be left to the user (Nikolova, 2009), or done randomly (Araki et al., 2016; Gutl et al., 2011).

While automatic distractor generation has been widely explored for vocabulary exercises, distractors for grammar exercises have received less attention. With closed class grammatical constructions such as prepositions, many of the approaches used for vocabulary distractors are applicable. However, this greatly underrates the importance of linking distractors to the pedagogical learning goal as good distractors characterize the space of options that a learner needs to weigh against each other. Since the focus of form-based grammar exercises is not on semantics but on form, they usually rely on

ungrammatical distractors (Volodina et al., 2014). Goto et al. (2010) illustrate that for closed class target answers, the initial candidate pool consists of all types belonging to the class, whereas for open class target answers, transformations may produce suitable distractors. For the closed class of prepositions, Lee et al. (2016) start with the defined set of prepositions as candidates. For ranking, they consider co-occurrence of the candidates with either the prepositional object or the head, and their frequency as annotated errors or learner-corrected tokens in a learner corpus. Suitable for open class types, Chen et al. (2006) use distractor generation rules for a defined set of construct patterns which introduce modifications of the target answer such as morphological or syntactic variants. Aldabe et al. (2007) present an approach to generate morphological transformations of the target answer as distractor candidates and filter out those whose morpho-syntactic pattern can be found in a corpus. For verb exercises, Aldabe et al. (2009) filter the verbs from the Academic Word List by transitiveness, tense and person, and rank them according to semantic similarity and distributional data. Heck and Meurers (2022) apply NLP- as well as rule-based transformations to generate well- and ill-formed variations of the target answer.

Lee et al. (2016) found distractor generation based on learner errors to yield the most plausible distractors. While their approach is closest to what we suggest, it relies on a manually annotated corpus. The resulting, statically determined distractors may be sufficiently representative for the learner population that provided the error corpus, yet they are likely to be unsuitable when more widely applied and do not allow to adapt to an individual learner's abilities. We therefore illustrate how automatic annotations obtained from a system's error diagnosis mechanism can effectively be used to generate and dynamically select valid and plausible distractors.

3 Approach

We evaluated a dataset of learner answers to form-based grammar exercises with the aim of answering the following research questions:

- RQ.1 Can the creation of learning goals, distractors and JS chunks be automated through learning analytics?
- RQ.2 Does human perception of relevant misconceptions align with relevant misconceptions derived from learning analytics?
- RQ.3 Do errors made in half-open exercises constitute plausible distractors of closed exercises?

In order to answer RQ.1, in the following we indicate which steps of the evaluations could not be based on automated processing of the data but instead required manual labour. In addition, we determined the ability of the system's error diagnosis module to identify relevant errors automatically. This on the one hand outlines the status quo of possible automization and on the other hand indicates future directions for extending the module in order to support the envisioned learning analytics based adaptivity.

In order to answer RQ.2, we first identified the most frequent errors made in half-open exercises. To this end, we determined misconceptions of interest by freely annotating the entire dataset once without any reference set of potential labels. Of the thus compiled labels, those specific to questions in the simple past were included in the final label set. In order to develop an annotated learner corpus from the learner answers, we relied on two sources: (a) automatic annotations provided by the system's error diagnosis module, and (b) manual annotations. The automatic annotations provide the single most relevant error for each learner answer. They were refined into more fine-grained labels if simple string matching was sufficient and mapped to the label set. We used these annotations whenever available ($n = 1,778$) and manually annotated the remaining learner answers ($n_{answers} = 3,058$, $n_{labels} = 6,576$) if the system could not diagnose the nature of the error. Five annotators with backgrounds in computational linguistics annotated the learner answers independently with an unconstrained number of labels. Inter-Annotator Agreement (IAA) for the multi-label annotations of all annotators was calculated as Krippendorff's alpha at $\alpha = .2075$. For the evaluations, the union set of manual and automatic annotations was used in order to not miss any potential errors. Although this might introduce some noise, it serves the purpose of identifying distractor candidates best.

In a second step, we contrasted the learner errors against misconceptions judged relevant by human exercise creators. To this purpose, we analyzed the available exercises, distractors and JS chunks of with respect to the errors for which they provide opportunities. We annotated the exercises with the

same labels used for learner error annotations. A label was assigned if it is in principle possible to make the associated error in the exercise.

Errors made in half-open exercises can only inform distractor generation if learners tend to choose the associated distractors in SC exercises. Similarly, separating constituents into individual chunks only supports learning if learners fail to put these chunks into the correct order in JS exercises. In order to answer RQ.3, we therefore analyzed whether the identified most frequent errors were also made in SC and JS exercises if the exercises provided opportunities to make them.

4 Data

The evaluations are based on data obtained in the I4S project. The study collected data from 7th grade learners of English as a second language in German secondary schools who worked with the *FeedBook* over the course of a school year. The ILTS offers practice exercises in a task based setting with intelligent feedback provided to the learners as they work on the exercises. The subset of the data used for the pilot evaluations consists of the exercises on questions in the simple past.

The resulting dataset is based on 132 exercise items of the four exercise types illustrated in Figures 8–11 of Appendix A: 27 Jumbled Sentences (JS) whose chunks learners have to put into the correct order; 27 SC items for which learners need to select the correct option from the dropdown; 58 Fill-in-the-Blanks (FiB) items with input fields into which learners must write the target form; and 20 Short Answer (SA) items which require learners to write a sentence in response to a prompt. 10 of the FiB items present all correct forms to insert into the blanks as bags of words in the exercise instructions instead of giving lemmas in parentheses behind the blanks. As this renders them more similar to SC exercises, we treat them as such. FiB and SA exercises constitute half-open exercise types while SC and JS exercises are closed types. A total of 4,836 incorrect learner answers to an actionable element of the exercises was collected from 199 learners who submitted at least 1 of the exercises. An actionable element is defined as the blank of a FiB or SC exercise, a chunk of a JS exercise, or an answer to a SA exercise. All submissions were considered so that there may be multiple answers per learner and actionable element if a learner re-submitted a revised answer.

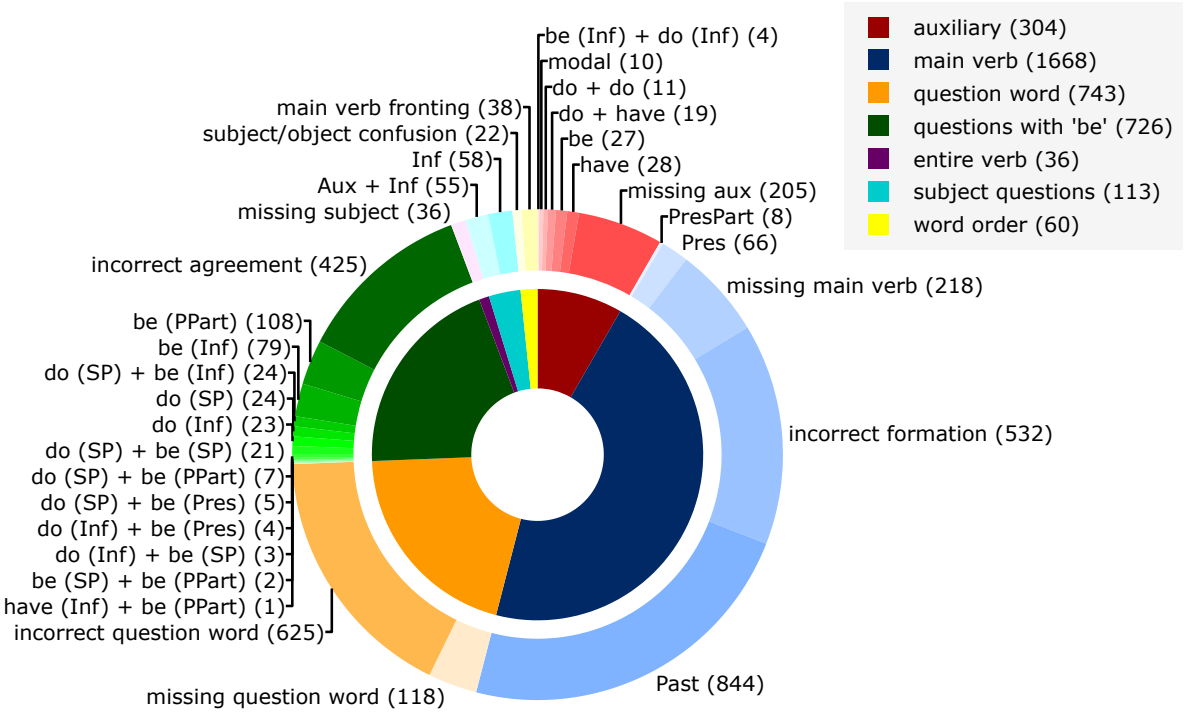

Figure 1: Frequencies of error types

5 Evaluation

While the focus of the analyses is on distractor generation, we also evaluated the feasibility of using the system's error diagnosis module to determine relevant learning goals and generate chunks of JS exercises.

5.1 Learning goal selection

Learning goals comprise pedagogically motivated groupings of learner errors. We therefore manually identified linguistically and pedagogically related groups of error labels.

The resulting seven groups of errors that constitute important learning goals are illustrated in Figure 1: Auxiliary errors, main verb errors, errors targeting the entire verb, question word errors, word order errors, errors in questions with 'be', and errors in subject questions. The latter two constitute interesting special cases since question formation rules for them differ from the general rule. Their relevance as separate learning goals is strikingly emphasized when normalizing the error frequencies by the opportunities to make the respective error, as illustrated in Figure 2. Exercise generation should thus ensure to generate exercises targeting these seven learning goals for questions in the simple past.

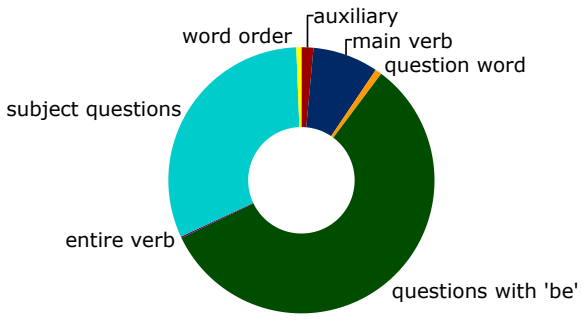

Figure 2: Normalized frequencies of error types

Focusing on RQ.1, we verified how well the system's existing error diagnoses reflect the labels identified as relevant to exercises on questions in the simple past. To this purpose, we determined the overlap between the labels used in manual and in automatic annotations. In addition, we manually annotated a subset ($n = 491$) of the automatically annotated learner errors and calculated multi-label IAA between the automatic and the joint manual annotations.

The automatic annotations cover 34 of the 63 labels found relevant for exercises on questions in the simple past. Although this leaves substantial potential for extensions of the error diagnosis module, it also provides a solid starting point for further analyses. Automatic annotations include only a single label per error, yet IAA with the manual annotations was even slightly higher than that for the human annotators at $\alpha = .2175$. The error diagnosis module can therefore be used for purposes of automatic exercise generation, although both applications would benefit from extending the coverage of diagnosed learner errors.

In order to address RQ.2, we examined the exercises in the system. They evidently provide practice opportunities for all identified misconceptions as the errors were observed in the ILTS' learner records. Yet the numbers of opportunities might differ from one misconception and exercise type to the other. In order to evaluate the available exercises' coverage of the identified learning goals, we determined the exercise annotations' coverage of the error labels.

The analysis reveals that not all exercise types offer practice for all misconceptions. Figure 3 illustrates that not all learning goals relevant according to the learner records can currently be practiced both with closed and half-open exercises. Thus, there is no perfect overlap between learning goals introduced by human exercise creators and those identified through learning analytics.

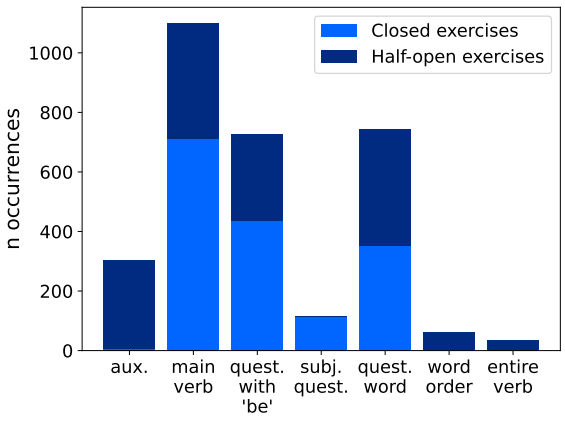

Figure 3: Error frequencies per exercise type

5.2 Distractor generation

While feedback generation aims to cover as many learner errors as possible, distractor generation needs to focus on the most frequent learner errors. This requires to filter the output of the answer hypotheses generated for feedback provision. Tversky (1964) found 3-option SC exercises to be

the most reliable. We therefore aimed to determine the two most frequent errors made in half-open exercises as distractors for SC exercises. Since not all error types can be made in all exercises, we normalized the occurrences of misconceptions by the number of exercise items that provided opportunities to make the error.

Figures 4–6 present normalized error frequencies per exercise type, indicating (through coloured dots next to the frequency bars) whether the system provides exercises with opportunities to make the error.

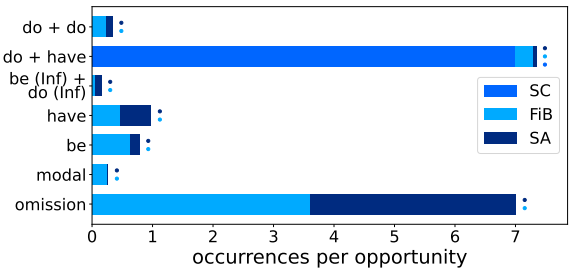

Figure 4: Frequencies of errors targeting the auxiliary

The most frequent error with respect to **auxiliaries** made by learners (see Figure 4) consists in leaving it out (e.g., Example 1a). Of the remaining errors observed in half-open exercises, using *be* (e.g., Example 1b) or *have* (e.g., Example 1c) instead of the auxiliary *do* are most frequent. Combinations of multiple auxiliaries (e.g., Example 1d) are also observed, but only in occasional submissions of half-open exercises.

(1) What did Mr. Connor bake?

a. *What baked Mr. Connor?
b. *What was Mr. Connor bake?
c. *What had bake Mr. Connor?
d. *What does Mr. Connor have bake?

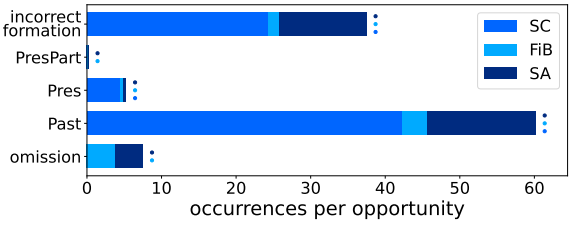

Figure 5: Frequencies of errors targeting the main verb

With respect to the **main verb** (see Figure 5), the most frequent learner error consists in using the simple past or past participle form instead of the infinitive (e.g., Example 2a). Omitting the main verb altogether (e.g., Example 2b), using simple present – identifiable through the third person singular 's' – (e.g., Example 2c), or incorrectly forming the main verb (e.g., Example 2d) were also observed rather frequently in learner answers. The latter error constitutes a special case in that it occurs only in combination with other misconceptions. Since infinitives do not transform the verb, learners always give the correct form if they intend to provide the verb in this mood. Other misconceptions appear only occasionally in learner answers.

(2) Did you enjoy them?

a. *Did you enjoyed them?
b. *Did you them?
c. *Did you enjoys them?
d. *Did you enjoyd them?

Only a single misconception, omitting the subject, is relevant to the learning goal practicing the **entire verb**. This error can be found with FiB as well as SA exercises, which constitute the two exercise types providing opportunities for the error.

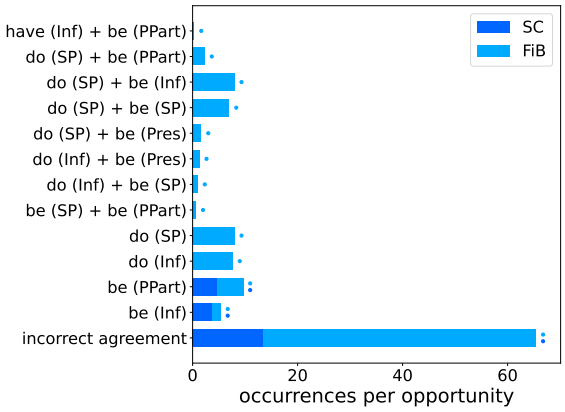

Figure 6: Frequencies of errors in questions with 'be'

The most frequent error concerning **questions with 'be'** (see Figure 6) by far constitutes incorrect agreement with the person of the subject (e.g., Example 3a). With FiB exercises, additional do-support (*did be*, e.g., Example 3b), *did was/were* (e.g., Example 3c), *did* (e.g., Example 3d), and *do* (e.g., Example 3e) are also frequent and should therefore be considered for distractor generation.

(3) Were you scared?

a. *Was you scared?
b. *Did you be scared?

c. *Did you was scared?
d. *Did you scared?
e. *Do you scared?

As there are no half-open exercises available for **subject questions**, it is not possible to determine from the data what kind of errors learners would produce on their own. Observed misconceptions are therefore restricted to those offered by the SC distractors. They consist in using only the infinitive of the main verb (e.g., Example 4a) or else the infinitive with do-support (e.g., Example 4b).

(4) Who persuaded you to come to the party?
a. *Who persuade you to come to the party?
b. *Who did persuade you to come to the party?

Exercises on **question words** constitute a special case in that misconceptions are specific to the target question word. The bar chart in Figure 7 illustrates that although almost all question word confusions are present in the dataset, there are clearly discernable, predominant misconceptions in the use of question words. These are, however, not bidirectional. While *where* is often incorrectly substituted by *what* or *when* in the normalized dataset, the most frequently used question words instead of *what* are *how* and *which*, and omitting the question word altogether or using *where* is the most frequent error with *when*. Instead of *why*, learners most often used *how* or *who*, whereas the most frequent question word instead of *how* is *what* or sometimes *why* in the dataset.

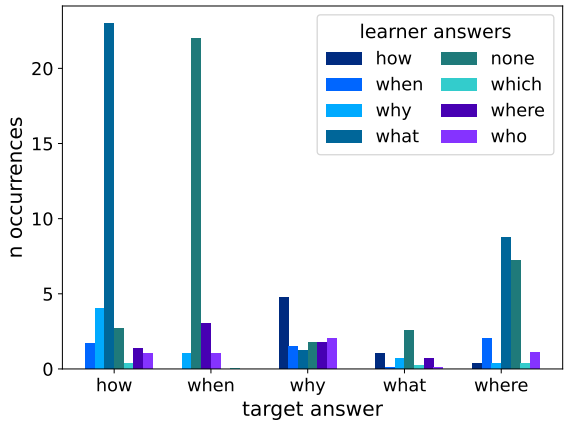

Figure 7: Frequencies of question word confusions

Turning to RQ.3, we analyzed whether the identified most frequent errors of half-open exercises appear in SC exercises as well if according distractors are available.
The system's distractors do not cover the most frequent errors for all learning goals. With respect to the main verb, they support three of the most frequent misconceptions identified in half-open exercises: using a past form, simple present, or an incorrectly formed variant of the main verb. These distractors are selected frequently by learners in SC exercises. Concerning questions with 'be', SC exercises offer distractors targeting the most frequent misconception. These distractors are also selected frequently by learners. This indicates that errors observed in half-open exercises constitute plausible distractors for SC exercises, although available distractor coverage is too scarce to confirm general validity of this assumption.

With respect to RQ.1, we determined whether the automatic annotations support the labels of the two most frequent errors in half-open exercises.
Looking at the individual learning goals, the most frequent misconceptions with auxiliaries – omission and the use of *be* – are both supported by the automatic annotations so that the error diagnosis module is already able to generate such distractors. The automatic annotations do not yet cover any of the misconceptions concerning the main verb, the entire verb, or questions with 'be'. They do, however, support all labels for question word errors, thus providing the means to automatically generate according distractors.

Focusing on RQ.2, we compared the distractors introduced by human exercise creators with the most frequent errors in the learner records.
For **auxiliaries**, available distractors cover only the use of *do* + *have* out of the identified misconceptions. Although this distractor was selected very frequently in SC exercises, the error appears only occasionally in half-open exercises. The most frequent error with respect to this learning goal, leaving it out, is not covered by any of the SC distractors in the system. This makes sense considering that according SC exercises focusing only on the auxiliary would require an empty distractor. This exercise type thus does not lend itself well for practice of omission errors. However, the system's SC exercises do not cover any of the remaining observed misconceptions either.

The distractors cover three of the most frequent misconceptions identified in half-open exercises practicing the **main verb**: using a past form, simple present, or incorrectly formed variant of the

main verb. Only omitting the main verb altogether, which was also observed rather frequently, is not covered for the above mentioned reason.

The system does not provide any SC exercises to practice the **entire verb**.

Concerning **questions with 'be'**, SC exercises offer distractors targeting the most frequent misconception, incorrect agreement, as well as the use of the past participle (*been*, e.g., Example 5a), and of the infinitive of *be* (e.g., Example 5b) instead of its simple past form. However, the latter two are not among the most frequent learner errors.

(5) Were you scared?

a. *Been you scared?

b. *Be you scared?

In general, while the distractors do not cover all misconceptions found in the learner submissions, coverage of the identified most frequent errors is high. Only those targeting word order, which is better practiced with JS exercises, and omission errors are not covered by the distractors. Concerning their pedagogical validity, solely misconceptions that are only covered by SC but not half-open exercises, i.e., those of subject questions, do not appear at all with half-open exercises. The same holds for co-occurrences of labels, indicating that the available distractors only integrate combinations of misconceptions that learners also tend to make jointly in production exercises. The manually created distractors therefore seem to be pedagogically valid since the system does not expose learners to misconceptions they would not develop of their own accord. However, in addition to the misconceptions covered by the error labels, the distractors encompass errors that have not been identified as pedagogically relevant in the manual annotation and selection process. Although both distractor creation and learning goal identification constituted manual processes, they thus put different foci on targeted misconceptions. This might indicate that exercise creators do not intuitively choose distractors that are relevant to the learning goal.

In order to compare manually created distractors to those informed by learning analytics in terms of plausibility, we followed Haladyna and Downing (1993)'s approach which states that at least 5% of all incorrect answers to the question need to correspond to a distractor in order for it to be plausible. We calculated the ratio of n times a distractor was selected over m times any of the item's incorrect options was selected. Distractors obtaining a ratio lower than .05 are thus considered implausible. The evaluation shows that all distractors were selected at least once, although with differing frequencies. Only two instances of distractors were beneath the 5% threshold. While the incorrect form *forgat* may indeed be implausible, there is no clear indication as to why the form *been* was selected so rarely in the distractor group *be - was - been*, which appears in the same constellation in various other (preceding and succeeding) items, where this distractor was selected more frequently.

5.3 Sentence chunking

Jumbled Sentences are a natural choice of exercise type for controlled practice of word order. In order to constitute useful practice material, the chunks should fulfill two criteria: (a) They should be small enough to separate the challenging constituents that learners may struggle to assemble in the correct order. (b) On the other hand, the chunks should only be as small as necessary so as not to distract from the learning goal. We therefore analyzed word order errors with the goal of identifying constituents that should be extracted into individual chunks.

The errors particular to questions in the simple past and targeting word order concern fronting of the main verb before the subject (e.g., Example 6a), as well as interchanging the subject and the object of the sentence (e.g., Example 6b). Relevant chunks for JS exercises therefore comprise a chunk for the main verb, for the subject, and for the object.

(6) Did Mr. Jones see a doctor?

a. *Did see Mr. Jones a doctor?

b. *Did a doctor see Mr. Jones?

With respect to RQ.1, the automatic annotations do not further distinguish between word order errors. Thus, the current error diagnosis cannot determine the most appropriate chunking for a learner.

Addressing RQ.2, we analyzed the JS exercises in the system. For the first criterion concerning sentence chunking, we determined whether the exercises provide opportunities to make the word order errors observed in half-open exercises. In the exercises, 10 out of the 27 items merge the main verb with the succeeding token, thus not supporting main verb fronting errors. Only 11 items have individual chunks for the subject and the object, while the remaining 16 items have either no object or merge it with the preceding preposition or suc-

ceeding main verb.
For the second criterion, we determined the number of remaining chunks not corresponding to a constituent involved in any of the errors. To this end, we subtracted the general number of word order relevant constituents from the number of the exercise item's chunks. Allowing for some preceding and succeeding co-text, we defined results greater than two as indicative of excessive chunking. The sentences in the system are split into a mean of 5.33 chunks ($\sigma = .88$) so that according to the criterion of $n(= 3)$ relevant chunks $+2$, the overall number of chunks is only sightly higher than the optimal number. Considering that most exercise items merge some of the relevant chunks with preceding or succeeding tokens or do not incorporate them at all, however, the exercises do contain substantial excessive chunking.

Regarding RQ.3, the learner error data reveals that while JS exercises offer potential for all observed relevant word order errors, none of the learners made any main verb fronting errors in these exercises, indicating that this is only an issue in more open exercises. Subject/object confusion, on the other hand, was only observed with JS and FiB, but not with SA exercises, although all three exercise types offer opportunities for this error. Since it is of a more semantic nature, this could suggest that learners do not put much effort into semantically parsing sentences in less open-ended exercise types, rendering subject/object errors careless mistakes rather than misconceptions. Thus, neither subject/object confusion nor main verb fronting seem to be relevant for JS exercises. This might suggest that JS exercises are not relevant for practicing question formation and that word order issues arise mostly in combination with formation issues so that learners cannot practice these issues with form-controlling JS exercises. On the other hand, the fact that learners only make the errors in exercises where they have to focus on multiple linguistic aspects at once could also indicate that they lack proceduralization which would allow them to overcome processing overload. In this case, JS exercises could provide opportunities to practice each aspect in isolation.

6 Conclusion

We outlined a data-driven approach to determine relevant learning goals, distractors and sentence chunking for the generation of form-based grammar exercises.

Addressing our first research question, we demonstrated the feasibility of using a system's error diagnosis mechanism to automatically annotate learner errors made in half-open exercises in order to dynamically adapt distractors to a learner's misconceptions. Although not all of the most frequent errors are automatically annotated in the piloted system, it is possible to extend the error diagnosis module to generate all relevant answer hypotheses. Distractor generation and error diagnosis can work hand in hand to this end. We also highlight the relevance of human involvement in the selection of pedagogically valid misconceptions. Pre-filtering of distractor templates should be manual and pedagogically motivated, while ranking of the candidates is best informed by learning analytics. The presented evaluations of most frequent learner errors based on the entire learner corpus serve as exemplary application to an adaptation module, and at the same time may be used as initial settings while the system still lacks learner records for individually adapted exercise configurations.

With respect to the second research question, we found that while there is substantial overlap between human intuition and learning analytics based exercise generation, they also differ in the focus they put on different misconceptions. Since this focus is inconsistent in human output depending on the specific task at hand, human exercise creators might benefit from explicitly specifying the learning goal in a first step. Our evaluations suggest that highest pedagogic validity of exercises can be achieved by relying on human effort to define learning goals, and on learning analytics based, automatic processing for exercise generation.

The third research question cannot be answered conclusively since the exercises do not cover all potential misconceptions for all exercise types. Where no learner data from half-open exercises is available, no conclusions can be drawn about the pedagogical validity of learner errors as distractors. This constitutes a limitation of the presented evaluations. Future work will therefore need to determine whether the errors that learners make in half-open exercises are also good distractors for SC exercises or whether learners instantly perceive them as incorrect when contrasted against the correct option. It is also yet unclear to what extent the most frequent misconceptions differ between and within learners over extended periods of time.

References

Itziar Aldabe and Montse Maritxalar. 2014. Semantic Similarity Measures for the Generation of Science Tests in Basque. *IEEE Transactions on Learning Technologies*, 7(4):375–387.

Itziar Aldabe, Montse Maritxalar, and Edurne Martinez. 2007. Evaluating and improving the distractor-generating heuristics. In *Workshop on NLP for Educational Resources. In conjuction with RANLP07*, pages 7–13.

Itziar Aldabe, Montse Maritxalar, and Ruslan Mitkov. 2009. A Study on the Automatic Selection of Candidate Sentences Distractors. In *Artificial Intelligence in Education*, volume 200, pages 656–658.

Tahani Alsubait, Bijan Parsia, and Uli Sattler. 2015. Generating Multiple Choice Questions From Ontologies: How Far Can We Go? In *Knowledge Engineering and Knowledge Management: EKAW 2014 Satellite Events, VISUAL, EKM1, and ARCOE-Logic, Linköping, Sweden, November 24-28, 2014. Revised Selected Papers. 19*, pages 66–79. Springer.

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 *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1125–1136.

Carmen Becker and Jana Roos. 2016. An approach to creative speaking activities in the young learners' classroom. *Education Inquiry*, 7(1):27613.

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*, HLT '05, pages 819–826, USA. Association for Computational Linguistics.

Chia-Yin Chen, Hsien-Chin Liou, and Jason S. Chang. 2006. FAST – An Automatic Generation System for Grammar Tests. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 1–4.

Tao Chen, Naijia Zheng, Yue Zhao, Muthu Kumar Chandrasekaran, and Min-Yen Kan. 2015. Interactive Second Language Learning from News Websites. In *Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications*, pages 34–42.

David Coniam. 1997. A Preliminary Inquiry Into Using Corpus Word Frequency Data in the Automatic Generation of English Language Cloze Tests. *CALICO Journal*, 14.

Rui Correia, Jorge Baptista, Maxine Eskenazi, and Nuno J. Mamede. 2012. Automatic Generation of Cloze Question Stems. In *PROPOR*, pages 168–178. Springer.

Lingyu Gao, Kevin Gimpel, and Arnar Jensson. 2020. Distractor Analysis and Selection for Multiple-Choice Cloze Questions for Second-Language Learners. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 102–114.

Donna Marie Gates. 2011. How to Generate Cloze Questions from Definitions: a Syntactic Approach. In *2011 AAAI Fall symposium series*.

Takuya Goto, Tomoko Kojiri, Toyohide Watanabe, Tomoharu Iwata, and Takeshi Yamada. 2010. Automatic Generation System of Multiple-Choice Cloze Questions and its Evaluation. *Knowledge Management & E-Learning: An International Journal*, 2:210–224.

Qi Guo, Chinmay Kulkarni, Aniket Kittur, Jeffrey P. Bigham, and Emma Brunskill. 2016. Questimator: Generating Knowledge Assessments for Arbitrary Topics. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 3726–3732, New York, New York, USA. AAAI Press.

Christian Gutl, Klaus Lankmayr, Joachim Weinhofer, and Margit Hofler. 2011. Enhanced Automatic Question Creator–EAQC: Concept, Development and Evaluation of an Automatic Test Item Creation Tool to Foster Modern e-Education. *Electronic Journal of e-Learning*, 9(1):23–38.

Thomas M. Haladyna and Steven M. Downing. 1993. How Many Options is Enough for a Multiple-Choice Test Item? *Educational and Psychological Measurement*, 53(4):999–1010.

Tanja Heck and Detmar Meurers. 2022. Parametrizable exercise generation from authentic texts: Effectively targeting the language means on the curriculum. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 154–166, Seattle, Washington. Association for Computational Linguistics.

Jennifer Hill and Rahul Simha. 2016. Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and Google n-grams. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 23–30.

Shu Jiang and John Lee. 2017. Distractor Generation for Chinese Fill-in-the-blank Items. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148, Copenhagen, Denmark. Association for Computational Linguistics.

Nikiforos Karamanis, Ruslan Mitkov, et al. 2006. Generating Multiple-Choice Test Items from Medical Text: A Pilot Study. In *Proceedings of the fourth international natural language generation conference*, pages 111–113.

Susanne Knoop and Sabrina Wilske. 2013. WordGap - Automatic Generation of Gap-Filling Vocabulary Exercises for Mobile Learning. In *Proceedings of the second workshop on NLP for computer-assisted language learning at NODALIDA 2013; May 22-24; Oslo; Norway. NEALT Proceedings Series 17*, 086, pages 39–47. Citeseer.

Girish Kumar, Rafael Banchs, and Luis D'Haro. 2015. RevUP: Automatic Gap-Fill Question Generation from Educational Texts. In *BEA@NAACL-HLT*, pages 154–161.

John Lee and Stephanie Seneff. 2007. Automatic Generation of Cloze Items for Prepositions. In *Eighth Annual Conference of the International Speech Communication Association*.

John Lee, Donald Sturgeon, and Mengqi Luo. 2016. A CALL System for Learning Preposition Usage. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 984–993, Berlin, Germany. Association for Computational Linguistics.

Chen Liang, Xiao Yang, Neisarg Dave, Drew Wham, Bart Pursel, and C. Lee Giles. 2018. Distractor Generation for Multiple Choice Questions Using Learning to Rank. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 284–290, New Orleans, Louisiana. Association for Computational Linguistics.

Chen Liang, Xiao Yang, Drew Wham, Bart Pursel, Rebecca Passonneaur, and C. Lee Giles. 2017. Distractor Generation with Generative Adversarial Nets for Automatically Creating Fill-in-the-Blank Questions. In *Proceedings of the Knowledge Capture Conference*, K-CAP 2017, New York, NY, USA. Association for Computing Machinery.

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

Ming Liu, Vasile Rus, and Li Liu. 2017. Automatic Chinese Multiple Choice Question Generation Using Mixed Similarity Strategy. *IEEE Transactions on Learning Technologies*, 11(2):193–202.

Montse Maritxalar, Elaine Dhonnchadha, Jennifer Foster, and Monica Ward. 2011. Quizzes on Tap: Exporting a Test Generation System from One Less-Resourced Language to Another. In *Human Language Technology Challenges for Computer Science and Linguistics*, volume 8387, pages 502–514, Cham. Springer International Publishing.

Detmar Meurers. 2012. Natural Language Processing and Language Learning. In *The Encyclopedia of Applied Linguistics*. John Wiley & Sons, Ltd.

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 NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 10–18, Los Angeles, California. Association for Computational Linguistics.

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

Ruslan Mitkov, Andrea Varga, Luz Rello, et al. 2009. Semantic similarity of distractors in multiple-choice tests: extrinsic evaluation. In *Proceedings of the workshop on geometrical models of natural language semantics*, pages 49–56.

Josef Robert Moser, Christian Gütl, and Wei Liu. 2012. Refined Distractor Generation with LSA and Stylometry for Automated Multiple Choice Question Generation. In *Australasian Conference on Artificial Intelligence*, pages 95–106. Springer.

Jack Mostow and Hyeju Jang. 2012. Generating Diagnostic Multiple Choice Comprehension Cloze Questions. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 136–146.

Ivelina Nikolova. 2009. New Issues and Solutions in Computer-aided Design of MCTI and Distractor Selection for Bulgarian. In *Proceedings of the Workshop Multilingual resources, technologies and evaluation for central and Eastern European languages*, pages 40–46.

Andreas Papasalouros, Konstantinos Kanaris, and Konstantinos I. Kotis. 2008. Automatic Generation Of Multiple Choice Questions From Domain Ontologies. In *e-Learning*.

Cora Parrisius, Ines Pieronczyk, Carolyn Blume, Katharina Wendebourg, Diana Pili-Moss, Mirjam Assmann, Sabine Beilharz, Stephen Bodnar, Leona Colling, Heiko Holz, et al. 2022a. Using an Intelligent Tutoring System within a Task-Based Learning Approach in English as a Foreign Language Classes to Foster Motivation and Learning Outcome (Interact4School): Pre-registration of the Study Design.

Cora Parrisius, Katharina Wendebourg, Sven Rieger, Ines Loll, Diana Pili-Moss, Leona Colling, Carolyn Blume, Ines Pieronczyk, Heiko Holz, Stephen Bodnar, et al. 2022b. Effective Features of Feedback in an Intelligent Tutoring System-A Randomized Controlled Field Trial (Pre-Registration).

Rajkumar Patil, Sachin Bhaskar Palve, Kamesh Vell, and Abhijit Vinod Boratne. 2016. Evaluation of multiple choice questions by item analysis in a medical college at Pondicherry, India. *International Journal of Community Medicine and Public Health*, 3(6):1612–1616.

Naiara Pérez and Montse Cuadros. 2017. Multilingual CALL Framework for Automatic Language Exercise Generation from Free Text. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 49–52, Valencia, Spain. Association for Computational Linguistics.

Ildikó Pilán and Elena Volodina. 2014. Reusing Swedish FrameNet for training semantic roles. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1359–1363, Reykjavik, Iceland. European Language Resources Association (ELRA).

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 22–32.

Pei Quan, Yong Shi, Lingfeng Niu, Ying Liu, and Tianlin Zhang. 2018. Automatic Chinese multiple-choice question generation for human resource performance appraisal. *Procedia Computer Science*, 139:165–172.

Björn Rudzewitz, Ramon Ziai, Kordula De Kuthy, Verena Möller, Florian Nuxoll, and Detmar Meurers. 2018. Generating Feedback for English Foreign Language Exercises. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 127–136, New Orleans, Louisiana. Association for Computational Linguistics.

Keisuke Sakaguchi, Yuki Arase, and Mamoru Komachi. 2013. Discriminative Approach to Fill-in-the-Blank Quiz Generation for Language Learners. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–242.

Karim Shabani, Khatib Mohammad, and Saman Ebadi. 2010. Vygotsky's Zone of Proximal Development: Instructional Implications and Teachers' Professional Development. *English Language Teaching*, 3.

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

Simon Smith, P. V. S. Avinesh, and Adam Kilgarriff. 2010. Gap-fill Tests for Language Learners: Corpus-Driven Item Generation. In *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*, pages 1–6. Macmillan Publishers India.

Simon Smith, Adam Kilgarriff, Gong Wen-liang, Scott Sommers, and Wu Guang-zhong. 2009. Automatic Cloze Generation for English Proficiency Testing. In *Proceedings of the LTTC Conference*.

Nina Spada and Yasuyo Tomita. 2010. Interactions Between Type of Instruction and Type of Language Feature: A Meta-Analysis. *Language Learning*, 60(2):263–308.

Katherine Stasaski and Marti A. Hearst. 2017. Multiple Choice Question Generation Utilizing An Ontology. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 303–312.

Eiichiro Sumita, Fumiaki Sugaya, and Seiichi Yamamoto. 2005. Measuring Non-native Speakers' Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions. In *Proceedings of the second workshop on Building Educational Applications Using NLP*, pages 61–68.

Yuni Susanti, Ryu Iida, and Takenobu Tokunaga. 2015. Automatic Generation of English Vocabulary Tests. In *CSEDU (1)*, pages 77–87.

Yuni Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. 2018. Automatic distractor generation for multiple-choice English vocabulary questions. *Research and Practice in Technology Enhanced Learning*, 13(1):15.

Amos Tversky. 1964. On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology*, 1(2):386–391.

Elena Volodina, Ildikó Pilán, Lars Borin, and Therese Lindström Tiedemann. 2014. A flexible language learning platform based on language resources and web services. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3973–3978, Reykjavik, Iceland. European Language Resources Association (ELRA).

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing Multiple Choice Science Questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Masaru Yamada. 2019. Language learners and non-professional translators as users. In Minako O'Hagan, editor, *The Routledge handbook of translation and technology*, chapter 11, pages 183–199. Routledge.

Chak Yan Yeung, John Sie Yuen Lee, and Benjamin Ka-Yin T'sou. 2019. Difficulty-aware Distractor Generation for Gap-Fill Items. In *Australasian Language Technology Association Workshop*.

Torsten Zesch and Oren Melamud. 2014. Automatic Generation of Challenging Distractors Using Context-Sensitive Inference Rules. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148.

Branko Žitko, Slavomir Stankov, Marko Rosić, and Ani Grubišić. 2009. Dynamic test generation over ontology-based knowledge representation in authoring shell. *Expert Systems with Applications*, 36(4):8185–8196.

A Exercise types

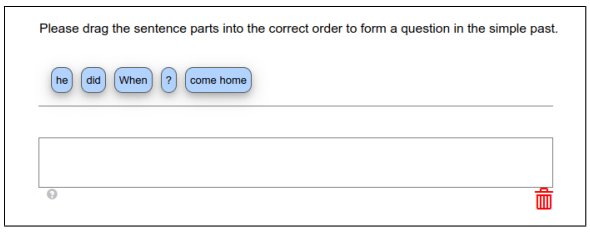

Figure 8: Jumbled Sentences

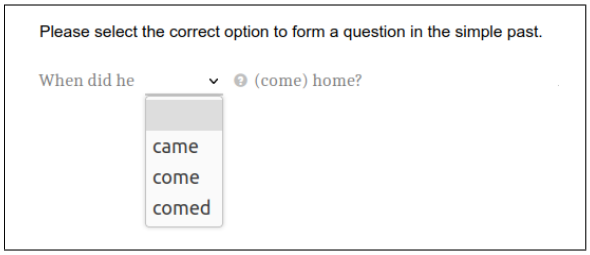

Figure 9: Single Choice

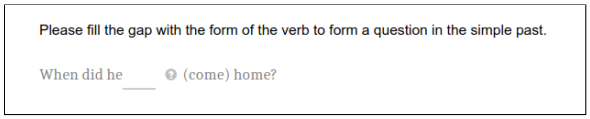

Figure 10: Fill-in-the-Blanks

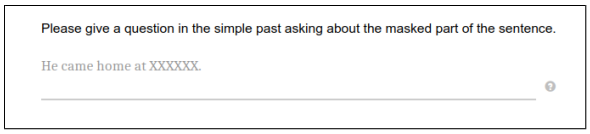

Figure 11: Short Answer