Large Language Models and Artificial Intelligence, the End of (Language) Learning as we Know it—or not quite?

Cerstin Mahlow
School of Applied Linguistics
ZHAW Zurich University of Applied Sciences
cerstin.mahlow@zhaw.ch

2023-02-12

Abstract

The rapid advancements in large language models (LLM) and artificial intelligence (AI) have been a subject of recent significant interest and debate. This paper explores the impact of these developments on language learning. I discuss the technology underlying AI-based tools and the natural language processing (NLP) tasks they were originally designed for. This will help us to identify opportunities and limitations regarding their use in the context of language learning. I then examine how such technology can be used efficiently and effectively in language teaching and learning. The availability of such tools will require language teaching to focus on the non-mechanical aspects of writing. Similarly, automatically produced personalized teaching and learning materials will not replace human teachers, but give space for and support human–human interaction.

1 Introduction

For some years now, language technology and Natural Language Processing (NLP) improves ever faster and seems to have achieved a degree of maturity to be effectively and efficiently included into almost all daily communicative situations. For some applications like translation, it is already used as an accepted agent to take over some tasks previously done by humans. It has been proven again and again, however, that machine translation (MT) cannot replace human-lead translation but rather changes established procedures and processes. (Bowker and Ciro, 2019) The quality of current MT systems also raises questions on whether or not it is actually useful or necessary to actively learn a foreign language, and how to include MT as an element into the Common European Framework of Reference for Languages (CEFR) (Delorme Benites and Lehr, 2022).

Applications using language technology are often called Artificial Intelligence (AI) or AI-based tools. The sudden availability of such applications to the general public in late November 2022, with ChatGPT from OpenAI as the most prominent example, seems to indicate that indeed automatic production of texts in human-like quality is now possible. Millions of users tried out ChatGPT in December 2022, the underlying resources got integrated into existing or new applications. Social and traditional media were flooded with astonishing examples of texts produced by ChatGPT, but slowly skeptical voices appeared, pointing to flaws or even errors. Previous to ChatGPT, Meta had released Galactica, which swiftly had to be deactivated as users discovered and showed serious issues that had not been properly communicated or addressed by Meta. ChatGPT in contrast appeared to "have learnt" from this and made clear that not everything in the texts produced should be taken too serious; the application also refuses to produce clearly offensive or malicious texts.

Technology of any kind is created to serve certain purposes. Language technology in general is designed and implemented to solve specific language-related tasks. ChatGPT and similar applications produce texts according to prompts in an interactive way while the user is in the lead.

The sudden availability of mature language technology to the general public in late 2022 together with advertisement explicitly highlighting its capabilities to produce essays as expected to be written by students and academics alike, started fierce discussions on social media, in ad hoc organized online conferences and talks in December 2022, and in newspapers.

Considering the assumed and actual relevance for learning and teaching, this discourse constitutes an example of what Mahlow and Hediger (2021) elaborate as loosely coupled system of technology and pedagogy. In this case general-purpose technology can be used for teaching and learning by leveraging certain aspects and making use of affordances (Mahlow and Hediger, 2021). It is also a very nice example of a view Norman (1993) expressed: Technological progress has massive effects on societal progress wich leads to further purposeful development of technology (Norman, 1993, 7-8). Interpreting "society" as educational institutions, we are just witnesses of developments both in pedagogy as well as in technology which are connected in a flexible way forming a "loosely coupled system" (Weick, 1976). As technology has made massive progress, we are now faced with massive effects in learning and teaching. Understanding both this coupling as a general principle and its effects as well as the concrete technology it affects, will allow for purposeful but far-ranging controlled transformations in (language) learning and teaching. Which in fact could be called digital transformation of language learning, as whole processes are involved going beyond mere digitization and digitalization (Mahlow and Hediger, 2019).

In this paper I focus on how to use current language technology for language learning with an emphasis on writing. Writing is both an everyday activity and a significant competence learners acquire to actively and successfully participate in communicative situations. Writing and learning how to write in a first as well as in a foreign or second language is thus an important skill. Writing competencies are specifically addressed as written production in the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001, 2020). Producing written texts involves cognitive and linguistic competencies. Cognitive competencies address knowledge about genres and organizing relevant information or facts to achieve a certain communicative goal. Linguistic competencies on the other hand cover general morphological and syntactical knowledge, but also refer to a repertoire of vocabulary, phrasal structures, and discourse devices language learners acquire. Both aspects are interconnected: What is usually covered by "style" might be interpreted and described as appropriate genre-specific linguistic features.

Learning how to write thus should address all aspects and offer opportunities to explore and train cognitive, pragmatic, as well as linguistic competencies. Written texts produced by language learners are often used as evidence of learners' performance and serve assessments purposes. They supposedly show whether or not learners are able to correctly recognize specific communicative situations and apply appropriate linguistic features to master this scenario and thus prove competence at a certain level.

If we agree that similar to earlier experiences in the field of translation, language technology changes established writing processes but never replaces humans completely, learning how to write will remain in curricula. Available applications as well as potential new ones—given the general capabilities of the underlying resources—will become an integral part of language learning. Focusing on opportunities coming with almost disruptive changes to learning and teaching will allow to focus on specific pedagogical aspects (again) and support learners in new ways.

The rest of this paper is structured as follows: First, I look at the underlying technology for AI-based applications like ChatGPT and show what they are designed for in section 2. This will help to understand general chances and limits. Section 3 looks at opportunities for automation in learning and teaching for which language technology in fact can be used efficiently and effectively.

2 LLM as core of language-processing and language-producing AI applications

Specific applications of language technology are the focus and the center of enthusiastic optimism as well as of fierce debates about danger of all kinds. Single applications are used as proxy in discourses—at the time of this writing, "ChatGPT" is used when actually discussing AI-based language technology at large—where we rather should look at the underlying resources and technology to be able to evaluate and assess general possibilities and limits. This certainly requires good understanding of the nature of these resources, general scenarios they could be used in, and assumptions of what would be needed to make it useful for the intended task (e.g., percentage of correct results or performance compared to humans).

The core of today's AI-based applications in language learning and teaching are Large Language Models (LLM), which I address in this section. If we understand how language models are constructed and what they can be used for in general with reasonable confidence, we will be able to determine the role of existing and future applications based on these resources. A crucial aspect when talking about language technology are general approaches and tasks in Natural Language Processing (NLP), which clearly profit from improving LLM and which can be used as resources and tools for activities in learning and teaching. Clearly, the currently most impressive use of LLM is the production of texts in reaction to user prompts. From a technical point of view, ChatGPT is just a chatbot using a specific LLM with a very convincing and easy to use interface; its success underlines the importance of the user interface.

2.1 Large Language Models

Large language models are resources for language technology applications that use deep neural networks with a huge number of features and parameters trained on massive amounts of text data. Examples are Google's BERT (Devlin et al., 2019) with 340 million parameters and PaLM (Chowdhery et al., 2022) with 540-billion parameters, and OpenAI's GPT-3 trained on 45 tera bytes of text with 175 billion parameters (Brown et al., 2020). OpenAI's ChatGPT is a chatbot based on an advanced version of GPT-3—GPT-3.5—with additional fine-tuning and the capability to store and make use of previous utterances both from the bot and from the user. GPT-4 is announced to be released in early 2023 with even more parameters.

wordformFor training—i.e., for creating the models in the first place—, these models use unlabeled text. The texts have not been preprocessed or annotated for any information: they use so-called unsupervised learning. Which also means that there is no significant control or knowledge of the sources and original intentions of these texts: one would perhaps like to exclude machine-produced texts (either generated or automatically translated texts), to verify the language of texts and probably treat multilingual texts separately, use weighting to address and level out biases, etc. Modeling only takes into account what are usually called left and right contexts of words, i.e., what is before and after each single wordform.

In a broad linguistic abstraction: language models entail only *co-text* of language units but have no access to *context*. The power of a model is influenced both by the amount of training data and by the number of parameters the model uses to first classify co-text and later retrieve possible or plausible co-text by using a new architecture called transformers (see Tay et al., 2022). *Large* language models are simply language models trained on large amounts of texts and use large amounts of parameters.

Language models and their use could be understood as an actual implementation of Firth's well-known statement: "You shall know a word by the company it keeps" (Firth, 1957, 11) which influenced lexical semantics and was used to allow for drawing conclusions not only on the meaning of words but also on the concepts behind words or sequences of words. However, words and their co-texts do not allow to deduce all facts, readers as well as writers are aware of and use. Obvious examples include the need for providing additional context when reading and processing historical texts (Piotrowski, 2012) as well as the still hard task to unambiguously determine specific persons and places which needs additional resources in named entity recognition (Wang et al., 2021).

However, LLMs—and language models in general—are used in various specific tasks in Natural Language Processing (NLP) and are one reason for significant performance gains in these tasks.

2.2 Natural Language Processing Tasks and Potential Use for Language Learning and Teaching

Some established NLP tasks serve as basis for applications in language learning and teaching. The field of Computer-Assisted Language Learning (CALL) has progressed a lot since Levy (1997) established the term, which later changed to Technology-Enhanced Language Learning (TELL). Zhang and Zou (2022) provide a comprehensive overview on the current state of the art focusing on five major types of technology for second and foreign language learning: mobile-assisted, multimedia, socialized, speech-to-text and text-to-speech assistance, and gamification. Here we take a different perspective and show which NLP tasks can be incorporated in TELL, focusing on generally available technology for non-educational purposes whose affordances allow for use in educational applications and settings (Mahlow and Hediger, 2021).

NLP is used for text indexing and keyword extraction (e.g., Hulth, 2003), which can be used just as such: to create an index of several texts or a set of keywords for a text. In addition, there is a long tradition on formally modeling the underlying structure of a text (e.g., Hobbs, 1984; Hobbs et al., 1982) used for answer extraction (AE) (e.g., Schwitter et al., 2000; Seonwoo et al., 2020) and question answering (QA) (e.g., Adlakha et al., 2022; Kwiatkowski et al., 2019), where users can input a question and are either provided with a fitting text snippet (AE) or with a generated answer based on extracted small text snippets. These techniques can be used for the construction of reading comprehension tests. Extracted keywords (or concepts) can also be provided as seeds for higher level systems to look up suitable references or to point readers to more elaborate explanations of these concepts considering the individual learner's level of competence (Meurers et al., 2010; Chinkina et al., 2016).

Summarization of texts build on these tasks: they are provided as extraction of the most relevant sentences (e.g., Jia et al., 2020) or by applying abstractive summarization where sentences with the most important information are generated (e.g., Wang et al., 2020). Irrespective of the technique applied, summarization can be used as part of a feedback system: the writer is provided with a summary of the text produced, they then have to decide themselves whether or not this summary fits their original communicative intention and identify passages in the text to be revised.

Approaches used in summarization can also be applied to shorten texts by keeping the original message intact. This could be used by learners to make their writing fit the formal constraints of a writing task, and it can be used by teachers to provide learners with an abbreviated version of a longer text from a newspaper or the like.

Checkers for spelling, grammar, and style have been around since the 1970s (Macdonald et al., 1982; Heidorn et al., 1982). Originally designed to support expert writers (Fontenelle, 2005; Heidorn, 2000), they have been adapted and advanced to address learners' needs (e.g., Gamon, 2010; Tschichold et al., 1997); this development is still ongoing (e.g., Sjöblom et al., 2021; Yuan and Bryant, 2021).

Google introduced a new approach to grammar correction in 2019 by interpreting grammatical error correction as "translation from an ungrammatical to a grammatical sentence" (Kumar and Tong, 2019). This allows for the application of machine translation—which is also based on language models—to this task (Lichtarge et al., 2019). Machine translation can also be used for round-trip-translations (Somers, 2005), which allows writers to get suggestions on alternative formulations of sentences and paragraphs of their texts: a text in language A is automatically translated into another language B, and the resulting text is then immediately translated back into language A. Due to the properties of LLMs, the resulting text often uses more idiomatic syntax, some words and phrases may be replaced by synonyms or plausibel alternatives respecting the co-text. This way, the writer receives a revised version for their original text which can be used as starting point to make learners aware of and discuss vocabulary choice, grammar, style, etc.

Co-reference resolution aims to identify and link entities (persons, organizations, places, etc.) mentioned in texts by variants of their full names or by pronouns (e.g., Peng et al., 2016; Roesiger and Kuhn, 2016). When learning to write in a foreign language, these co-reference chains can be used to produce fill-in-the-blank variants of a given text and ask writers to input appropriate variants of the respective entity. They can also be used to check for consistency when writers revise their texts.

Text simplification generates new texts simpler in vocabulary and syntactic structure by keeping the information and communicative goal of the original text (e.g., Grabar and Saggion, 2022). Simplification can be applied to produce plain/simple language versions of texts—e.g., for Wikipedia (Coster and Kauchak, 2011)—or to adapt texts for specific language levels.

Text-to-speech (automatic rendering of written texts into audio) and speech-to-text (automatic transformation of spoken utterances into written texts) can be used for dictation to help writers struggling with spelling or writer's block, and for having written texts read aloud to the writer to help them detect issues for revising.

Sentence similarity measures how similar sentences are with respect to syntactical and morphosyntactical structures (Das and Smith, 2009), or to vocabulary and semantic (Fernando and Stevenson, 2009). This approach can be used to detect topical relevance in learner essays (Rei and Cummins, 2016) as starting point for feedback. It can also be used to create phrasebooks from authentic learner input, to produce textbook examples to be presented to writers as acceptable instances of rhetoric patterns, and as ad hoc created variants of sentences or phrases the writer produced themselves. Another application is detection of (paraphrased) sentences or text passages produced by someone else, as used in plagiarism detection.

Additionally, any combination of the tasks briefly described above is possible to serve specific purposes. One crucial aspect here is the quality or correctness of results achieved by respective applications. Users will have to evaluate whether current state-of-the-art results are *good enough* to be used in real-world settings; it is nearly impossible to determine this without considering concrete context and tasks as shown by Mahlow and Piotrowski (2009) for automatic lemmatization and morphosyntactic analysis. The combination of resources and tasks need thorough evaluation to determine the overall quality and decide on suitability for language learning purposes.

2.3 Automated Text Production vs. Natural Language Generation

The output of applications using LLM is often called "generated text," which points towards the field of Natural Language Generation (NLG). NLG is the production of "understandable texts in [...] human languages from some underlying non-linguistic representation of information" (Reiter and Dale, 2000) and has been an active research field for several decades. In analogy to text-to-speech, it has also been named *data-to-text* (Schneider et al., 2022).

The most important aspect of NLG systems: data (i.e., measure points, facts in any kind of knowledge base, entire databases, etc.) are the explicit context of the text to be generated. This context is made available in a structured form to the machine that generates the text. Additionally, all information about intended audience, genre, text length, etc., has to be made explicit as well and is used as features and parameters for generation. Dale (2020) gives an overview of the (commercial) state of the art on NLG; Schneider et al. (2022) describe categories of NLG systems. They emphasize that:

> "data-to-text systems in real-world applications still require such a share of human configuration and control and the creative contribution share of the software [...] is still so limited that it would not be adequate to claim creative autonomy of the software in the process." (Schneider et al., 2022)

Current language technology is thus not an autonomous creative agent taking part in the writing process. There is no *interaction* between a human writer and a text-producing machine, only very sophisticated *interactivity*. Language technology is used for writing support in settings of automated text production as defined by Mahlow and Dale (2014). Dale and Viethen (2021) provide an overview on writing assistance based on state-of-the-art NLP resources—i.e., LLM—and approaches. In 2021, GPT-3 was already available and had been incorporated into various tools aimed at supporting writers by co-authoring, not only (copy-)editing. These applications addressed specific genres like blog posts and poetry, as well as specific writing tasks like expanding, rewriting, and shortening texts (Dale and Viethen, 2021). Some months later, they got included as writing aids into experimental editors (e.g., Dang et al., 2022; Yuan et al., 2022). However, they were not used widely and did not trigger the same discussions that we see two years later.

The more powerful LLMs get, the more users forget that texts automatically produced by any application using these resources are just plausible extensions of existing co-text. These co-texts are either existing (parts of) texts, which are then expanded, or prompts the system seems to "react" to: "GPT-3 is still very capable of generating nonsense, but on the whole it's more plausible nonsense; and with appropriate fine tuning and prompting, the texts it generates can be eerily convincing." (Dale and Viethen, 2021, 516)

In an abstract view, automated text production based on LLM are just very sophisticated and powerful further developments of predictive texting known from input-support on mobile phones (Ganslandt et al., 2009). No facts and information determine the next words or sentences, but only known co-texts of already existing words and sentences. The resulting text is correct from a linguistic point of view: it does not contain spelling and grammar issues, it is coherent and consistent. But it is plausible and acceptable only when focusing on the *language* produced.

All information that seems to be included in the text is currently merely due to frequent cooccurrences of words and thus not trustworthy. No references to any context can be made, no conclusions on underlying knowledge, understanding, or intention of an assumed author can be drawn. Depending on the genre the LLM is asked to produce, a text may contain facts—e.g., dates, places, or even correctly formatted bibliographical references. But also these are only *plausible*, they are not true; they could, but most often do not exist. The LLM just produces sequences of words that follow general patterns for bibliographical entries consisting of strings that are names, followed by an arbitrary four-digit-number, a string looking like a title, followed by a string mimicking a publishing house, etc. While references produced by a LLM are thus most probably non-existing, one can very confidently use them to check whether the *format* of a list of bibliography entries adheres to a certain citation style or is compatible with a list of example entries.

These systems are not creative: they do not invent anything as they have no agency. They just react to arbitrary input—be it a prompt from a human or previously produced text. All creativity, all surprisal is only in the mind of human readers interpreting these texts, while ignoring authorship and not being aware of the circumstances—i.e., the process—of their production.

2.4 User interfaces

User interfaces (UI) play an important role in perception of usefulness and trustworthiness; they are constructed to employ specific aspects. The design of user interfaces is a serious business; how to successfully address user needs depends on various aspects. The implementation of user interfaces for computer applications, in contrast, today is easier than ever before: current programming languages and toolkits together with vast amounts of tutorials support fast development and roll-out of responsive applications of any kind. Although it is not entirely trivial to design and implement a convincing user interface to allow using complex resources in specific ways, the essential aspect is to carefully define potential uses in a more abstract way in the first place.

In June 2022, Sharples (2022) reflected on automated essay writing. He describes access to the GPT3 Playground as "straightforward," but it actually requires several deliberate actions to be carried out in a specific order:

> Anyone with internet can sign up to the OpenAI website, gain an account, click the "Playground" tab, type a prompt such as the title of an essay, set the maximum length of output (up to 4000 language "tokens," or approximately 3000 words) and click the Submit button. A few seconds later, the system produces a typed and formatted text. (Sharples, 2022)

In 2021 and 2022, various teaching and learning materials would contain similar instructions for tasks intended to encourage students to explore AI-based text production. Since then, applications appeared with UIs that makes interacting with the language model much easier and builds on familiar user experiences (UX) known from similar applications. As an example, the UI of ChatGPT is similar to generic messengers or chat applications: Users type in a dedicated field, hit "send," and then receive a response which they answer by typing the next sequence of words in this field again. All parameters—e.g., length of output—are pre-set to fit most users needs.

LLM can also be integrated into existing writing applications like Microsoft Word, to allow for more advanced support . We also see re-implementations of popular applications with integrated access to LLM. One such example is lex[1], developed by Nathan Baschez, intended as a "Google docs-style editing experience" using the JavaScript framework Yjs[2] by Kevin Jahns with less formatting features compared with Google-Docs. It is extended with access to GPT-3 to let writers call up the language model for producing plausible continuations of the text produced so far (TPSF), taking into account everything before the current cursor position (i.e., the left co-text). (Justin, 2023) This can be used as a means to overcome writer's block or to produce possible ideas on how to continue. What makes this use case intriguing is the seamless integration of general LLM capabilities into a specific application users are already familiar with. This way language technology becomes just another feature, although a very powerful one.

3 Automation in language learning and teaching

Section 2 covered language technology originally intended to solve certain NLP tasks, but whose affordances also allow their use in language learning. In this section, I focus on language technology that specifically addresses the educational needs (Mahlow and Hediger, 2021) of language learners and teachers.

In general, major technological developments, as can be described in waves of "industrial revolutions," enable and support the consistent, almost automatic execution of tasks that previously required significant human effort. Automation of any kind frees up capacity that humans can use for other activities.

In learning to write, one such example is mastering handwriting or using a keyboard, i.e., some of the "mechanics" of writing. In the beginning, learners have to do everything very carefully and consciously. (Dowling, 1994) Once these processes have been internalized, writers are usually no longer aware of them. (Fayol, 1999; Kellogg, 2008) Consequently, they have more cognitive capacity (Piolat et al., 2004) available to focus, for example, on spelling—another skill that can be mastered almost automatically (Fayol, 1999)—coherence, and in particular the development of creative ideas (Torrance and Galbraith, 1999). As for spelling and punctuation, it could be argued that writers do no longer need to practice these skills, since today's checking programs are capable of cor-

1. lex.page
2. https://github.com/yjs/yjs

recting almost all errors, and the use of dictation would prevent spelling errors from the start. Dictation would also theoretically make it possible to avoid learning how to enter text altogether—either with a pen or a keyboard.

This can be understood as the delegation of certain tasks to a trusted entity that is able to perform these tasks with high quality, so that the user can focus on other aspects. Automation offers consistency and standardization. This aspect of digital transformation in turn enables personalization at scale (Mahlow and Hediger, 2019).

Considering these general possibilities of current digital technology, we can identify some areas in the learning and teaching of writing that indeed require consistent execution and that, on the other hand, would benefit from standardization and personalization. One such area is the provision of feedback, another is the creation of assignments and prompts.

3.1 Feedback

Providing immediate and existing feedback is an accepted strategy for supporting learning. There have been attempts to automate feedback, beginning a hundred years ago with Pressey's mechanical devices for multiple-choice tests (for an overview, see Petrina, 2004), which were developed into drill machines by Skinner (1958). With respect to writing, automatic essay scoring (AES) (Shermis and Burstein, 2002) emerged and generated heated debate, while automatic writing evaluation (AWE), with its focus on providing feedback to improve learners' writing skills, was much more positively received (Fu et al., 2022).

As mentioned earlier, LLM can recognize and produce plausible writing, which is error-free, consistent, and coherent in terms of spelling and grammar. It is also possible to access text output that is specific to a particular language level and for a particular purpose or genre. This, in turn, means that LLM can be confidently used to immediately and consistently evaluate learner input in terms of linguistic features and provide feedback on how texts can be improved to better meet these criteria. For learners, this represents a more acceptable or comfortable experience: Feedback will be consistent but adaptive to the learner—which has been shown to be effective (Leontjev, 2014)—and the machine is not emotionally involved as teachers might be (or as learners might assume their teachers would be)

These scenarios and their corresponding applications already exist, their performance and capabilities will only benefit from the use of increasingly better underlying language technology, in our case LLM. The power of LLM combined with the widely available computing power of laptops, tablets, and even smartphones will also make it possible to design and implement writing support that helps learners think about their texts and discover passages that they can revise and edit themselves. As described above, one task in NLP is summarization. Applications that use LLM are able to provide a summary of a paragraph, section, or even an entire text within milliseconds as an additional feature (e.g., Dang et al., 2022). These summaries contain the most important information of the text that the system detected. The writer can then verify whether or not this summary is consistent with their original communicative intent—discrepancies can be due to both a lack of linguistic competence (the writer does not know how to express something appropriately in a particular language) and the fact that the text does not contain the intended information and needs to be expanded. If other linguistic features are needed, writers could ask the system for examples of how to express something—in which case the system would use phrase/sentence inventories or even create them on the spot—or provide them with machine translation tools if they can competently express themselves in another language. This setting is similar to a coaching session in which the writing consultant tells the writer what they understand after reading their draft, which serves as a starting point for discussion about text structures, cohesion, etc. Again, the writing consultant could use the LLM to offer alternative wording, etc. These scenarios are the more empowering for learners, the more consistent and replicable explanations and suggestions provided by the machine are. Teachers can focus on specific aspects, make connections to previous writing tasks, etc., while being relieved of the cognitively demanding task of finding plausible alternatives to show variations and patterns in writing.

3.2 Create tasks and check solutions

The creation of appropriate tasks and exercises for learners is an important aspect of language teaching, particularly to support internal differentiation, which requires personalized instruction. Automatic exercise generation can help to ensure comparability of individual tasks with respect to topic, difficulty, language level, etc. Taking into account that LLM are able to produce plausible texts meeting specific criteria, they can be used to produce writing prompts, fill-in-the-blank exercises, and even more sophisticated types of tests, which have been shown to be helpful for learners, but very challenging to create (Mahlow et al., 2010; Mahlow and Hess, 2004).

LLM can produce examples of specific phenomena in grammar or vocabulary even in ad hoc situations. There is one obvious caveat concerning examples: phrases and sentences produced are linguistically acceptable and plausible, but they are not real-world examples, they are not authentic. These sentences *could* appear in texts and would be considered valid, but they cannot necessarily be found in any actually *existing* text. In contrast, examples extracted from corpora are instances of actual language use as it has been observed, and one can access the co-text of these phrase and sentences. Here, language teachers have to carefully weigh advantages and disadvantages and decide which approach to apply in which situation.

However, it is important to note that the automatic generation of exercises still requires manual specification, as demonstrated by Heck et al. (2022). The use of language models in a co-creative way can make this process easier. For instance, Zou et al. (2022) showed that language models can be used to automatically generate true/false reading comprehension questions from texts. In both cases, further progress on the power of LLM and hybrid neuro-symbolic approaches—i.e., the combination of LLM with human-understandable knowledge bases—will improve availability and quality in the coming months. We will also see combinations with templates that allow for seamless integration into learning management systems (LMS) and the like—one already existing example is QuestionAid[3].

LLM can provide teachers and learners with detailed instructions for communicative situations in role-plays or writing prompts. Most often, role-plays are used in language learning to practice the use of discourse devices, where it does not matter whether the "facts" used in weather reports or tourist recommendations about museums are actually true. Similarly, writing prompts ask learners to use appropriate linguistic devices to write a text in a particular genre. Chatbots based on LLM—e.g., ChatGPT—can be used as partner for interactive text-adventure games, they can take over the role of a chat partner who also points the writer to challenges and corrects grammar and spelling.

4 Conclusion

In conclusion, the integration of large language models and artificial intelligence in language learning and teaching in general—and in learning how to write in particular—has the potential to enhance the learning experience by providing new and innovative ways of learning. Automation can be used to handle reactive, repetitive, and routine tasks in teaching while freeing up time and attention of teachers for human-human interaction. Language technology-based tools can act as supportive writing buddies, tailored to the learner's language level and needs, making suggestions and helping them to overcome challenges.

The generated output from these models can also be used for reflective purposes, allowing learners to reflect on language-related processes. This approach prioritizes human-human interaction and delegates repetitive and routine tasks to the machine, providing a supportive and efficient learning experience. As language technology continues to evolve, it is important to find a balance between automation and human interaction, leveraging the strengths of both to provide a comprehensive language learning experience.

We are now in a situation, where from the technological point of view (almost) everything seems feasible: pedagogy at large—and language learning and teaching in particular—feels under pressure to react and

3. https://www.question-aid.com

to make good use of the tools available now. As emphasized before, technology always also *reacts* to societal developments. Which puts us in a very promising situation to *drive* transformation in learning and teaching by demanding the design and implementation of tools according to specific needs of learners and teachers alike.

References

[Adlakha et al. 2022] Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries and Siva Reddy. TopiOCQA: Open-domain conversational question answering with topic switching. In *Transactions of the Association for Computational Linguistics*, 10:468–483, 2022. doi:10.1162/tacl_a_00471. URL https://aclanthology.org/2022.tacl-1.27.

[Bowker and Ciro 2019] Lynne Bowker and Jairo Buitrago Ciro. *Machine translation and global research: Towards improved machine translation literacy in the scholarly community*. Emerald Group Publishing, 2019.

[Brown et al. 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan and H. Lin (eds.) *Advances in Neural Information Processing Systems*, vol. 33, 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[Chinkina et al. 2016] Maria Chinkina, Madeeswaran Kannan and Detmar Meurers. Online information retrieval for language learning. In *Proceedings of ACL-2016 System Demonstrations*, 7–12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016. doi:10.18653/v1/P16-4002. URL https://aclanthology.org/P16-4002.

[Chowdhery et al. 2022] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov and Noah Fiedel. Palm: Scaling language modeling with pathways. 2022. doi:10.48550/ARXIV.2204.02311. URL https://arxiv.org/abs/2204.02311.

[Coster and Kauchak 2011] William Coster and David Kauchak. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 665–669. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. URL https://aclanthology.org/P11-2117.

[Council of Europe 2001] Council of Europe. *Common European framework of reference for languages*. Cambridge University Press, 2001.

[Council of Europe 2020] Council of Europe. *Common European framework of reference for languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing, 2020.

[Dale 2020] Robert Dale. Natural language generation: The commercial state of the art in 2020. In *Natural Language Engineering*, 26(4):481–487, 2020. doi:10.1017/S135132492000025X.

[Dale and Viethen 2021] Robert Dale and Jette Viethen. The automated writing assistance landscape in 2021. In *Natural Language Engineering*, 27(4):511–518, 2021. doi:10.1017/S1351324921000164.

[Dang et al. 2022] Hai Dang, Karim Benharrak, Florian Lehmann and Daniel Buschek. Beyond text generation: Supporting writers with continuous automatic text summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22. New York, NY, USA: Association for Computing Machinery, 2022. doi:10.1145/3526113.3545672. URL https://doi.org/10.1145/3526113.3545672.

[Das and Smith 2009] Dipanjan Das and Noah A. Smith. Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 468–476. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. URL https://aclanthology.org/P09-1053.

[Delorme Benites and Lehr 2022] Alice Delorme Benites and Caroline Lehr. Neural machine translation and language teaching: possible implications for the cefr. In *Bulletin suisse de linguistique ap-*

pliquée, 2021(114):47–66, 2022. doi:10.21256/zhaw-24716. URL https://digitalcollection.zhaw.ch/handle/11475/24716.

[Devlin et al. 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[Dowling 1994] Carolyn Dowling. Word processing and the ongoing difficulty of writing. In *Computers and Composition*, 11(3):227–235, 1994. doi:10.1016/8755-4615(94)90015-9.

[Fayol 1999] Michel Fayol. From on-line management problems to strategies in written composition. In Mark Torrance and Gaynor Jeffery (eds.) *The Cognitive Demands of Writing: Processing Capacity and Working Memory Effects in Text Production*, vol. 3 of *Studies in Writing*, 13–23. Amsterdam: Amsterdam University Press, 1999.

[Fernando and Stevenson 2009] Samuel Fernando and Mark Stevenson. A semantic similarity approach to paraphrase detection. In *Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics*, 2009.

[Firth 1957] John Rupert Firth. *Studies in Linguistic Analysis*. Oxford: Blackwell, 1957.

[Fontenelle 2005] Thierry Fontenelle. Dictionnaires et outils de correction linguistique. In *Revue française de linguistique appliquée*, X(2):119–128, 2005. URL http://www.cairn.info/revue-francaise-de-linguistique-appliquee-2005-2-page-119.htm.

[Fu et al. 2022] Qing-Ke Fu, Di Zou, Haoran Xie and Gary Cheng. A review of awe feedback: types, learning outcomes, and implications. In *Computer Assisted Language Learning*, 0(0):1–43, 2022. doi:10.1080/09588221.2022.2033787. URL https://doi.org/10.1080/09588221.2022.2033787.

[Gamon 2010] Michael Gamon. Using Mostly Native Data to Correct Errors in Learners' Writing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 163–171. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. URL http://www.aclweb.org/anthology/N10-1019.

[Ganslandt et al. 2009] Sebastian Ganslandt, Jakob Jörwall and Pierre Nugues. Predictive text entry using syntax and semantics. In *Proceedings of the 11th International Conference on Parsing Technologies*, IWPT '09, 37–48. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009.

[Grabar and Saggion 2022] Natalia Grabar and Horacio Saggion. Evaluation of automatic text simplification: Where are we now, where should we go from here. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, 453–463. Avignon, France: ATALA, 2022. URL https://aclanthology.org/2022.jeptalnrecital-taln.47.

[Heck et al. 2022] Tanja Heck, Detmar Meurers and Florian Nuxoll. Automatic exercise generation to support macro-adaptivity in intelligent language tutoring systems. In *Intelligent CALL, granular systems, and learner data: short papers from EUROCALL 2022*, 162–167. 2022. doi:10.14705/rpnet.2022.61.1452.

[Heidorn et al. 1982] G. E. Heidorn, K. Jensen, L. A. Miller, R. J. Byrd and M. S. Chodorow. The epistle text-critiquing system. In *IBM Systems Journal*, 21(3):305–326, 1982. doi:10.1147/sj.213.0305.

[Heidorn 2000] George E. Heidorn. Intelligent writing assistance: Techniques and applications for the processing of language as text. In Robert Dale, Herman Moisl and Harold Somers (eds.) *Handbook of Natural Language Processing*, 181–207. New York, NY, USA: Marcel Dekker, 2000.

[Hobbs 1984] Jerry R. Hobbs. Building a large knowledge base for a natural language system. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, 283–286. Stroudsburg, PA, USA: Association for Computational Linguistics, 1984. doi:10.3115/980491.980550. URL https://aclanthology.org/P84-1059.

[Hobbs et al. 1982] Jerry R. Hobbs, Donald E. Walker and Robert A. Amsler. Natural language access to structured text. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*. 1982. URL https://aclanthology.org/C82-1020.

[Hulth 2003] Anette Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, 216–223. 2003. URL https://aclanthology.org/W03-1028.

[Jia et al. 2020] Ruipeng Jia, Yanan Cao, Hengzhu Tang, Fang Fang, Cong Cao and Shi Wang. Neural extractive summarization with hierarchical attentive heterogeneous graph network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3622–3631. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.emnlp-main.295. URL https://aclanthology.org/2020.emnlp-main.295.

[Justin 2023] Justin. Lex.page – eliminate writers block using an ai writing assistant. 2023. URL https://goldpenguin.org/blog/introducing-lex-page-ai/.

[Kellogg 2008] Ronald T. Kellogg. Training writing skills: A cognitive developmental perspective. In *Journal of Writing Research*, 1(1):1–26, 2008.

[Kumar and Tong 2019] Shankar Kumar and Simon Tong. Using neural machine translation to correct grammatical faux pas in Google Docs. 2019. URL https://workspace.google.com/blog/ai-and-machine-learning/using-neural-machine-translation-to-correct-grammatical-in-google-docs?hl=en.

[Kwiatkowski et al. 2019] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le and Slav Petrov. Natural questions: A benchmark for question answering research. In *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi:10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026.

[Leontjev 2014] Dmitri Leontjev. The effect of automated adaptive corrective feedback: L2 english questions. In *APPLES: Journal of applied language studies*, 8(2), 2014.

[Levy 1997] Michael Levy. *Computer-assisted language learning: Context and conceptualization*. Oxford University Press, 1997.

[Lichtarge et al. 2019] Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar and Simon Tong. Corpora generation for grammatical error correction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3291–3301. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019. doi:10.18653/v1/N19-1333. URL https://aclanthology.org/N19-1333.

[Macdonald et al. 1982] Nina H. Macdonald, Lawrence T. Frase, Patricia S. Gingrich and Stacey A. Keenan. The Writer's Workbench: Computer aids for text analysis. In *IEEE Transactions on Communication*, 30(1):105–110, 1982. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1095380.

[Mahlow and Dale 2014] Cerstin Mahlow and Robert Dale. Production Media: Writing as Using Tools in Media Convergent Environments. In Eva-Maria Jakobs and Daniel Perrin (eds.) *Handbook of Writing and Text Production*, vol. 10 of *Handbooks of Applied Linguistics*, 209–230. Berlin, Germany: De Gruyter Mouton, 2014.

[Mahlow and Hediger 2019] Cerstin Mahlow and Andreas Hediger. Digital transformation in higher education—buzzword or opportunity? In *ELearn*, 2019(5), 2019. doi:10.1145/3329488/3331171.

[Mahlow and Hediger 2021] Cerstin Mahlow and Andreas Hediger. Education as loosely coupled system of technology and pedagogy. In

On Education. Journal for Research and Debate, 4(12), 2021. doi: 10.17899/on_ed.2021.12.7.

[Mahlow and Hess 2004] Cerstin Mahlow and Michael Hess. Sentence Completion Tests for Training and Assessment in a Computational Linguistics Curriculum. In *COLING-2004 workshop on eLearning for Computational Linguistics and Computational Linguistics for eLearning*, 61–70. 2004.

[Mahlow and Piotrowski 2009] Cerstin Mahlow and Michael Piotrowski. A target-driven evaluation of morphological components for German. In Simon Clematide, Manfred Klenner and Martin Volk (eds.) *Searching Answers – Festschrift in Honour of Michael Hess on the Occasion of his 60th Birthday*, 85–99. Münster: MV-Verlag, 2009.

[Mahlow et al. 2010] Cerstin Mahlow, Michael Piotrowski and Wolfram Fenske. A solid foundation: Why e-assessment should be based on a systematic typology of test items. In Miguel B. Nunes and Maggie McPherson (eds.) *Proceedings of the IADIS International Conference E-Learning 2010, Freiburg, Germany, July 26 – 29, 2010*, vol. 2, 47–51. IADIS Press, 2010.

[Meurers et al. 2010] Detmar Meurers, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf and Niels Ott. 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*, 10–18. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. URL https://aclanthology.org/W10-1002.

[Norman 1993] Donald A. Norman. *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine*. Boston, MA, USA: Addison-Wesley, 1993.

[Peng et al. 2016] Haoruo Peng, Yangqiu Song and Dan Roth. Event detection and co-reference with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 392–402. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016. doi:10.18653/v1/D16-1038. URL https://aclanthology.org/D16-1038.

[Petrina 2004] Stephen Petrina. Sidney Pressey and the automation of education, 1924–1934. In *Technology and Culture*, 45(2):305–330, 2004. URL http://www.jstor.org/stable/40060743.

[Piolat et al. 2004] Annie Piolat, Jean-Yves Roussey, Thierry Olive and Murielle Amada. Processing time and cognitive effort in revision: Effects of error type and of working memory capacity. In Linda Allal, Lucile Chanquoy and Pierre Largy (eds.) *Revision. Cognitive and instructional processes*, vol. 13 of *Studies in Writing*, 21–38. Boston, Dordrecht, London: Kluwer, 2004.

[Piotrowski 2012] Michael Piotrowski. *Natural Language Processing* *for Historical Texts*. San Rafael, CA, USA: Morgan & Claypool, 2012. doi:10.2200/S00436ED1V01Y201207HLT017.

[Rei and Cummins 2016] Marek Rei and Ronan Cummins. Sentence similarity measures for fine-grained estimation of topical relevance in learner essays. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 283–288. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016. doi:10.18653/v1/W16-0533. URL https://aclanthology.org/W16-0533.

[Reiter and Dale 2000] Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge, UK: Cambridge University Press, 2000. URL http://www.worldcat.org/isbn/052102451X.

[Roesiger and Kuhn 2016] Ina Roesiger and Jonas Kuhn. IMS HotCoref DE: A data-driven co-reference resolver for German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 155–160. Paris, France: European Language Resources Association (ELRA), 2016. URL https://aclanthology.org/L16-1024.

[Schneider et al. 2022] Adela Schneider, Andreas Madsack, Johanna Heininger, Ching-Yi Chen and Robert Weißgraeber. Data-to-text systems as writing environment. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, 1–10. Dublin, Ireland: Association for Computational Linguistics, 2022. doi:10.18653/v1/2022.in2writing-1.1. URL https://aclanthology.org/2022.in2writing-1.1.

[Schwitter et al. 2000] Rolf Schwitter, Diego Molla, Rachel Fournier and Michael Hess. Answer extraction towards better evaluations of NLP systems. In *ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*. 2000. URL https://aclanthology.org/W00-0604.

[Seonwoo et al. 2020] Yeon Seonwoo, Ji-Hoon Kim, Jung-Woo Ha and Alice Oh. Context-aware answer extraction in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2418–2428. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.emnlp-main.189. URL https://aclanthology.org/2020.emnlp-main.189.

[Sharples 2022] Mike Sharples. Automated Essay Writing: An AIED Opinion. In *International Journal of Artificial Intelligence in Education*, 32(4):1119–1126, 2022. doi:10.1007/s40593-022-00300-7. URL https://doi.org/10.1007/s40593-022-00300-7.

[Shermis and Burstein 2002] Mark D. Shermis and Jill C. Burstein (eds.) *Automated Essay Scoring*. Routledge, 2002. doi:10.4324/9781410606860.

[Sjöblom et al. 2021] Eetu Sjöblom, Mathias Creutz and Teemu

Vahtola. Grammatical error generation based on translated fragments. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 398–403. Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, 2021. URL https://aclanthology.org/2021.nodalida-main.44.

[Skinner 1958] B. F. Skinner. Teaching machines. In *Science*, 128(3330):969–977, 1958. URL http://www.jstor.org/stable/1755240.

[Somers 2005] Harold Somers. Round-trip translation: What is it good for? In *Proceedings of the Australasian Language Technology Workshop 2005*, 127–133. Sydney, Australia, 2005. URL https://aclanthology.org/U05-1019.

[Tay et al. 2022] Yi Tay, Mostafa Dehghani, Dara Bahri and Donald Metzler. Efficient transformers: A survey. In *ACM Computing Surveys*, 55(6), 2022. doi:10.1145/3530811. URL https://doi.org/10.1145/3530811.

[Torrance and Galbraith 1999] Mark Torrance and David Galbraith (eds.) *Knowing What to Write: Conceptual Processes in Text Production*, vol. 4 of *Studies in Writing*. Amsterdam University Press, 1 ed., 1999. URL http://www.worldcat.org/isbn/9053563075.

[Tschichold et al. 1997] Cornelia Tschichold, Franck Bodmer, Etienne Cornu, Francois Grosjean, Lysiane Grosjean, Natalie Kubler, Nicolas Lewy and Corinne Tschumi. Developing a new grammar checker for English as a second language. In *From Research to Commercial Applications: Making NLP Work in Practice*. 1997. URL https://aclanthology.org/W97-0902.

[Wang et al. 2021] Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang and Kewei Tu. Improving named entity recognition by external context retrieving and cooperative learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1800–1812. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021. doi:10.18653/v1/2021.acl-long.142. URL https://aclanthology.org/2021.acl-long.142.

[Wang et al. 2020] Zhengjue Wang, Zhibin Duan, Hao Zhang, Chaojie Wang, Long Tian, Bo Chen and Mingyuan Zhou. Friendly topic assistant for transformer based abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 485–497. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.emnlp-main.35. URL https://aclanthology.org/2020.emnlp-main.35.

[Weick 1976] Karl E. Weick. Educational organizations as loosely coupled systems. In *Administrative Science Quarterly*, 21(1):1–19, 1976. URL http://www.jstor.org/stable/2391875.

[Yuan et al. 2022] Ann Yuan, Andy Coenen, Emily Reif and Daphne Ippolito. Wordcraft: Story writing with large language models. In *27th International Conference on Intelligent User Interfaces*, IUI '22, 841–852. New York, NY, USA: Association for Computing Machinery, 2022. doi:10.1145/3490099.3511105. URL https://doi.org/10.1145/3490099.3511105.

[Yuan and Bryant 2021] Zheng Yuan and Christopher Bryant. Document-level grammatical error correction. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, 75–84. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.bea-1.8.

[Zhang and Zou 2022] Ruofei Zhang and Di Zou. Types, purposes, and effectiveness of state-of-the-art technologies for second and foreign language learning. In *Computer Assisted Language Learning*, 35(4):696–742, 2022. doi:10.1080/09588221.2020.1744666. URL https://doi.org/10.1080/09588221.2020.1744666.

[Zou et al. 2022] Bowei Zou, Pengfei Li, Liangming Pan and Ai Ti Aw. Automatic true/false question generation for educational purpose. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, 61–70. Stroudsburg, PA, USA: Association for Computational Linguistics, 2022. doi:10.18653/v1/2022.bea-1.10. URL https://aclanthology.org/2022.bea-1.10.