Auto-generated language learning online courses using generative AI models like ChatGPT

Sylvio Rüdian [1] and Niels Pinkwart [2]

Abstract: Generating online courses is always a trade-off between possibilities, technical limitations, and quality. State-of-the-art generative models can assist teachers in the creation process. However, generating learning materials is highly complex. Hence, teachers mainly create them manually. In this paper, learning content for a concrete micro-learning template is generated focusing on the field of language teaching. It intends that learners can find correct responses by logical thinking. Teachers provide a topic as input. Then, the approach asks for the required information using GPT3.5 with instructional prompts and combines responses to form a language learning unit. The quality of the resulting learning content, focusing on correctness, and appropriateness, is evaluated and discussed to examine the practicability of the tool, and alternatives are given.

Keywords: Language learning, generative AI models, auto-generated course units.

1 Introduction

Language learning apps to learn foreign languages become popular within the last decades. Course suppliers operate worldwide to teach languages. Duolingo, Rosetta Stone, Babbel, Busuu, or Lingoda are just a few examples of well-known brands. The main success of such brands teaching languages is sourced in the scalability to create course materials and the desire to learn languages by a large target group. In language teaching, learning content is highly structured. It consists of vocabulary, translations, grammar, dialogs, and more. The core competencies are pre-defined as language proficiency levels [Co21]. For each level, certain skills are defined, including vocabulary sets. Based on that source, language learning courses can be designed by teachers.

Recently, generative AI models become prominent. Models like GPT3.5, better known as ChatGPT [Ko23], allow teachers to generate texts, word lists with semantic relations, or even tasks. From the first view, such generative models are a great base to create rich teaching content. At the current state, those outputs are text-only, but responses of such models can be used to create texts as a base for learning materials. Further, technologies to create online courses become highly flexible. Exemplarily, H5P allows teachers to

[1] Humboldt-Universität zu Berlin, Department of Computer Science, Berlin, Germany, ruediasy@informatik.hu-berlin.de, https://orcid.org/0000-0003-3943-4802

[2] German Research Center for Artificial Intelligence (DFKI), Educational Technology Lab, Berlin, Germany, niels.pinkwart@dfki.de, https://orcid.org/0000-0001-7076-9737

doi: 10.18420/delfi2023-14

create tasks without the need to code tools. Due to the integration into learning management systems, such tasks can easily be integrated into courses [Pe16]. A course generator, as introduced by Rüdian et al. [RP21], allows teachers to fill H5P items with content, but the approach requires coding skills to apply. In combination, the existing technology allows generating course materials covering reading, and writing skills.

Text-to-speech approaches allow the generation of speeches with high quality, in any language where trained models exist [Ren20]. Any text can be read out. Generated audio files allow the creation of listening tasks. Tasks that are used to enhance speaking skills can be created using pre-defined models like the H5P dictating tool [HH19]. Alternatively, speech-to-text approaches can be used to identify what learners have spoken and compare that with existing patterns to identify possible divergences [Zh20].

Altogether, many experimental solutions allow teachers to create language learning courses for any topic. However, there is still no solution that generates the baseline: vocabulary, and texts, based on a given topic. This paper aims to bridge that gap by focusing on the research question, of whether an auto-generated language learning course unit using GPT3.5 creates correct and appropriate vocabulary, sentences, and texts to form learning material. The examination in this paper focuses on:

1. Identifying vocabulary sets,
2. Combining vocabulary,
3. Generating sentences, and
4. Generating texts.

Being able to generate language learning online courses is of high interest due to different reasons. First, creating learning units requires expertise for the considered domain. Second, creating a variety of units depends on the creativity of the creator. Third, learners have different learning goals, and fourth, they differ in pre-knowledge. Hence, courses should be personalized to suit learner needs. Personalized course sequencing aims to "dynamically select the most appropriate resource at any moment, based on the current needs and goals of the learner" [Ul07]. Item sequences can be arranged in different ways to fit learner needs. Normally, courses are created while tutors have some assumptions, e. g. about learners' pre-knowledge, on that basis a course progression is designed. This pre-defined course sequence may not be optimum for all learners. One challenge is to design online courses in a way to be engaging and to hold an appropriate difficulty level. However, the one-size-fits-all solution is still often preferred as this is the solution with minimal effort. But if course sequences suit learner needs, they can be highly engaging [Co02]. Further, tutors' assumptions can be wrong, or learner needs change over time. Then, adjustments are required. Yu et al. highlight that "it may be advantageous to allow students more freedom to access the learning contents without strictly following the pre-defined sequence" [Yu17]. The basis to personalize item sequences is the existence of micro-learning units, containing fixed blocks of learning materials, which can be arranged differently due to their independence [Yu17]. If learning material, using state-of-the-art approaches can be generated, those contents are correct and useful, we have the basis to generate personalized language learning online courses.

2 Methodology

To examine, whether generated item sequences are useful, a pipeline is designed which allows teachers to provide textual input (like a topic), which is then processed to generate learning content. Therefore, appropriate prompts for the generative model GPT3.5 are explored that lead to the aimed outputs. Then, the resulting content is rearranged and combined to simulate suitable tasks of micro-learning units. In this paper, the idea of language immersion is used, intending that the learner is situated in a foreign-language environment. The learner learns the language through visual relationships, without the need for translations from one into another language. This is comparable to learning a language in childhood. In computer-based dynamic immersion language learning, typical daily-life situations are created, and learners must find connections between images and texts, or spoken voice. The supplier “Rosetta Stone” uses the approach in its application [Ro07].

Within an immersion-based course, a set of lexical categories L is used, like subjects S, or predicates P. More categories are existing, but for better readability, the paper focuses on those two. Subjects can be nouns, with $\{A, B, C, D\} \in S$, predicates can be verbs, with $\{a, b, c, d\} \in P$. For all items $x \in \{S, P\}$ of a task T, a set of corresponding images $I_{x1} \dots I_{xn}$ with $n \in \mathbb{N}$ exists. Hence, for noun A, there is a set of related images $I_{A1} \dots I_{An}$. Different combinations of $\{S, P\}$ are possible and they can be related to images as well. A micro-learning unit consists of a task’s set following the conditions, 1) that in each task new vocabulary is taken from one lexical category only, which can be combined with another category if related words are not new, and 2) that for new words, there is a logical inference to find related images. Thus either new vocabulary for S, and/or P is introduced within single tasks. The introduction of new words is done by showing images with related words, or by single-choice tasks. The presentation of images with related ones is straightforward. Single-choice tasks follow the pattern, that a subset of words with related images is shown, e.g. $[A, I_{A1}]$ and $[B, I_{B1}]$. Then, X in $[X, I_{A2}]$ must be determined by selecting the correct answer, A or B. By definition, only $[A, I_{A2}]$ exists, thus $X = A$. In the following task, X in $[X, I_{B2}]$ must be determined, etc. More complex single-choice questions are possible. Exemplarily, A is presented by two images I_{A1}, I_{A2} first. Then, X in $[A, X]$ must be determined by selecting the correct image from $\{I_{A3}, I_{B2}, I_{C1}\}$, where I_{A3} is the correct answer for the word A. Next, B is presented by one image I_{B1}, and X in $[B, X]$ must be determined using the set $\{I_{A3}, I_{B2}, I_{C1}\}$. Here, I_{B2} is the only correct answer. Finally, C is represented by an image, and X in $[C, X]$ must be selected from the set $\{I_{A3}, I_{B2}, I_{C1}\}$. Due to the logical inference $X = I_{C1}$. If some words are known due to previous tasks, they can be used for more complex logical inferences. One example follows. Assuming nouns $\{A, B\}$ are already known and two verbs $\{a, b\}$ need to be introduced. Then, $[\{A, a\}, I_{Aa1}]$ and $[\{B, b\}, I_{Bb1}]$ are presented. X_1 in $[X_1, I_{Ab1}]$ and X_2 in $[X_2, I_{Ba1}]$ must be determined by selecting the correct answers from $\{\{A, b\}, \{B, a\}\}$. As the relation of an image to A or B is already known, hence $[A, I_{A1}]$ and $[B, I_{B1}]$ exist, the task can be solved without knowing the meaning of a and b, because for I_{A1} , only the answer that contains A can be the correct one, which is I_{Ab1} and vice versa.

For sure, highly complex combinations are possible. Based on the concept of immersion, templates can be defined, which have to be filled with learning content. For the concrete experiment, 8 nouns $\{A, B, C, D, E, F, G, H\} \in S$, and 4 verbs $\{a, b, c, d\} \in P$ are used for each micro-learning unit, assuming that related images $I_{x1} \ldots I_{xn}$ exist. The unit consists of 8 sub-tasks (Fig. 1). In tasks 1-4, $\{A, B, C, D\}$, and $\{a, b, c, d\}$ are introduced. Therefore, a subset of both sets is selected ($\{A, B, C\}$ and $\{a, b, c\}$), used to introduce the relations of words to images (like $\{A\}$ to $\{I_A\}$), which are presented to the learner (tasks (1), and (2)). Then, $\{D\}$ and $\{d\}$ can be found in tasks (3), and (4). Learners can identify them by logical thinking, as those are the new words, where only one new word and new image exist. To repeat a verb like $\{\alpha\}$, it can be combined with all nouns in task (3), but it is not mandatory. Introducing $\{D\}$ only, is also sufficient (like in task (4)). In task (5), words are combined to form short sentences. Nouns' combinations are created. As all nouns are semantically related to verb $\{\alpha\}$, their combination is also related to $\{\alpha\}$. Further, verbs $\{b, c\}$ are combined with $\{b\} + \{AB\}$, and $\{c\} + \{CD\}$. Related images must be found by detecting combined nouns that are related to given verbs. In (6), four new nouns $\{E, F, G, H\}$ are introduced, while for $\{E, F, G\}$, images of $\{A, B, C, D\}$ are used as distractors, so that the learner must recognize previous images to detect new ones. $\{H\}$ can be found due to logical thinking similar to tasks (3), or (4), but limited to a noun. Tasks (7), and (8) contain a text/story consisting of 8 short sentences, which include a subset of $\{A, B, C, D, E, F, G, H\}$, $\{a, b, c, d\}$, and related images. The set of 8 tasks is used as the sequence model in the experiment. The challenge is to fill that template with content.

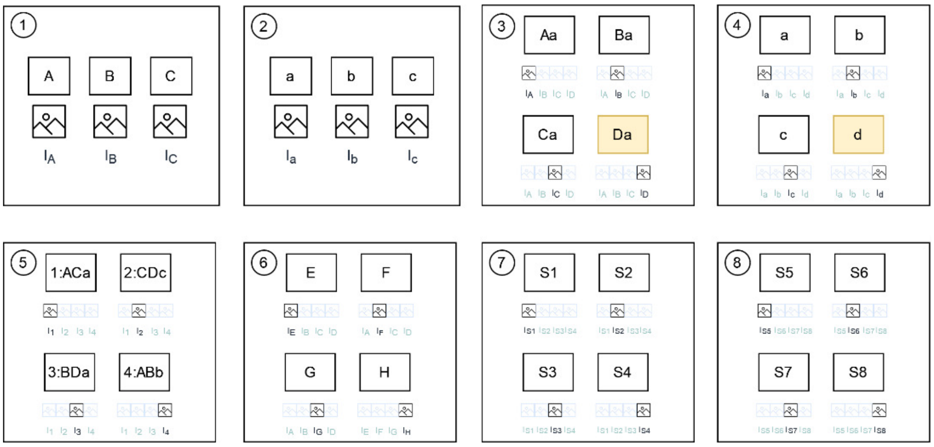

Fig. 1: Prepared template to generate 8 tasks based on language immersion

Micro-learning units are created based on this template. Contents are combined using rules as each task of the unit contains a fixed number of words, which must be identified, and combined. The approach allows generating tasks if required images exist. The order of images or answers within a task can be randomized. The investigation does not focus on personalization, which is not intended in this paper as the focus remains on generating texts for micro-learning units. Those units are then evaluated. As courses can be evaluated on different criteria, the most crucial: correctness and appropriateness are analyzed and evaluated by a teacher.

3 Architecture

Next, the architecture is described. The teacher or the learner provides a topic as input. Based on that, lists of related verbs, and nouns are asked for. Some of these nouns and verbs can be combined. Thus, a new list is generated by asking to merge both lists to create connections if there is a semantical relation. For the collected words, it must be determined whether all constraints are fulfilled so that the template can be filled with content. Therefore, the dataset is validated on the existence of words, with relations $\{A, B, C, D\}$ to $\{a\}$, $\{AB\}$ to $\{b\}$, and $\{CD\}$ to $\{c\}$. If that combination exists, the process continues. Otherwise, new word lists must be generated. Then, sentences are formulated based on combined word sets as defined in Fig. 1. Sentence lengths can be limited by word counts, and language level to avoid using words that might be unknown to the learner. Finally, a short text is generated based on the vocabulary used in the unit. To avoid a potential language mix, all prompts are formulated in German, intending to get responses in the same language. Principally, an instruction may contain the aimed target language, but trials have shown that this is not always the case. All used prompts are given in Tab. 1.

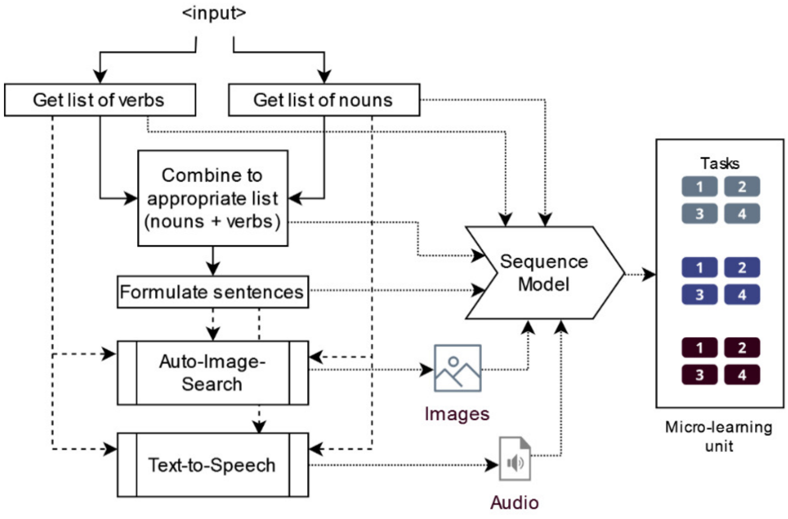

Fig. 2: Pipeline to auto-generate micro-learning units

For vocabulary (verbs, and nouns), and sentences, images can be generated using a generative model, which returns images based on textual inputs. All words, sentences, and files can be submitted to a sequence model, as visualized in Fig. 2. The same inputs can be used to derive sound files using a text-to-speech engine. As speech can already be generated easily with state-of-the-art tools, it is assumed that learning materials using appropriate textual contents are either appropriate combined with speech.

Types	Prompts
Nouns $\{A, B, C, D, E, F, G, H\}$	Erstelle eine Liste von 20 Substantiven zum Thema [#topic]. Sprachlevel A1. (*Make a list of 20 nouns related to the topic [#topic]. Language level A1.)*
Verbs $\{a, b, c, d\}$	Erstelle eine Liste von 20 Verben zum Thema [#topic]. Sprachlevel A1. *(Make a list of 20 verbs related to the topic [#topic]. Language level A1.)*
Relations $Nouns \Leftrightarrow Verbs$	Kombiniere Wörter [#nouns] mit [#verbs]. Die Liste soll nur das Substantiv und mindestens 4 Verben in einer Zeile enthalten. Verwende keine Nummern und keine Aufzählung. *(Combine words [#nouns] with [#verbs]. The list should contain only the noun and at least 4 verbs in one line. Do not use numbers or enumeration.)*
Sentence	Erstelle einen Satz, der die Wörter [#noun1,#noun2,#verb] enthält. Der Satz darf maximal 12 Wörter enthalten. Sprachlevel A1. *(Create a sentence that contains the words [#noun1,#noun2,#verb]. The sentence can contain a maximum of 12 words. Language level A1.)*
Text	Erstelle einen zusammenhängenden Text aus 10 Sätzen. Nutze folgende Wörter: [#words]. Jeder Satz darf maximal 7 Wörter enthalten. Sprachlevel A1. Jeder Satz in einer Zeile. Verwende keine Zahlen. *(Create a coherent text of 10 sentences. Use the following words: [#words]. Each sentence can contain a maximum of 7 words. Language level A1. Each sentence in one line. Do not use numbers.)*

Tab. 1: Prompts used to gather information

The approach generates a unit for a given topic. Then, the next topic must be provided. To automate that step, a “random surfer model” is used [CM08]. Hence, a random word is selected from $\{A, B, C, D, E, F, G, H\}$ of the current unit, which is used as topic to generate the next unit. The approach runs until 200 units are generated, which are then evaluated.

Principally, the evaluation of an online course (or its units) can cover a wide range of criteria, ranging from considering pre-knowledge [AMF03], quality of learning materials [XLZ20], to coherence [MJP15]. The evaluation in this study is limited to words, sentences, and texts, with a focus on the correctness, and appropriateness of the generated learning content using a generative GPT model (GPT-3.5-turbo). Appropriateness is an essential element when evaluating a course [AMF03]. For the study, words, sentences, or texts are appropriate, if they are coherent, with logical order, and sentence lengths as demanded that are suitable for a language level A1, without hallucinated contents [Ba23]. The latter is important to help learners to find the correct word meanings, and relations to contexts without confusing them. It must not be emphasized, that course material for language learners must be correct (at least spelling, and grammar in a language learning course) [XLZ20]. Incorrect learning materials are crucial as it is a knockout criterion to apply the approach in real-world scenarios. Hence, the two criteria are examined, appropriateness ("Is the output appropriate for the context?"), and correctness ("Is the output a correct response?"). Units are rated by an experienced language teacher on a binary scale (yes/no). To compare the resulting tasks within units, words, sentences, and texts are examined separately to uncover potential limitations.

4 Results

200 learning units are generated, limited to words (1, 2, 3, 4, 6 in Fig. 1), sentences (5 in Fig. 1), and texts (7, 8 in Fig. 1). A visualization of the evaluation is given in Fig. 3. Considering words, most of them are correct (.985), and appropriate (.97). If single sentences are formulated based on two, or three given words with maximum sentence length of 12 words, 4/5 are correct (.805), and appropriate (.835). Texts are mainly correct (.9), and most texts are appropriate (.935). In general, the evaluation shows that selected words and generated texts outperform in correctness, and appropriateness, while generated sentences have some weaknesses. Considering generated samples that are not erroneous at all, hence, words, sentences, and texts are correct, and appropriate within the unit, results in 56.5%. This subset can be directly processed without the need for further adjustments.

First, typical mistakes of selected, and combined words are examined. Exemplarily, "kaufne" (to buy), or "Kücken" (chicken) are misspelled. Such erroneous words are directly provided by the generative model. Also, semantical uncommon relations are found. The approach asks the generative model to create a list of related nouns with verbs. Exemplarily, words like "dip + cow" are used to form the sentence "The cow is dipped." (de: *Die Kuh wird portioniert.*), or "practice + bag" is combined to "I practice in the bag" (de: *Ich übe in der Schultasche.*). Those chunks must not be incorrect but can be unusual without further contextual embedding.

Second, generated sentences consisting of three given words are often grammatically incorrect or do not make sense. Tab. 2 shows some examples of uncommon sentences, that are not appropriate, or erroneous in German. Further, but rarely, a language switch can be found in the response. Exemplarily, for the word "fruit", the Finnish word

“Hedelmä” is used. Also, complete sentences are generated in another language, but with a given translation, exemplarily “Minä maistan banaania ja vadelmaa. (Ich schmecke Banane und Himbeere.)”, which means “I taste banana and raspberry”, but given in Finnish.

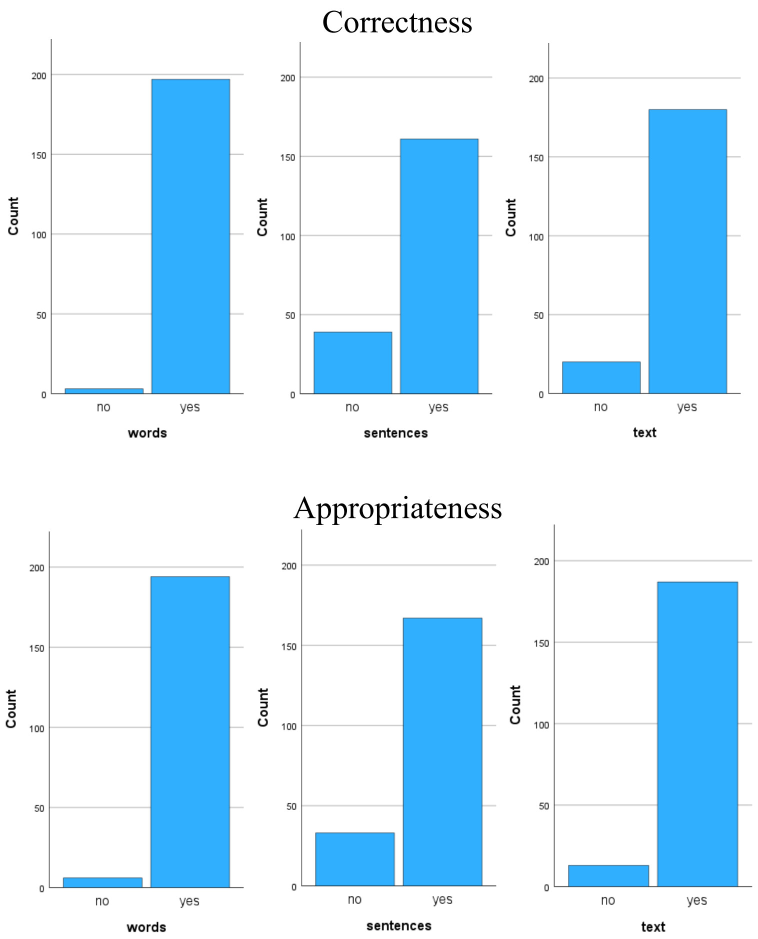

Fig. 3: Results of the evaluation (correctness and appropriateness)

If the model cannot generate an appropriate sentence with given conditions, the response looks like this: “I cannot create valid sentences that contain "[words]" and have language level A1”, although the model previously claimed that there is a relation between those words. To give an example, words [chocolate, cheese, liquefy] cannot be combined by the model, but the model previously claimed that there is a relation between them. Further, responses are not deterministic. The response containing the information of not being able to create a sentence is manifold. Besides, generated sentences can directly be listed in the response, or a preamble is added like “I can create this sentence:”, followed by quotes. Also, after some sentences, the number of words is given in brackets. Such information must be removed.

Types	Sentences
Uncommon	- Ich übe in der Schultasche. *(I practice in the school bag.)*
	- Ich bete im Taufbecken. *(I pray in the baptismal font.)*
	- Der schwüle Donner rollte heran. *(The sultry thunder rolled in.)*
	- Ich werde das Nest füttern. *(I will feed the nest.)*
	- Der Apfel ist voll. *(The apple is full.)*
	- Die Kuh servieren wir frisch. *(We serve the cow fresh.)*
	- Die Einbruchmeldeanlage beleuchten den Raum. *(The burglar alarm lights the room.)*
	- Ich kaufe eine Leseratte. *(I'll buy a bookworm.)*
	- Die Kuh wird portioniert. *(The cow is portioned.)*
	- Die Teppiche wachsen im Kaufhaus. *(The carpets grow in the department store.)*
	- Am Tag der Deutschen Einheit schmücken wir auch Heiligabend. *(On German Unity Day, we also decorate Christmas Eve.)*
	- Ich serviere Kaffeetrinkern in der Kaffeepause. *(I serve coffee drinkers during the coffee break.)*
	- Die Konservierungsmittel lagern in Zahnheilkunde. *(The preservatives are stored in dentistry.)*
Grammar mistakes (German-only)	- Einen Tischplatte abschrauben ist schwierig.
	- Ich suche den Autorenname.
	- Ich möchte Kirsche probieren.
	- Ich liebe es, Hühnchen scharfzu machen, indem ich es in Zitronensaft marinieren.
	- Ich muss den IBAN des Kontoinhabers prüfen.
	- Bitte ausstecken Sie das USB-Ladekabel vom USB-Netzteil.

Tab. 2: Common mistakes of generated sentences
with equivalent translation including erroneous parts

Next, generated texts are examined. They achieve the highest scores in the evaluation. Most of the texts are coherent, and they follow a logical order. Nevertheless, examples are found which are grammatically correct but do not make sense. Tab. 3 provides two examples in German. The first combines “heat” with “black ice”. The second describes baking a “cake”, but then, “paprika” is added – and at the end, everything must be “flavored”, which is not typical when baking cakes.

Further, texts in responses can be formatted differently. Instead of creating one sentence per line as requested, sometimes no line breaks are included, or double line breaks are used. The last sentence of the generated text may contain: “these 20 nouns and verbs help with vocabulary training”, which is not asked for. Again, such information must be removed to be useful in a learning unit. Nevertheless, texts perform remarkably better than generated single sentences.

Example 1	Example 2
Die **Hitze** drückt, ich brauche Abkühlung.	Ich koche gern.
Ein kühles Bad im See wäre super.	Heute backe ich einen **Kuchen**.
Aber das Wasser ist eiskalt, **Glatteis droht**.	Dafür brauche ich Backpulver, Mehl, Butter und Zucker.
Also doch zurück zur Klimaanlage im Büro.	Ich rühre alles zusammen.
Doch der Ventilator tut es auch.	Dann füge ich Eier hinzu.
Er sorgt für Luftbewegung und Ventilation.	Nach dem **Rühren** gebe ich **Paprika** hinzu.
Das Kühlen hilft bei der Konzentration.	Ich schneide die Zutaten klein und gebe Öl hinzu.
Entspannen und Relaxen ist einfach so möglich.	Alles vermische ich gut.
Chillen ist trotz der Hitze kein Problem.	Das Ganze kommt in den Ofen.
Hauptsache, ich bleibe kühl und entspannt.	Zum Schluss **abschmecken** und genießen.

Tab. 3: Semantical errors in generated texts

5 Limitations & Challenges

One challenge is the formatting of the response using generative large language models. An instruction-tuned GPT model is trained using existing prompts, and responses. More complex prompts are based on a variety of templates, which are filled with contents (responses from sub-queries). Such templates can be structured as proposed by Wang et al. [Wa22], exemplarily: “*Come up with a series of tasks: Task 1: {instruction for existing task 1}, Task 2: {instruction for existing task 2},…*”. As long as those templates are not publicly known, there is the chance that a response does not meet the expected format, so it cannot be processed as expected. During the experiment, enumerations were given in responses, although the query explicitly includes not using numbers when generating texts. Further, instead of providing one sentence as requested, the model generates multiple sentences, or the response starts with “I will try three variants:”. Such results are unexpected and must be handled separately, as we cannot be certain that the model will follow all instructions even if previous results were valid. Hence, even if instructions include that the response should contain one sentence per line, texts without line breaks are often seen. Despite the variety, most responses are formatted as expected to be useful.

Even if topics are selected from language level A1, resulting sentences, and texts can often be too complex for beginners. To name an example, for the topic “eye”, the selected word list which fulfills all conditions, starts with words like “intraocular pressure, eye inflammation, or ophthalmology”. From an educational perspective, those are not the words to begin with. If the random surfer model is applied, and one of the non-conventional words is selected, new units become unnecessarily complex. Hence, results should be filtered by language level afterward. The experiment is limited to A1 contents. Further research can examine other levels as well. Besides, generated units are evaluated by one teacher. Hence, the result may be biased due to his/her quality perception.

Generating course units using the proposed method leads to low costs. At the current state, using GPT-3.5, $0.002 per 1K tokens are charged [Op23]. Based on those costs, the final costs for the experiment focusing on words, sentences, and texts are 1.60$, resulting in 200 language learning units of different topics. However, an evaluation of whether generated images, or speech, are appropriate, is not part of the paper and is examined in further research.

For the experiments, GPT-3.5 is used, those responses are transferred into a structured format and processed to form a learning unit. Nevertheless, the GPT model is closed, and cannot further be controlled by researchers. From the privacy perspective, using such a model is acceptable as no personal data is used, or processed. Still, all prompts can be accessed by a company that provides the model, which may not be acceptable if prompts, and results are the base for further products. Alternative open models are required, which mimic the functionality of generating words, sentences, or coherent texts. Exemplarily, an open model like GPT-J [WK21] could be fine-tuned, using prompts, and responses from GPT-3.5. Then, the model can mimic responses for the concrete instructions, but with new contexts. Nevertheless, the resulting quality must be examined as the GPT-J model consists of 6B parameters, and GPT-3.5 covers 175B [Ko23]. Alternatively, collections like the “wordnet” [Fe12] can be used to get words with a semantical relation. Further research can examine, which approach leads to the best result.

6 Conclusion

In this paper, the generative language model GPT3.5 is used to create learning content for a concrete template. Results have shown that the model performs best in generating texts. Some weaknesses are found if sentences must be formulated using concrete words. All in all, fully auto-generated language learning online courses with high quality cannot be expected using the approach at the current state. A teacher-in-the-loop method must be favored. However, for teachers, auto-generation can be helpful, as they just need to select which result is acceptable and can generate new versions at low cost. For teachers, the procedure of creating course units with a given template is less time-consuming. They must not be that creative, as they can concentrate on providing topics, while the remaining part, including generating interactive H5P elements, can be done automatically. Further research will show whether open models achieve comparable results, which opens the door for generating highly varying adaptive language learning units that best suit learner needs.

Acknowledgments: This work was supported by the German Federal Ministry of Education and Research (BMBF), grant number 16DHBKI045 (IMPACT).

Bibliography

[AMF03] Achtemeier, S. D.; Morris, L. V.; Finnegan, C. L.: Considerations for developing evaluations of online courses. In Journal of Asynchronous Learning Networks 7.1, University of Georgia, pp. 1-13, 2003.

[Ba23] Bang, Y. et al.: A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. arXiv preprint, 2023.

[CM08] Chebolu, P.; Melsted, P.: PageRank and the random surfer model. In SODA (8), 2008.

[Co02] Conrad, D. L.: Engagement, excitement, anxiety, and fear: Learners' experiences of starting an online course. In The American journal of distance education 16(4), pp. 205-226, 2002.

[Co21] Council of Europe: A Common European Framework of Reference for Languages: Learning, Teaching, Assessment. In Council for Cultural Co-operation. Strasbourg: Cambridge University Press, 2001.

[Fe12] Fellbaum, C: WordNet. In The Encyclopedia of Applied Linguistics, C. Chapelle, 2012.

[HH19] Homanová, Z.; Havlásková, T.: H5P interactive didactic tools in education. In EDULEARN19. IATED, 2019.

[Ko23] Koubaa, A.: GPT-4 vs. GPT-3.5: A Concise Showdown, 2023.

[MJP15] McGahan, S. J.; Jackson, C. M.; Premer, K.: Online course quality assurance: Development of a quality checklist. In I. A. (10), 2015.

[Op23] OpenAI: Pricing. Retrieved from https://web.archive.org/web/20230305235115/https://openai.com/pricing, accessed: 2023 03 05.

[Pe16] Petterson, F. et al.: Activities: Interactive Content – H5P. Retrieved from https://moodle.org/plugins/mod_hvp, accessed: 03 01 2023, 2016.

[Ren20] Ren, Y. et al.: Fastspeech 2: Fast and high-quality end-to-end text to speech, 2020.

[Ro07] Rosetta Stone Ltd.: Rosetta Stone User's Guide (2). US, 2007.

[RP21] Rüdian, S.; Pinkwart, N.: Generating adaptive and personalized language learning online courses in Moodle with individual learning paths using templates. In International Conference on Advanced Learning Technologies (ICALT), pp. 53-55, 2021.

[Ul07] Ullrich, C: Course Generation as a Hierarchical Task Network Planning Problem. In Dissertation. Saarbrücken, 2007.

[Wa22] Wang, Y. et al.: Self-Instruct: Aligning Language Model with Self Generated Instructions. arXiv preprint, 2022.

[WK21] Wang, B.; Komatsuzaki, A.: GPT-J-6B: A 6 billion parameter autoregressive language model, 2021.

[XLZ20] Xu, D.; Li, Q.; Zhou, X.: Online course quality rubric: a tool box. In Online Learning Research Center, 2020.

[Yu17] Yu, H. et al.: Towards AI-powered personalization in MOOC learning. In npj Science Learn 2 (15), pp. 1-5. Nature, 2017.

[Zh20] Zhang, L. et al.: End-to-end automatic pronunciation error detection based on improved hybrid ctc/attention architecture. In Sensors, 20(7), p. 1809, 2020.