The instructional effectiveness of automatically generated exercises for learning French grammatical gender: preliminary results

Stephen Bodnar
Department of Linguistics, University of Tübingen, Germany
stephen.bodnar@uni-tuebingen.de

Abstract

Research into Automatic Exercise Generation (AEG) contributes new tools aimed at reducing the barrier to creating practice material, but few have been deployed in actual instruction with real learners. The present study extends previous work by employing AEG technology in instruction with L2 learners to evaluate its pedagogical effectiveness. Thirty-two second language learners of French were assigned to either a treatment condition, who practised with generated exercises, or a control condition that did no extra work. Both groups completed pre-, post-, and delayed post-tests. Participants in the treatment condition also completed questionnaires that elicited data on their in-practice emotions and the situations in which they arose. Our preliminary results suggest that AEG-based instruction can be pedagogically effective and support positive learning experiences, help to identify aspects of the instruction that could be improved, and suggest that a peer review mechanism could have an important role in future CALL platforms that use generated exercises.

1 Introduction

Despite the success of artificial intelligence in many aspects of our daily lives, sightings of Intelligent Computer Assisted Language Learning (ICALL) systems outside of the research lab remain rare. One barrier to their more wide-spread adoption is that equipping these systems with enough exercises for sustained practice is costly, requires a special skill set, and is beyond the scope of many projects. Fortunately, a growing body of research is investigating technology for Automatic Exercise Generation (AEG), which employs language technologies and linguistic resources to automatically generate practice materials.

The present study extends work in AEG by exploring the feasibility of using generated exercises with real learners. The learning context is an e-learning tool we have developed called COLLIE. While previous work tends to focus on English, COLLIE targets French grammatical gender, a linguistic target that learners find difficult (Lyster and Izquierdo, 2009). COLLIE scaffolds learning of gender-predictive noun suffixes with nine exercise types, including three spoken exercises, all of which can be generated automatically from arbitrary French texts, and an instruction sequence adapted from an effective human-led intervention.

We evaluated COLLIE by recruiting 32 French L2 learners from three North American universities. Half of the participants were assigned to a control condition, while another half completed an automated version of the instructional treatment in Lyster and Izquierdo's (2009) study adapted to online self-study and featuring only automatically generated exercises. In this paper we report on our findings showing positive learning outcomes from pre-test to delayed-posttest, suggesting that AEG can provide an effective context for learning a challenging element of French grammar. Self-reports from learners who practised with COLLIE report largely positive emotional experiences, and responses to open item questionnaires pinpoint sources of frustration, related to speech recognition and the instructional format, with the majority of negative experiences not attributable to the use of AEG.

2 Background

Research on tools aimed at reducing the burden of creating learning materials for ICALL systems has taken place since at least the early 2000s (e.g., Heift and Toole, 2002) and since then its value has continued to be recognised (e.g., Presson et al., 2013). Studies in the area aim at developing computational methods, often based on underlying natural language processing technology but not always (e.g., Malafeev, 2014), for automatically creating L2 practice exercises of different types.

An important aspect of research into AEG is

Stephen Bodnar. The instructional effectiveness of automatically generated exercises for learning French grammatical gender: preliminary results. *Proceedings of the 11th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2022)*. Linköping Electronic Conference Proceedings 190: 10–22.

evaluation. In our experience, the evaluations in the literature can be grouped into three main concerns:

- evaluations of the technology underlying the exercise generation (e.g., Heift and Toole, 2002; Chalvin et al., 2013; Freitas et al., 2013; Aldabe et al., 2006; Beinborn, 2016; Colin, 2020; Ferreira and Pereira Jr., 2018; Malafeev, 2015; Perez-Beltrachini et al., 2012; Zilio et al., 2018; Baptista et al., 2016; Zanetti et al., 2021);

- human expert judgments of exercise quality along different dimensions (e.g., Chinkina and Meurers, 2017; Chinkina et al., 2020; Burstein and Marcu, 2005; Antonsen et al., 2013; Chalvin et al., 2013; Pilán et al., 2017; Pilán, 2016; Slavuj and Prskalo, 2021; Malafeev, 2014; Freitas et al., 2013); and

- reports from actual tool use by students (e.g., Chinkina et al., 2020; Malafeev, 2014; Antonsen and Argese, 2018; Antonsen et al., 2013; see also Galvan et al., 2016) or instructors (Toole and Heift, 2002; Burstein and Marcu, 2005; Antonsen and Argese, 2018).

Most evaluations fall into the first two categories, and while the third type is certainly related to our interest in the readiness of AEG for deploying to real learning situations, none of the studies investigate the instructional effectiveness of generated exercises.

A crucial step in evaluating AEG is establishing that new algorithms can deliver real value to L2 language instructors and learners. To this end, the present study explores the extent to which practice with automatically generated exercises delivers effective L2 learning. In this context, it is clearly important to study how learners' proficiency on target linguistic features changes as a result of practice. Alongside L2 proficiency, learners' affective (i.e., emotion-related) experiences in the practice activities are also important to consider because of the potential for different emotional experiences to influence learning outcomes. The pathways between emotions and L2 proficiency development are theorised to be dynamic and bidirectional (Shao et al., 2020), and to interact with personal goals as well as the environment (Dörnyei, 2009), and so are very complex, but evidence for the important role of emotions in SLA is emerging: Teimouri et al. (2019) in their meta-analysis of SLA research on anxiety found strong support for a negative relationship between anxiety and L2 achievement, suggesting that feelings of "tension, apprehension, nervousness, and worry" (p. 2, as cited in Spielberger, 1983) hinder L2 learning at a macro level. In a longitudinal study of classroom learning, Saito et al. (2018) found evidence for the facilitating effects of positive emotions on practice behaviour and L2 development.

In the context of self-study CALL practice using AEG, anxiety is perhaps less likely to play an important role, but due to the uncertain readiness of the emerging technology, other negative emotions such as confusion, frustration or boredom could hinder learning. Similarly, in-practice feelings of enjoyment, interest, curiosity, or confidence could "facilitate holistic thinking and creative problem solving, broaden the scope of attention and cognition, ... and enhance intrinsic motivation and long-term efforts" (Shao et al., 2020, p. 8). Thus, affective experiences play an important role in L2 instruction and so are a valuable dimension of evaluating instructional effectiveness. For these reasons, in the present study we target the following two research questions:

- To what extent can instruction based on automatically generated practice exercises improve learners' L2 grammatical accuracy?

- To what extent does AEG-based instruction support positive learning experiences?

3 The present study

The study proceeded in three phases. In the design phase we searched the SLA literature for an instructional approach to provide a solid pedagogical basis for the to-be-generated exercises that at the same time appeared technically feasible to automate. The approach we identified is a practice sequence developed by Lyster (2016, 2018) with three types of activities: *noticing activities* expose learners to written and spoken language carefully chosen to draw their attention to L2 features that are difficult to learn; *awareness activities* stimulate learners to reflect on the patterns they see in the language; *output activities* prompt learners to test their hypotheses by producing written and spoken language and receiving feedback. A successful intervention study in the SLA literature

(Lyster and Izquierdo, 2009) provided a concrete example of the practice sequence and its exercises. The intervention guided learners to notice and use noun suffixes in French that predict grammatical gender (e.g., most nouns ending in *-ette* tend to be feminine), and we adopted it as a reference for the technology we would develop. In what follows, we refer to this study as the *original intervention*.

In the development phase, we developed the technology to automatically generate the 9 different exercises used in the original intervention. One or two could not be used because they were technically too challenging, and we replaced those with similar ones that were technically feasible. This included an exercise generation pipeline using NLP, various linguistic resources, and a learner model (see Section 4.1). To be able to collect data on the effectiveness of the exercises with real learners, we also developed user interfaces for the exercises, a Learning Management System (LMS) for researchers to carry out experiments, and a number of research instruments (see Section 4.2).

In the evaluation phase we arranged a new intervention modelled closely after the original intervention we identified in the design stage. Keeping our planned evaluation similar to the original intervention had two advantages: 1) we could be confident our instructional treatment had validity, and 2) the human-led study could serve as a gold standard against which we could compare learning outcomes of a second intervention that used automatically generated exercises (see Malafeev, 2014 and Chinkina et al., 2020, who use a similar approach of comparing ratings of exercises to a human gold standard).

4 Materials and Methods

4.1 Exercise generation pipeline

Among the existing methods for exercise generation (see Perez-Beltrachini et al., 2012 for a discussion of different methods), our approach has the most in common with the systems developed by Heift and Toole (2002) and Heck and Meurers (2022) as our pipeline relies on NLP tools to handle arbitrary documents as input (as opposed to being based on static corpora, e.g. Pilán et al., 2017, or automatically generated language, e.g. Perez-Beltrachini et al., 2012; Verweij, 2020). Our approach differs from (Heift and Toole, 2002) because the pipeline requires another NLP component, namely a dependency parser, and unlike the work by Heck and Meurers (2022), our pipeline does not accept HTML but is limited to plain text, as preserving the original look and feel of the document was not an important requirement to realise the instructional approach we selected.

The pipeline can be divided into two stages, an intake stage, and a generation stage (see Figure 1).

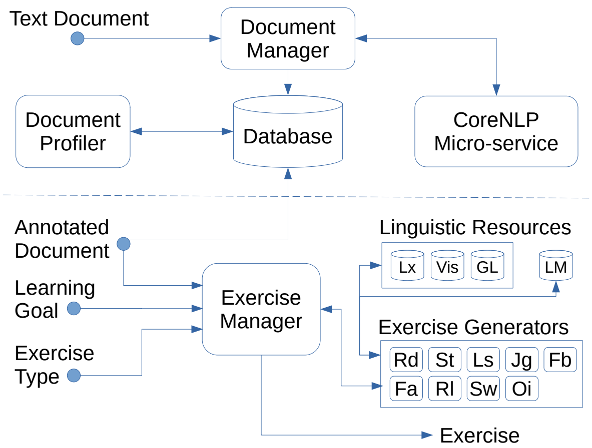

Figure 1: The exercise generation pipeline consists of an intake stage (top) and a generation stage (bottom). The system supports generating nine exercise types for practising French grammatical gender: Reading / Noticing (Rd); Sorting (St); Listing (Ls); Judging (Jg); Fill-in-the-Blanks targeting determiners (Fb) and determiners and adjectives (Fa); Riddles (Rl); Say the Word that Fits (Sw); and Object Identification (Oi). During generation the system makes use of three linguistic resources, Lexique (Lx); a database of readily visualisable words (Vis); and GLAWI (GL), as well as a Learner Model (LM).

In the intake stage, the pipeline stands ready for processing. Documents can be submitted to a Document Manager component by instructors or researchers through a web authoring tool; alternatively, larger numbers of documents can be batch processed using a script utility. Once received, the document is parsed into a structured data representation with Part of Speech tags and Dependency Parsing annotations. These annotations are obtained automatically using the Stanford CoreNLP toolkit (Manning et al., 2014), which we have implemented as a remote micro-service. Following parsing, the document and its corresponding parsed data structure is saved to a database for later retrieval. Immediately after being stored, a Document Profiler component analyses the document annotations to search for instances of language that are suitable for a given learning goal; for each supported learning goal, matching instances are counted, and the results of this document intake are also saved to the database as a cached profile to support efficient ranking of documents according to a learning goal of interest, or to discover which learning goals are supported by a particular document.

For the generation stage, we implement on-the-fly exercise generation. This means that after documents are fully processed no exercises are generated immediately. Instead, the system waits until there is a request from a user. Requests specify three arguments, 1) a document, 2) a learning goal and 3) an exercise type. These arguments are received by an Exercise Manager component, which orchestrates the generation of the exercise. First, the manager looks in the database to see if an exercise for this triplet has already been generated, and if found it is returned. If not, the manager searches in its registry of exercise generators. Exercise Generators in the system are specialized components that know how to build exactly one exercise type for exactly one learning goal. At pipeline initialisation time, the system searches through its codebase and registers all components implementing an Exercise Transformer interface. Then, when a request to generate an exercise arrives, the Exercise Manager searches this registry to see if it has a suitable generator and if so, delegates the exercise generation request.

Exercise generators all have in common that they iterate over the lines of a document to perform checks on each line for pedagogical relevance. These checks involve searching for dependency parse relations as well as additional linguistic criteria. Each line is processed differently depending on the results of the checks: if the generator determines that a line contains an instance of the learning goal (e.g., French grammatical gender), the line is transformed to a data structure with a format dictated by the particular exercise type being generated (e.g., a fill-in-the-blank). Otherwise, the generator either includes the raw text as is for meaningful context, as with the Reading exercise generator, or discards the line, as in the Riddle exercise. Depending on the exercise, generators employ additional linguistic resources for different purposes:

- Checking for adherence to predicted grammatical gender: When a generator encounters a target word with a gender-predictive suffix (e.g., words ending in *-ette* are nearly always feminine), it must verify that the gender predicted by the suffix is indeed the word's actual gender (there can be exceptions, such as *squelette* "skeleton" which has masculine grammatical gender). The pipeline integrates a Lexique database (New et al., 2004) to look up the gender of words with predictive suffixes and avoid words that are exceptions.

- Identification of readily visualisable words: In the case of Object Identification exercises, the generator must determine if a word with a gender-predictive suffix can be easily depicted in an image. For this, we developed an in-house resource that distinguishes between words that can be visualised easily (e.g., *une éruption* "an eruption") or with difficulty (*e.g., une abstraction "an abstraction"*). The resource draws on three English-language databases in psycholinguistics containing ratings for words related to their ease of visualisation (Wilson, 1988; Brysbaert et al., 2013; Scott et al., 2018). In these partially overlapping databases, each word has a score related to how readily it can be visualised. As an approximation we first automatically translated headwords to French and then combined available ratings from the different databases by taking their mean. Finally, we set a threshold by experimenting with different values and choosing the lowest value that did not return unsuitable words. [1]

- Clue creation for Riddle exercises: Our approach to clue creation is straightforward. We integrate a linguistic resource called GLAWI (Hathout and Sajous, 2016), which is a lexical database containing definitions (among other information) derived from Wiktionnaire. The riddle generator obtains clues for a target word by loading the relevant definitions from GLAWI. Because the target word can sometimes appear in the returned definitions, in a second step we replace all occurrences of the target with underscore characters to ensure the riddle is not too easy.[2]

[1] We also manually reviewed the images to mark content that was inappropriate for an educational context (e.g., nudity, violence) but a detailed presentation of this is beyond the scope of this paper.

[2] For a more creative approach to clue generation for those working with English as a target language, see Galvan et al., 2016

- Selecting previously unseen words: an additional resource used by some generators is the system's Learner Model. The learner model tracks which nouns a learner has seen and how often. For example, in Reading exercises, when a target word appears on the page, the word's appearance is logged and registered with the learner model. This data is then an important resource during the generation of the Judgment exercise, which aims to help the learner generalise their knowledge from words they have already seen to words with the same gender predictive suffixes that they have not yet encountered.

Currently the exercise pipeline supports generation for French grammatical gender with nine exercises (see Figures 2 - 4 below for examples), and support for more languages is planned for the future (e.g., grammatical gender for Dutch and German). The pipeline is implemented in Java, and is designed to be modular so that it can be integrated into any back-end web application based on the Java virtual machine.

4.2 COLLIE e-learning platform

To support our evaluation of AEG, we developed a web-based e-learning platform that learners could use and into which the exercise pipeline described above could be embedded. The platform we have developed is called COLLIE, an abbreviation of *Counter-balanced Language Learning & Instruction made Easier*. The name and platform draws inspiration from Lyster's (2007) approach to balancing meaning-focused classroom learning, which can fall short of pushing learners to become fully accurate speakers, with accuracy-focused practice where they are pushed to notice L2 features that are difficult to learn, reflect on and apply patterns in the L2, and practice producing written and spoken language. The vision for COLLIE is to make it easier for teachers to supplement their classroom-based activities, which are usually about communicating, with accuracy-focused exercises that students can practice on their own time, using content related to their classroom activities.

In its current implementation, COLLIE supports written and spoken practice with immediate feedback. Scaffolded feedback is feasible because of the system's closed exercise design and narrow focus on grammatical gender, though expanding to support other learning goals or more open exercises would require a more sophisticated feedback mechanism (c.f. Rudzewitz et al., 2018). To support feedback in spoken exercises, we rely on a commercial Automatic Speech Recognition (ASR) service provided by Google Cloud (Google Cloud, nd) for transcribing recordings before they are processed by the system's feedback module. The recognition model used by the system is for European French (language tag 'fr-FR'), and we use a mechanism offered by the service to provide a set of hints consisting of all possible combinations of French singular definite and indefinite determiners and a target noun (e.g. *le squelette* | *la squelette* | *un squelette* | *une squelette*). This setting helps to guide the ASR towards transcriptions that are most likely for a given practice item.

As a web application, COLLIE consists of a back-end web server based on the Grails framework and a front-end set of user interfaces that communicate using HTTP requests. The front-end interface generates requests, which are received at specific URL endpoints by the back-end for processing by different application modules. The modules are implemented in Java and Groovy as object-oriented classes and deliver core functionalities from the AEG and LMS domains related to entities such as Document, Exercise, User, LogEvent, SpeechRecording and so on. All modules have accompanying unit tests to support refactoring.

Results from back-end processing are serialised to JSON and returned to the front-end for rendering. All user interfaces for the platform are implemented using the React.js single page web application framework. User interface elements are modular and paremeterised into reusable components (e.g., a VoiceRecorder for audio recording).

4.3 Instruments

To measure changes in French grammatical accuracy, we adopted three proficiency tests used in the original intervention, two oral production measures and a binary-choice test. As annotations for the oral production recordings are not yet complete, in this paper we present the binary-choice test results as a preliminary indicator of instructional effectiveness. Participants completed the test on the COLLIE platform. They viewed 80 different items featuring words with gender-predictive suffixes one at a time (see Figure 5).

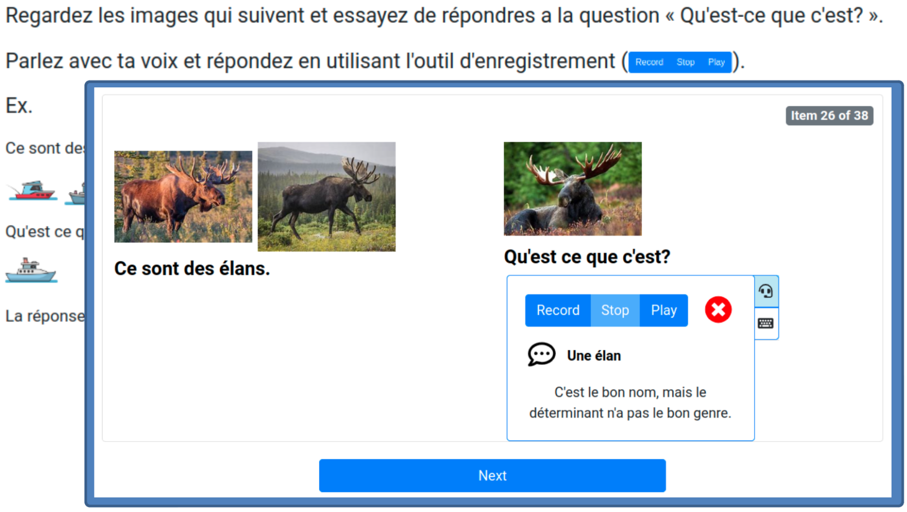

Figure 2: In the *Object Identification* exercise, learners must name the object they see on the right while using the correct determiner. This is a spoken exercise, where student's answers are first transcribed using speech recognition, and then evaluated for correctness.

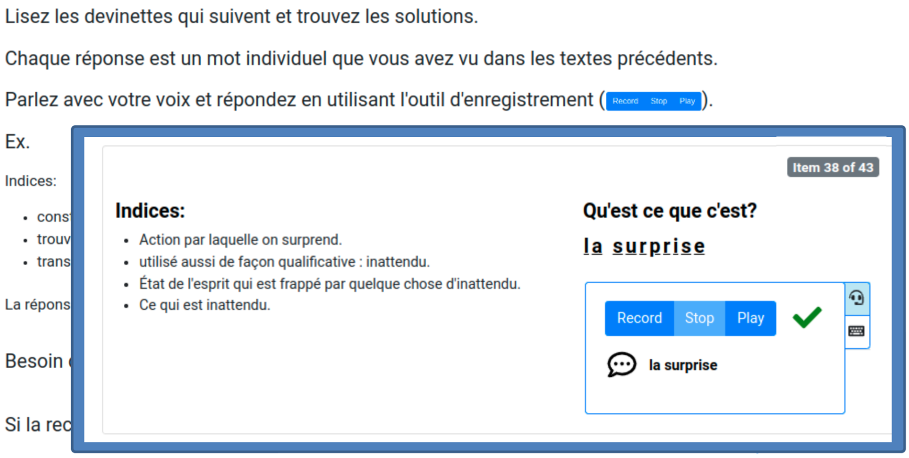

Figure 3: In *Riddle* exercises, clues appear on the left in bullet points, which learners must read to guess the determiner and noun on the right (in blanks). Clues are selected automatically, and because their helpfulness can vary, there is a hint button that if pressed reveals approximately half of the letters for the noun.

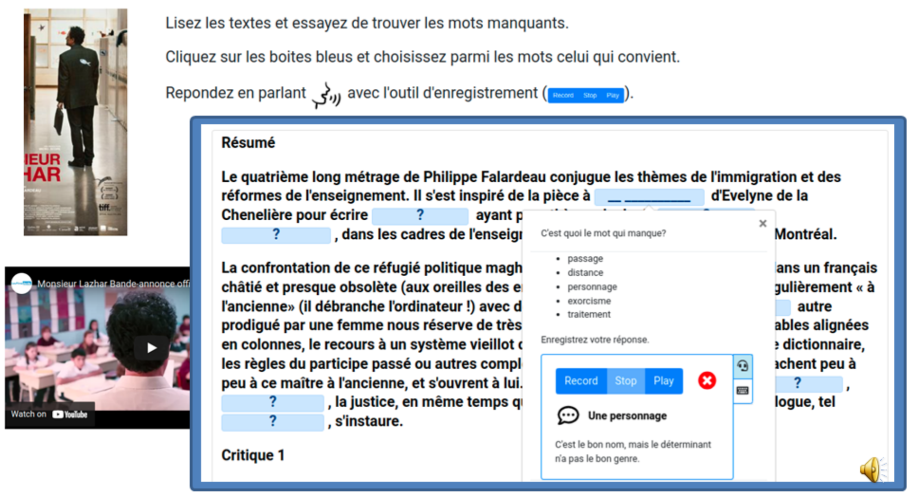

Figure 4: The *Say the Word that Fits* exercise asks students to fill blanks in a document by speaking the correct determiner and noun combination. In the cases where speech recognition does not work accurately, which can be the case sometimes for some participants, there is a keyboard icon they can press to type their answer.

Presented with each word were buttons they could click to choose between a masculine and feminine determiner. Participants received instructions and completed a short practice test before starting the actual test.

Figure 5: An example item from the binary-choice test.

Along with the proficiency measures, we also employed a questionnaire to measure affective experiences related to the practice exercises. The instrument we selected is based on the well-known Achievement Emotions Questionnaire (Pekrun et al., 2002) which is now gaining attention in SLA research (e.g., Shao et al., 2019). In our adapted version of the questionnaire (see Figure 6), participants reported how frequently they felt an emotion (Diener et al., 2009) in response to the following prompt:

> Please think about what you have been doing and experiencing during today's grammar exercises. Then report how often you experienced each of the following feelings, using the scale below. For each item, select a number from 1 to 5, and indicate that number with a mouse-click.

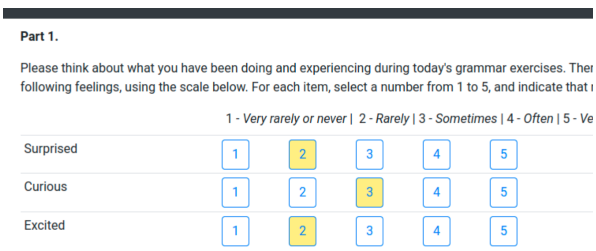

Figure 6: A close-up of the questionnaire used to elicit data on participants' emotions during practice.

The questionnaire included 7 positive valence emotions and 7 negative valence emotions, where *valence* refers to whether an emotion is positive, like feeling interested or curious, or negative, like feeling bored or confused. Participants responded using a 5-point scale, from very rarely (1), to rarely (2), to sometimes (3), to often (4), or very often and always (5).

Along with these quantitative items, the questionnaire also included open-ended items to prompt learners to describe in their own words the situations in which they felt the emotions they reported.

4.4 Data collection

For the evaluation of COLLIE and its AEG technology, we arranged an intervention closely modelled after the original intervention by Lyster and Izquierdo (2009). In Fall of 2021 we recruited 32 participants from three North American universities. The participants were intermediate-level learners of French and were actively attending a French course at the time.

The entire data collection took place over 9 weeks (see Figure 7). At week 1 participants completed the pretest. Immediately following the pretest they were assigned to a treatment or control condition. The mechanism used for assignment was an anticlustering algorithm available in R (Papenberg and Klau, 2021) which distributed participants between the two conditions based on their pretest scores in order to ensure the two conditions were balanced at the outset. Over the next three weeks participants in the treatment condition completed three practice sessions, once per week, and following practice an exit survey on approximately the fourth week. Both groups completed a post-test on the sixth week of the study, and a delayed post-test on the ninth and final week.

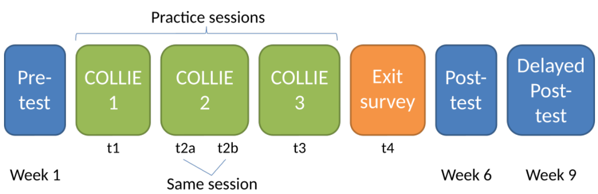

Figure 7: Data collection took place over 9 weeks.

The L2 instruction offered by COLLIE was online self-study and proceeded as follows. Each week participants received an invitation to register for a time slot to practice. Each time participants logged in, they were shown a home screen with a simple list of tasks for them to complete in the session. Tasks appeared as links that participants could click to launch a practice exercise. After participants completed a task, they were returned to their home screen and the completed task appeared with a line through it to indicate it had been completed and to help build a sense of progress. As participants worked on the exercises, their work was automatically stored and saved on the back-end so that in the event of a break or accidental page refresh their work was preserved. At the end of each session participants completed the learning experience questionnaire, after which a message appeared informing them they have completed their session and that they could safely log out.

Participants completed the learning experience questionnaire on five occasions: once at the end of session 1 (t1), two times in session 2 (t2a and t2b), as this was a longer session and we wanted to check the emotions halfway through the session and again at the end of the session; and again at the end in session 3 (t3). Finally, we also included the emotion questionnaire in the exit survey approximately 1 week after practice (t4), to see what kind of emotional experiences the participants would report after having not practised for some time.

4.5 Analysis

In our analysis of learning outcomes, we have the independent variable *condition* (either treatment or control) and the dependent variable *score*, which in this preliminary analysis is the raw score from the binary-choice tests. Participants completed pre-, post- and delayed post-tests. To investigate how test scores changed over time, our analysis compares scores for the two groups using a two-way repeated measures ANOVA with a 2 (*condition*: treatment or control) x 3 (*test*: pre, post or delayed post) design.

For our analysis of participants' learning experiences we take a slightly different approach. Rather than look at the frequency of the individual 14 emotions sampled by the questionnaire, we adopt a more coarse-grained view that compares the frequency of positive versus negative emotions. The independent variable is *valence* (either positive or negative), and the dependent variable is the self-reported *frequency*. Treatment condition participants completed the questionnaire five times, yielding a two-way repeated measures ANOVA with a 2 (*valence*: positive or negative) x 5 (*time*: t1, t2a, t2b, t3, t4) design.

The aim with eliciting information about particular learning situations was to gain insight into what the context or cause was for the emotions participants reported. We reviewed the open item responses and coded them with short one or two-word labels, for example *system error* or *exercise repetitiveness*. We then went over the labels and distinguished between situations resulting from the use of AEG technology and other causes. In the present study we focus on situations related to negative emotions, to detect any negative effects of using generated exercises.

5 Results

5.1 Learning gains

Our analysis of learning outcomes returned a main effect for test, $F(1.78, 53.42) = 10.82$, $MSE = 19.20$, $p < .001$, $\hat{\eta}^2_p = .265$, suggesting that the scores change from pretest to delayed posttest. There was no main effect of condition, $F(1, 30) = 3.60$, $MSE = 270.04$, $p = .068$, $\hat{\eta}^2_p = .107$. Also returned is a test by condition effect, $F(1.78, 53.42) = 14.38$, $MSE = 19.20$, $p < .001$, $\hat{\eta}^2_p = .324$.

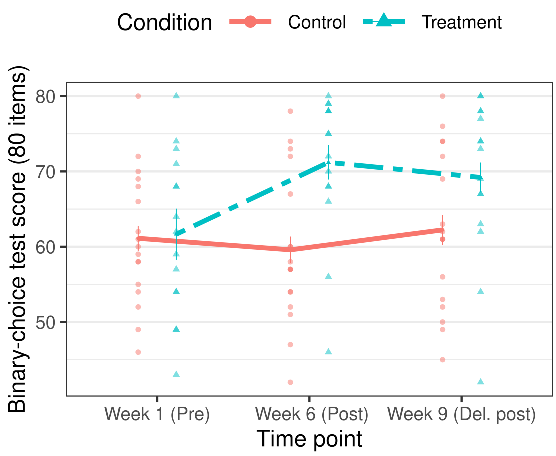

Figure 8: Participant binary-choice test scores on French grammatical gender at Week 1 (pre-test), Week 6 (post-test) and Week 9 (delayed post-test). The treatment group improves with time, while control group remains stable.

Post-hoc analysis of the test by condition effect (Sidak) points to important changes for the treatment group, returning a significant difference between pretest (M = 61.7) and posttest (M = 71.2), and no difference between posttest and delayed posttest (M = 69.2). This suggests the intervention helped the treatment group improve, and that this improvement had not faded three weeks later. For the control group, there were no significant differences, and their scores remained equivalent from pretest (M = 61.1) to posttest (M = 59.6) to delayed posttest (M = 62.2), suggesting that the control group remained stable. Together, these findings point to the pedagogical effectiveness of practising with the system.

5.2 Learning experience

Analysis of learning experience returned a main effect of valence, $F(1, 14) = 20.66$, $MSE = 2.14$, $p < .001$, $\hat{\eta}^2_p = .596$, suggesting that positive emotions were experienced more frequently than negative ones. There was no main effect for time, $F(3.17, 44.43) = 2.18$, $MSE = 0.16$, $p = .101$, $\hat{\eta}^2_p = .134$. We also observed a valence by time effect, $F(3.13, 43.85) = 15.94$, $MSE = 0.17$, $p < .001$, $\hat{\eta}^2_p = .532$, suggesting that positive and negative emotions had frequencies that changed differently in the practice sessions.

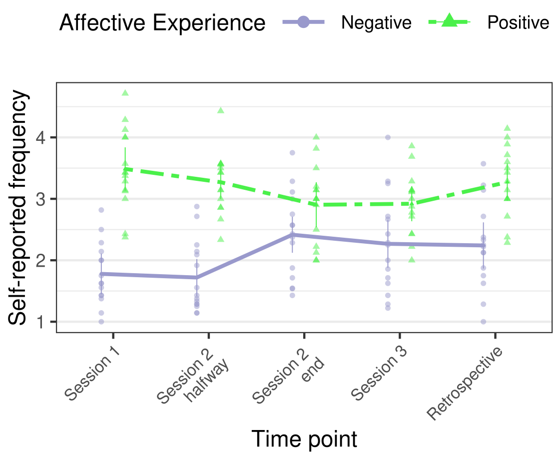

Figure 9: Self-reported frequency of positive vs. negative experiences over four weeks. Positive emotions (in green) follow a U-shape curve, while for negative emotions (in blue) there is a modest increase.

Post-hoc analysis indicates that positive emotions follow a U-shape curve; they start high (M = 3.49), but then drop significantly at the end of session 2 (M = 2.9) and do not change significantly in session 3, (M = 2.92). During this period positive and negative emotions occur equally frequently. At time point 4 there is again a significant increase (M = 3.27), when students had had a break from practice and were looking back. For the negative emotions, there is a modest but significant increase from the start to the end of the second session (M = 1.4 vs. M = 2.8).

5.3 Learning situations

In total we observed 169 instances of situations described in the open questionnaire data that were related to negative emotions, from which 37 unique categories emerged. Table 1 presents a subset of the most frequently occurring situations associated with negative emotions during practice, together with less common situations that are interesting because they can be attributed to the use of AEG technology for creating the practice materials.

From the entries in the table, we see that there are some clear links between negative emotions and certain situations. Participants reported feeling frustrated, discouraged or confused when the ASR failed to accurately transcribe their speech. Apparently the length of the exercises was sometimes too long, and this resulted in participants feeling bored or frustrated. In some cases participants appeared to have difficulty with learning the gender-predictive suffix patterns, despite the special instructional sequence, and this led to frustration and confusion.

Interestingly, there appear to have been relatively few situations directly related to the use of AEG technology, but from a pedagogical point of view those that we observed seem important and worth sharing here. First, a number of participants reported being unable to answer an item correctly even when they tried all possible answers. This occurred in a *Say the Word that Fits* exercise (see Figure 4), where the exercise generation pipeline created an item that had no target answer. The problem seems to have occurred due to a dependency parsing error that incorrectly assigned a determiner relation to the text *un peu* (a little). This caused the system to then look up the gender of *peu* in Lexique which failed and then resulted in no target answer being specified for the item. When students came across this and tried all possible combinations of determiner and noun without managing to have their answer accepted by the system they understandably reported feeling frustrated or confused.

A second AEG-related error occurred in the generation of Object Identification exercises (see Figure 2) in which participants reported being con-

Situation	Example	% Responses (169)	Related emotions (desc)	Related to AEG
Inaccurate speech recognition	"it took a really long time for the computer to recognize my voice on certain exercises"	13.6	frustration, discouragement, confusion	No
Length of exercises	"the exercises feel too long"	13.6	boredom, frustration	No
Learning challenges	"I was frustrated because I couldn't exactly figure out the pattern"	8.9	frustration, confusion	No
Unanswerable questions	"even when I copy pasted the answer in, it would not accept it"	8.9	frustration, confusion	Yes
System errors	"when I put un or une , it said something's not right"	7.7	frustration, confusion, discouragement	No
Repetitive exercises	"There was no variation in the format"	7.1	boredom	No
Unsuitable images	"some of the images which were meant to display singular [objects] showed multiple"	0.6	frustration	Yes
Inappropriate riddle clues	"definitions for the devinettes are sometimes very unhelpful ... and sometimes downright offensive"	0.6	frustration	Yes

Table 1: Example situations from open-items, showing negative emotions only.

fused by some of the images that appeared. To elicit the singular form of a target noun with its determiner the exercise requires that the right-most image shows a single instance of an object. The images used in this exercise were automatically downloaded and it appears in some cases the right-most image actually contained multiple instances. This led to confusion about whether the system expected an answer in singular or plural form.

A final issue occurred during the generation of a Riddle exercise (see Figure 3). As described above, clues for the riddles were created automatically by retrieving definitions from a lexical database called GLAWI (Hathout and Sajous, 2016), with content derived from Wiktionnaire. During the creation of a riddle for the target *une baleine* (a whale), the system regrettably included a colloquial and offensive definition from the database, which a small number of participants rightfully found unpleasant and frustrating.

6 Discussion

Tools that help to quickly author practice material for ICALL systems have the potential to help increase their impact in L2 instruction. Research into technology for AEG has demonstrated the feasibility of generating a variety of exercise types, and human experts tend to judge the output of these tools favourably, yet there has so far been relatively little research evaluating the effectiveness of L2 instruction with generated exercises. The present study aimed to address this gap by developing an exercise generation pipeline and e-learning platform targeting French grammatical gender with pedagogy informed by SLA research. Our evaluation of the platform investigated two dimensions of instructional effectiveness: 1) learning outcomes and 2) affective learning experiences. With regard to learning, our preliminary analysis of the binary-choice test scores showed that participants who completed the instruction improved significantly in comparison to a control group, suggesting that AEG can be an effective instructional tool. In the original intervention, Lyster and Izquierdo (2009) found that scores on two oral proficiency measures followed the same pattern as the binary-choice data. Currently we are working on completing annotations of the speech recordings from our own oral production measures, but we are optimistic that an analysis of the data will also show improvements and provide additional evidence for the effectiveness of practice with AEG.

Our analysis of participants' in-practice affective experiences indicates that positive emotions were experienced more frequently than negative ones, which is an encouraging finding. At the same time, we found that the frequency of positive and negative experiences changed over time, with positive emotions following a U-shaped curve in which they occurred less frequently in later sessions. These findings appear to indicate that the instruction delivered by COLLIE is on the right track but also that there is still room for improvement.

In this regard, the descriptions that participants provided of situations in which they experienced negative emotions are an important source of information for improving the instruction. It is somehow encouraging to find that the majority of negative experiences seem to result from situations unrelated to the use of AEG technology, being attributable instead to issues with the instructional format, such as repetitive or overly long exercises, which can be addressed relatively easily. Improving the accuracy of the speech recognition is also an important issue which can potentially be addressed by using an ASR engine trained on non-native speech (van Doremalen, 2014), though this is a much larger undertaking.

Although we observed relatively few negative experiences directly attributable to AEG, the three types of situations that did occur clearly will have a negative impact on learning and should be addressed. The issue with the inappropriate riddle clue is particularly concerning because it left at least one individual feeling uncomfortable for the rest of the practice session.

In the present study, only the images used in the generation of Object Identification exercises were reviewed to ensure their appropriateness for instruction, but the issue above suggests that human review of automatically generated content has a more important role in AEG than we initially anticipated, and which without assessing affective dimensions of learning might have gone undetected.

A recommendation, based on the study here, is for future work looking at exercise generation in the context of a CALL platform to consider exploring the idea of a peer review mechanism that encourages users to share the exercises they generate and to review each other's exercises. One can imagine a learning platform that shows the number of reviews for generated exercises, and possibly makes use of badges to clearly mark exercises that have been vetted by the community, to avoid some of the negative experiences that we saw here.

7 Conclusions and Future work

In conclusion, this study suggests that it is feasible to use automatic exercise generation to more easily create L2 practice exercises that are pedagogically effective and support positive learning experiences. At the same time, the affective data suggest that there is room for improvements to the instruction, and that a peer review mechanism could be an important feature of future CALL systems with AEG pipelines, to ensure more positive learning experiences.

In order to draw stronger conclusions about the efficacy of AEG, there are some limitations that need to be addressed. First, the current findings are based on just one proficiency measure, the binary-choice test. However, during the data collection we also gathered data from two oral production measures that, once annotated, could provide additional support. A second point would be to recruit additional annotators to analyse and label the open-item questionnaire data for a more robust qualitative analysis. Third, an interesting point to follow up on would be to compare the learning outcomes of the present study with those found in the original human-led intervention (Lyster and Izquierdo, 2009).

Finally, the present study focused on a single aspect of learning a foreign language, grammatical gender, over a relatively short time (three practice sessions). To obtain more support for the instructional effectiveness of AEG-based instruction in general, it would be interesting to carry out an evaluation with a system that supports a variety of linguistic targets, such as the system developed by Heck and Meurers (2022), over a longer period of time and with more participants.

Acknowledgements

Parts of this study were presented at the 2022 EUROCALL conference. I am grateful to two anonymous NLP4CALL reviewers for their helpful comments. I would like to thank Roy Lyster, who was my postdoctoral supervisor for this project, as well as the instructors who helped with recruiting participants. I would also like to thank Detmar Meurers for his support of my continuation of this project. This research was made possible by a grant from the *Fonds de recherche du Québec – Société et culture* (258852).

References

Aldabe, I., de Lacalle, M. L., Maritxalar, M., Martinez, E., and Uria, L. (2006). ArikIturri: An Automatic Question Generator Based on Corpora and NLP Techniques. In Ikeda, M., Ashley, K. D., and Chan, T.-W., editors, *Intelligent Tutoring Systems*, volume 4053 of *Lecture Notes in Computer Science*, pages 584–594. Springer.

Antonsen, L. and Argese, C. (2018). Using authentic texts for grammar exercises for a minority language. In *Linköping Electronic Conference Proceedings*, page 152.

Antonsen, L., Johnson, R., Trosterud, T., and Uibo, H. (2013). Generating Modular Grammar Exercises with Finite-State Transducers. In *Proceedings of the second workshop on NLP for computer-assisted language learning at NODALIDA*.

Baptista, J., Lourenco, S., and Mamede, N. J. (2016). Automatic generation of exercises on passive transformation in Portuguese. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 4965–4972.

Beinborn, L. M. (2016). *Predicting and Manipulating the Difficulty of Text-Completion Exercises for Language Learning*. PhD thesis, Technische Universität Darmstadt.

Brysbaert, M., Warriner, A. B., and Kuperman, V. (2013). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Burstein, J. and Marcu, D. (2005). Translation exercise assistant: automated generation of translation exercises for native-Arabic speakers learning English. In *HLT-Demo '05: Proceedings of HLT/EMNLP on Interactive Demonstrations*, page 16–17.

Chalvin, A., Eensoo, E., and Stuck, F. (2013). Mining a parallel corpus for automatic generation of Estonian grammar exercises. In *Third biennial conference on electronic lexicography (eLex 2013) "Electronic lexicography in the 21st century: thinking outside the paper"*, pages 280–295, Tallinn, Estonia.

Chinkina, M. and Meurers, D. (2017). Question Generation for Language Learning: From ensuring texts are read to supporting learning. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 334–344, Copenhagen, Denmark. Association for Computational Linguistics.

Chinkina, M., Ruiz, S., and Meurers, D. (2020). Crowdsourcing evaluation of the quality of automatically generated questions for supporting computer-assisted language teaching. *ReCALL*, 32(2):145–161.

Colin, É. (2020). *Traitement automatique des langues et génération automatique d'exercices de grammaire*. Theses, Université de Lorraine.

Diener, E., Sandvik, E., and Pavot, W. (2009). *Assessing Well-Being. Social Indicators Research Series*, chapter Happiness is the Frequency, Not the Intensity, of Positive Versus Negative Affect, pages 213–231. Springer, Dordrecht.

Dörnyei, Z. (2009). The L2 Motivational Self System. In Dornyei, Z. and Ushioda, E., editors, *Motivation, Language Identity and the L2 Self*, pages 9–42. Multilingual Matters.

Ferreira, K. and Pereira Jr., A. R. (2018). Verb Tense Classification And Automatic Exercise Generation. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, WebMedia '18, page 105–108, New York, NY, USA. Association for Computing Machinery.

Freitas, T., Baptista, J., and Mamede, N. J. (2013). Syntactic REAP.PT: Exercises on Clitic Pronouning. In *Proceedings of the 2nd International Symposium on Languages, Applications and Technologies (SLATE 2013)*, Porto, Portugal.

Galvan, P., Francisco, V., Hervás, R., and Méndez, G. (2016). Riddle Generation using Word Associations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia*.

Google Cloud (n.d.). Google Cloud - Speech-to-Text. https://cloud.google.com/speech-to-text. Retrieved 10 October 2022.

Hathout, N. and Sajous, F. (2016). Wiktionnaire's Wikicode GLAWIfied: a Workable French Machine-Readable Dictionary. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

Heck, T. and Meurers, D. (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.

Heift, T. and Toole, J. (2002). Task Generator: A Portable System for Generating Learning Tasks for Intelligent Language Tutoring Systems. In Barker, P. and Rebelsky, S., editors, *Proceedings of ED-MEDIA 2002–World Conference on Educational Multimedia, Hypermedia & Telecommunications*, pages p. 1972–1977. Denver, Colorado, USA.

Lyster, R. (2007). *Learning and teaching languages through content: A counterbalanced approach*. Amsterdam: John Benjamins.

Lyster, R. (2016). *Vers une approche intégrée en immersion*. Montréal : Les Éditions CEC.

Lyster, R. (2018). *Content-based language teaching*. New York: Routledge.

Lyster, R. and Izquierdo, J. (2009). Prompts Versus Recasts in Dyadic Interaction. *Language Learning*, 59(2):453–498.

Malafeev, A. (2014). Language Exercise Generation: Emulating Cambridge Open Cloze. *Int. J. Concept. Struct. Smart Appl.*, 2(2):20–35.

Malafeev, A. (2015). Exercise Maker: Automatic Language Exercise Generation. In *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"*, volume 14 of *21*, pages 441–452, Moscow.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

New, B., Pallier, C., Brysbaert, M., and Ferrand, L. (2004). Lexique 2 : A new French lexical database. *Behavior Research Methods, Instruments, &amp Computers*, 36(3):516–524.

Papenberg, M. and Klau, G. W. (2021). Using anticlustering to partition data sets into equivalent parts. *Psychological Methods*, 26(2):161–174.

Pekrun, R., Goetz, T., Titz, W., and Perry, R. (2002). Academic Emotions in Students' Self-Regulated Learning and Achievement: A Program of Qualitative and Quantitative Research. *Educational Psychologist*, 37(2):91–105.

Perez-Beltrachini, L., Gardent, C., and Kruszewski, G. (2012). Generating Grammar Exercises. In *NAACL-HLT 7th Workshop on Innovative Use of NLP for Building Educational Applications*, Montreal, Canada.

Pilán, I. (2016). Detecting Context Dependence in Exercise Item Candidates Selected from Corpora. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–161, San Diego, CA. Association for Computational Linguistics.

Pilán, I., Volodina, E., and Borin, L. (2017). Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation. *Traitement Automatique des Langues (TAL)*, 57(3):67–91.

Presson, N., Davy, C., and MacWhinney, B. (2013). *Innovative Research and Practices in Second Language Acquisition and Bilingualism*, chapter Experimentalized CALL for adult second language learners., pages 139–164. John Benjamins Publishing Company.

Rudzewitz, B., Ziai, R., De Kuthy, K., Möller, V., Nuxoll, F., and Meurers, D. (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.

Saito, K., Dewaele, J.-M., Abe, M., and In'nami, Y. (2018). Motivation, Emotion, Learning Experience, and Second Language Comprehensibility Development in Classroom Settings: A Cross-Sectional and Longitudinal Study. *Language Learning*, 68(3):709–743.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., and Sereno, S. C. (2018). The Glasgow Norms: Ratings of 5, 500 words on nine scales. *Behavior Research Methods*, 51(3):1258–1270.

Shao, K., Nicholson, L. J., Kutuk, G., and Lei, F. (2020). Emotions and Instructed Language Learning: Proposing a Second Language Emotions and Positive Psychology Model. *Frontiers in Psychology*, 11.

Shao, K., Pekrun, R., and Nicholson, L. J. (2019). Emotions in classroom language learning: What can we learn from achievement emotion research? *System*, 86:102121.

Slavuj, V. and Prskalo, L. N.and Bakaric, M. B. (2021). Automatic generation of language exercises based on a universal methodology: An analysis of possibilities. *Bulletin of the Transilvania University of Brasov. Series IV: Philology and Cultural Studies*, 16(43)(2):29–48.

Teimouri, Y., Goetze, J., and Plonsky, L. (2019). Second language anxiety and achievement: A meta-analysis. *Studies in Second Language Acquisition*, 41(2):363–387.

Toole, J. and Heift, T. (2002). The Tutor Assistant: An Authoring Tool for an Intelligent Language Tutoring System. *Computer Assisted Language Learning*, 15:373–386.

van Doremalen, J. (2014). *Developing automatic speech recognition-enabled language learning applications: from theory to practice*. PhD thesis, Radboud Universiteit Nijmegen.

Verweij, R. (2020). Automated Exercise Generation in Mobile Language Learning. Online: https://digitalcommons.bard.edu/senproj_s2020/297/.

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, &amp Computers*, 20(1):6–10.

Zanetti, A., Volodina, E., and Graën, J. (2021). Automatic Generation of Exercises for Second Language Learning from Parallel Corpus Data. *International Journal of TESOL Studies (2021)*, 3(2):55–70.

Zilio, L., Wilkens, R., and Fairon, C. (2018). SMILLE for Portuguese: Annotation and Analysis of Grammatical Structures in a Pedagogical Context. In *International Conference on Computational Processing of the Portuguese Language*.