

EMERGING TECHNOLOGIES



Partnering with AI: Intelligent writing assistance and instructed language learning

Robert Godwin-Jones, Virginia Commonwealth University

Abstract

In recent years, advances in artificial intelligence (AI) have led to significantly improved, or in some cases, completely new digital tools for writing. Systems for writing assessment and assistance based on automated writing evaluation (AWE) have been available for some time. That is the case for machine translation as well. More recent are synchronous feedback tools, such as Grammarly. That tool incorporates, as do others, predictive text technology, supplying automated sentence completion. Emerging writing assistance goes further, generating an entire text in response to a brief prompt. That capacity, along with significantly improved performance of both automated feedback systems and machine translation, is enabled through advances in AI, built on ever larger datasets and deep machine learning. While they differ in interface, functionality, and target audience, the available and emerging set of intelligent writing tools can be used to help learners improve the quality of their written texts. However, their use in instructional language learning has in some cases been controversial. In this column, we will be examining AI-enabled writing tools, reviewing the findings from research studies, and discussing their use in instructional settings. When integrated into writing instruction and practice, these digital tools have been found to offer significant benefits to both students and teachers. Teacher mediation aids learners in becoming informed consumers of language technology, as well as helping them to gain meta-linguistic knowledge. For researchers, intelligent writing tool use is optimally analyzed from a broad ecological perspective that examines the dynamic interplay of learner, software, and instructional environment.

Keywords: Automatic Corrective Feedback, Machine Translation, AI Tools, Second Language Writing

Language(s) Learned in This Study: English

APA Citation: Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology*, 26(2), 5–24. <http://doi.org/10125/73474>

Introduction

Advances in artificial intelligence (AI) in recent years have led to the availability of tools and services for writing that have radically reshaped the process of authoring and editing. Those developments are having a profound effect in education generally and, as we will explore in this column, on language learning and teaching as well. Tools for automatic writing evaluation (AWE), originally designed for essay assessment, have moved beyond conventional spelling and grammar checking to offer extensive help in identifying writing problems and suggesting areas needing revision. More recently, synchronous corrective feedback has become available in stand-alone tools or has been integrated into word processors (Koltovskaia, 2020; Ranalli 2018; Ranalli & Yamashita, 2022). As with editing tools, machine translation (MT) has improved exponentially in the last few years and now provides high-quality and reliable translations in many language combinations (Jolley & Maimone, 2022). The most dramatic development is the AI-enabled incorporation of auto-completion of phrases into text editors and online writing venues, as well as suggestions for alternative wording (Dale, 2021). Advances in all these areas continue through AI systems collecting massive sets of data and using deep machine learning based on artificial neural networks. This has led to significant breakthroughs in *natural language processing* (NLP), that is, turning text into structured data,

and therefore also advances in *natural language understanding*, the ability of AI systems to extract meaning from texts.

While digital writing tools have become widely used in everyday life and work environments, their use in instructed language learning has been controversial. As intelligent writing assistance moves beyond grammar and vocabulary help, to supply well-crafted translations, rewrites, or even the option to author extensive chunks of text, language educators may be concerned about the authenticity of submitted student writing, as well as the absence of learning accompanying copy and paste assignment completion. Some differentiation may come into play, with AWE tools being accepted but MT banned. Those distinctions are, however, becoming increasingly blurred, as AI systems built on *large language models*, such as GPT-3, enable a host of language services, from translation to essay writing (Dale, 2021). Moreover, the generated texts by these systems are unique, on-demand creations, making them undetectable by anti-plagiarism tools (Eaton et al., 2021).

Given that developments in AI will inevitably lead to ever improving performance in a variety of language tasks, it may be time to learn to live with the reality of students' access to advanced writing assistance and find ways to provide appropriate guidance (Carvalho et al., 2022; Hellmich & Vinall, 2021; Otsuki, 2020). The mediating presence of a knowledgeable teacher can help students discover the strengths and weaknesses of digital tools (Pellet & Myers, 2022). Critical analyses of AI-powered writing assistance can provide linguistic insights deriving from seeing the limitations in AI's understanding of human language in all its contextual and pragmatic complexity (Vinall & Hellmich, 2022). Through the interaction of software, learner interactions, and teacher mediation, a complex learning environment is created, which is best understood from a broader, ecological perspective, often missing in research studies in these areas (Hellmich & Vinall, 2021).

Current and Emerging AI-Enabled Writing Assistance

Automatic writing evaluation systems, such as [Criterion](#), [MY Access!](#), or [Pigai](#) are used principally in academic settings. They have been available for some time, so a base of research studies is available. Synchronous text editors are more recent. Tools such as [Grammarly](#) or [ProWritingAid](#) are widely used in educational, professional, and everyday environments. They offer automated written corrective feedback (AWCF), a descriptor beginning to be used in emerging research on their use (Guo et al., 2021; Koltovskaia, 2020; Ranalli 2018; Ranalli & Yamashita, 2022). Translation services such as [Google Translate](#) have improved dramatically in recent years and are now available in a variety of formats and on many different devices. The newest addition to intelligent writing assistance are automatic text generators, which, independent of the grammatical accuracy of a text string, suggest wording improvements ([Google Compose](#)) or even generate entire texts, when given a topic or prompt ([GPT-3](#)). We will examine each in turn.

Automated Writing Evaluation

AWE is widely used today, in both first language (L1) and second language (L2) educational environments and at all levels of instruction, from primary schools to universities. Writing is a complex endeavor, combining low level (spelling, mechanics) and higher-level skills related to content organization, logical sequencing, and stylistic appropriateness. Writing in an L2 presents its own special set of challenges, deriving from possible deficiencies in terms of lexical, syntactic, pragmatic, or rhetorical knowledge. Providing corrective feedback (CF) that is useful to writers is consequently a difficult task. How to provide that feedback has been a contentious issue in research on L2 writing (Bitchener & Ferris, 2012). Despite voices to the contrary (Truscott, 2007), there is a consensus that CF, if appropriately applied, is beneficial. (R. Ellis, 2016; Kang & Han, 2015; Link et al., 2020). However, there are so many contextual variables at play that generalizations are on shaky ground (R. Ellis, 2016; Hyland & Hyland, 2006).

For teachers, providing CF on student writing can be an extremely time-consuming task. Giving individually tailored feedback can be daunting, depending on class size. Compared to human readers and raters, AWE systems hold the promise of providing fast and consistent CF. Depending on the specific AWE system being used, the resources offered can be voluminous compared to teacher-supplied feedback, as

auxiliary writing resources are often integrated (Grimes & Warschauer, 2010; Hussein et al., 2019). A central element of good writing is the ability to edit and improve, with a final version being often quite different from the first draft (Flower & Hayes, 1981). Crucial in that process for beginning writers is feedback for each successive draft, leading to a series of revision cycles. Automated systems track revisions and offer CF for each rewriting, while maintaining data on changes, a task likely difficult for a teacher. In that way, an AWE system can identify areas of improvement, thus providing insight for the writer into characteristics of good writing, an important step in developing metacognitive knowledge about writing and language learning. Studies have shown that integration of AWE into writing instruction can supply a useful framework for deliberate practice (Palermo & Wilson, 2020). The improvements writers may see as a result can have a motivating effect (Camacho et al., 2021; Nunes et al., 2021).

Studies on the effectiveness of AWE vary considerably, showing inconsistent results (Link et al., 2020), but with a growing consensus that if implemented in a contextually appropriate manner, AWE programs can have a positive effect on the quality of student writing (Grimes & Warschauer, 2010; Nunes et al., 2021). Here again, generalizations are misleading. Research has shown how many variables are at play, including the nature of the instructional support, the attitudes and activities of the teacher towards the use of AWE, and the amount of practice afforded students (Nunes et al., 2021). One of the more telling variables in terms of effectiveness are student characteristics such as proficiency level, the writing/editing stage at which AWE is used, and student beliefs about the validity and usefulness of AWE (Grimes & Warschauer, 2010). Many studies have shown that AWE feedback is most useful at the early stages of language learning (Y. Huang & Wilson, 2021; Ranalli et al., 2017). The revisions most commonly made are in grammatical accuracy and lexical appropriateness rather than in content or structure (Bridgeman & Ramineni, 2017; Burstein et al., 2016). Furthermore, while individual texts may show improvement through AWE, there is often little evidence of sustained improvement or transfer to L2 learning generally (Y. Huang & Wilson, 2021). Pointing to that fact, Ranalli (2021) remarks that “AWE tools have fallen short of expectations” (p. 2). Instead of improving writing skills and L2 development, he asserts, they tend to be used by students in a proofreading orientation.

Automatic writing evaluation systems provide little assistance with improving areas such as argumentation strength, discourse coherence, or organization. That is due to the fact that these AI systems do not truly understand human language (Grimes & Warschauer, 2010; Y. Huang & Wilson, 2021). Natural language understanding in AWE-based systems is built on leveraging statistical analysis, large data collection, and machine learning to determine the likelihood of text sequencing. As a result of the parameters set in AWE systems, they tend not to value originality or creativity, but rather language mechanics, often privileging length and syntactical complexity over succinctness and clarity, or more intangible, humanistic features (Bridgeman & Ramineni, 2017; Y. Huang & Wilson, 2021). The emphasis on formal correctness translates into valuing formulaic conventions and surface-level accuracy over inventiveness (Stevenson, 2016). Another concern often raised in the use of AWE is the effect writing for a machine may have on the understanding of the nature of writing (Grimes & Warschauer, 2010). Writing does not occur in a vacuum, but rather is fundamentally a social practice; we write for specific purposes and for a particular readership. In using AWE there is the risk of conveying the impression to students that writing is an academic exercise, not a lifelong skill or a vital component of language learning (Hegelheimer & Lee, 2012; Link et al., 2020; Saricaoglu, 2019).

Mainstream AWE systems (i.e., Criterion, MY Access!) are calibrated for academic writing, with the essay being the genre best adapted to AWE analysis and feedback. It would enhance the utility of AWE systems were they more capable of dealing with a variety of genres, especially considering the proliferation of writing modes available online. As genres are associated with identifiable conventions and practices, it does not seem unreasonable that AWE systems could be trained to work within those defined genre parameters. The limitations in the kind of writing AWE can analyze extends to multimedia integration as well as to collaborative writing, the hallmarks of writing in today’s digital world (Godwin-Jones, 2018). In the model of traditional literacy built into AWE systems, their use “at face value seems to negate much of what we celebrate as liberating in new media use” (Warschauer & Ware, 2006, p. 176). One might add to that

concern the inability to deal with multilingualism, a feature of writing increasingly identified as wide-spread, particularly in online environments (Ortega, 2017).

Research on AWE is increasingly recognizing these limitations as problematic. Published studies themselves have been seen as partly responsible for the lack of progress pushing AWE into new areas of functionality (Hibert, 2019). One of the explanations for that situation has been the fact that a good number of researchers are themselves affiliated with the companies selling the products (Grimes & Warschauer, 2010). This influenced early studies which emphasize the reliability of the systems and their alignment with human raters. It also reflects the initial usage of AWE systems in formal assessment environments, in which the kind of writing required is tightly prescribed. Rather than taking company marketing materials at face value, researchers have been encouraged to validate claims so as to make software companies more accountable (Chapelle et al., 2015; Ranalli, 2021). Systematic, critical studies could improve the utility of AWE research. The contradictory claims in studies, along with missing details on the context of use and sometimes absence of control groups, make it difficult to draw any concrete conclusions about the use of AWE. Hibert (2019) laments the “theoretically and methodologically fragmented” nature of AWE research (p. 209). Given the questionable validity of company claims, as well as the lack of information on how their AI systems are configured, independent research studies are all the more important (Ranalli & Yamashita, 2022). Hibert’s (2019) systematic literature review points to the fact that AWE studies generally fail to take advantage of the data collection capabilities of AWE software. Methods used in data mining and clustering techniques could be effectively used in AWE studies (Godwin-Jones, 2021; Peng et al., 2020; Warschauer et al., 2019). This could enable identifying individual and contextual variables that influence the effectiveness of AWE feedback (Ranalli, 2021).

AWCF Tools (Text Editors Supplying Synchronous Feedback)

A set of tools similar to AWE are services which supply real-time automated written corrective feedback (AWCF), an under-investigated area in Computer-Assisted Language Learning (CALL; Ranalli, 2018; Ranalli & Yamashita, 2022). While AWE systems supply feedback after a text is written, AWCF tools such as Grammarly supply corrections and suggestions continuously and concurrently as a text is in the process of being written. Dale and Viethen (2021) supply a list of writing tools which operate in this way. In addition to Grammarly, well-known products include [Ginger](#) and ProWritingAid. Automated written corrective feedback tools do not attempt to provide feedback on organizational or structural factors in texts, but focus exclusively on lower-level writing issues, in particular grammatical and lexical errors. Whereas access to AWE systems is supplied by a web-based portal, Grammarly and similar tools are available in a variety of ways. Grammarly, for example, works as a standalone tool or is integrated into existing writing tools, such as Microsoft Word or Google Docs. Most recently, it is available as a web browser extension or as a virtual keyboard in smartphones. Given how widely Grammarly in particular is used (and how much it is hyped in marketing materials), more studies would be welcome, as it does indeed represent a new and distinct genre of writing-support technology (Dale & Viethen, 2021; Ranalli & Yamashita, 2022).

A number of studies have examined the use of Grammarly in English as a Foreign Language (EFL) settings. Several studies found Grammarly provided mostly accurate feedback, but also reported many false positives and false negatives (Dembsey, 2017; Dodigovic & Tovmasyan, 2021; John & Woll, 2020; R. O’Neill & Russell, 2020). As is the case with AWE feedback, there have been complaints that Grammarly’s feedback was often too repetitious (Dembsey, 2017) or too voluminous (R. O’Neill & Russell, 2019, 2020). An additional concern was how Grammarly words its feedback. In attempting to be easily understandable, it avoids explanations that might be too technical; in the process there is a danger of oversimplification (Dodigovic & Tovmasyan, 2021). On the other hand, other studies reported that the feedback from Grammarly was sometimes too technical (Dembsey, 2017; R. O’Neill & Russell, 2019). How CF is processed is complex, but one important factor is the learner’s proficiency level and knowledge of grammatical terminology. Zheng and Yu (2018) point out that limited linguistic knowledge can prevent students from fully processing feedback, thus preventing them from taking advantage of the information to make further revisions.

Studies have been inconsistent in their recommendations for how Grammarly should be used. While Dizon and Gayed (2021) recommend its use among beginners, Koltovskaia (2020) has the opposite view:

Based on the study's findings, students with low language proficiency may not be able to utilize Grammarly effectively as their lack of linguistic competence can prevent them from adequately understanding AWCF. Therefore, the use of Grammarly is recommended for students with more advanced English proficiency. (p. 11)

Given the concerns over accuracy and the nature of feedback, there is a consensus in studies of Grammarly in instructed language learning that its use should be a "starting point" (R. O'Neill & Russell, 2020, p. 102) and used in conjunction with teacher feedback and/or other monitoring. Its use as a standalone tool in instructed settings is not recommended (Dodigovic & Tovmasyan, 2021; H. W. Huang et al., 2020; John & Woll, 2020).

Studies found positive aspects of using Grammarly include its versatility in access options, its speed of feedback, and its free cost (there is also a paid premium version; H. W. Huang et al., 2020; Nova, 2018). Results in several studies showed improved writing after using Grammarly (Ghufron, 2019; Karyuatry et al., 2018; R. O'Neill & Russell, 2019). There were also reported gains in lexical diversity from its use (Dizon & Gayed, 2021). The Grammarly error categorization feature was found to be helpful (R. O'Neill & Russell, 2020). Whereas L2 texts may have so many errors that it is difficult for human correctors to categorize the exact nature of the problem, Grammarly's algorithmic analysis provides helpful error classification, producing "personalised, targeted feedback" (R. O'Neill & Russell, 2020, p. 102). The feature that identifies textual borrowings was found to be helpful in detecting plagiarism (Dodigovic et al., 2016).

Nazari et al. (2021) found that Grammarly helped students in developing self-regulation skills due to its multiplicity of access options and ease of use, as well as the opportunity to self-correct before summative assessments. Given its widespread use outside of formal education, it would be welcome to see studies of Grammarly use independent of instruction, particularly longitudinal studies that could show whether there is not only improvement in the writing task at hand, but also longer-term positive effects on writing skills. Also useful, as discussed in Ranalli and Yamashita (2022), would be studies which evaluate the effects of using synchronous writing tools on L2 students' self-initiated revisions, given how important that ability is as a writing skill. Viewed from an L2 learning perspective, writing tools which supply synchronous feedback offer a mixed bag (Ranalli & Yamashita, 2022). Immediate CF has been shown to be beneficial in terms of increases in grammatical accuracy, for example, when supplied by teachers or through chat interactions (Arroyo & Yilmaz, 2018). That has been applied to CF from AWE systems as well (Conijn et al., 2019). However, receiving continuous feedback while writing may be a source of distraction or frustration for writers (Ranalli & Yamashita, 2022). Constant messaging while writing may overload working memory and prevent writers from being able to process appropriately the feedback received. That is all the more likely when writing in an L2.

Tools like Grammarly do not differentiate in evaluating L1 or L2 texts. From a performance perspective, however, texts written by L2 learners pose special problems for tools such as Grammarly, as they tend to contain more complex, unpredictable errors, compared to L1 texts. That is the case, for example, for the most common error, misspellings (Ranalli & Yamashita, 2022). In dealing with texts containing a more complicated set of errors, AWCF systems may experience some delay in feedback. That is also due to the fact that the parsing and output is being done in the cloud, rather than locally (Ranalli & Yamashita, 2022). There is no setting in these tools that allows for varied feedback depending on user characteristics, such as being an L2 learner. If that were available, it could improve processing speed for L1 users, or in the case of L2 users, possibly signal the system to use a hybrid approach that adds access to a learner corpus to the normal machine learning methods (Leacock et al., 2014; Zomer & Frankenberg-Garcia, 2021). That might additionally add the option of differentiation in the nature and specificity of the CF provided. If users and/or instructors had such an option, that could accommodate varied CF for different kinds of writing tasks (Ranalli, 2018).

The feedback from Grammarly is specific rather than generic, explicitly addressing errors. While this kind of CF has been shown to be more effective in terms of revising (R. Ellis, 2016), continuous access to CF may run the “risk of reinforcing students’ low-level focus” (Ranalli & Yamashita, 2022, p. 14). The type of writing task at hand will play an important role in the usefulness and effectiveness of CF (Ranalli, 2018). The case can be made that while specific and explicit CF likely improves the quality of writing, implicit feedback may lead to longer-term L2 gains. S. Li’s (2010) meta-analysis of CF found that implicit, more generic feedback proved to be more effective in learning gains as measured by post-tests completed long after the instruction had taken place. Because of that finding, R. Ellis (2016) comments that “perhaps, then, the effects of implicit CF, like a good wine, need time to mature.” Whether that is the case as well for computer-supplied rather than teacher-supplied CF would be a good topic for longitudinal studies.

Machine Translation

One of the difficult issues in evaluating text editors is that companies reveal little about how their systems work and particularly what use is made of AI. One can assume, based on their functionality, that advanced machine learning is at their core (Dale & Viethen, 2021; Hussein et al., 2019). In the case of machine translation, proprietary methods are also common, although somewhat more technical information has become publicly available. We know, for example, that Google Translate transitioned in 2016 from primarily using statistical methods to machine learning based on artificial neural networks. This is an approach which requires less human programming, extracting patterns on its own from raw data. Importantly, it improves performance significantly in both speed and accuracy (Lewis-Kraus, 2016). The updated version can also handle longer texts better, as well as deal more effectively with unknown words (Jolley & Maimone, 2022). Its effectiveness, however, still relies on a sufficient store of data in the targeted languages.

The use of MT in educational settings has been controversial. Among digital language tools, MT software “stands out as particularly polemical” (Hellmich & Vinall, 2021, p. 1). It is common for language teachers to ban its use (Klekovkina & Denié-Higney, 2022). While many of those who forbid it consider its use “cheating,” others may feel its widespread use may spell the end of the need for FL teachers, as Crossley (2018) asserts. The idea that students simply copy and paste from Google Translate to complete assignments—and thereby fail to engage with the target language—is overly simplistic, as research into its use has shown (Jolley & Maimone, 2022; Klekovkina & Denié-Higney, 2022; Vinall & Hellmich, 2022). Typically, L2 students use MT to look up words or phrases, rather than have the service translate whole bodies of text. When surveyed as to why they used Google Translate, students cite the convenience and speed it provides (Fredholm, 2015) as well as the fact of it being free and available from multiple devices including mobile phones (Sukkhwan, 2014). Recent studies (Enriquez Raído & Sánchez-Torrón, 2020; Vinall & Hellmich, 2021) found that virtually all students surveyed used MT (mostly Google Translate) for learning tasks in instructed language learning. Along with Grammarly, Google Translate seems to have become a ubiquitous helper for students writing in an L2.

A number of studies have examined the use of MT in second language acquisition (SLA), especially in writing tasks (see Jolley & Maimone, 2022, for an overview). Most studies involve having students use MT for their first draft or asking them to compare their first draft with a machine-translated version. Several studies by Fredholm (2014, 2015, 2019) and by E. M. O’Neill (2016, 2019) show improved writing quality through the use of MT integrated into learning tasks. However, as is the case with AWE use, the studies generally examine the quality of writing samples, but not the larger question of transfer to general writing ability or to gains in grammatical or lexical knowledge. As is also the case with AWE use, multiple studies show that teacher mediation in the form of MT training makes a significant difference in effectiveness (Fredholm, 2019; S.-M. Lee, 2020; E. M. O’Neill, 2019). A study by H. Zhang and Torres-Hostench (2022) has shown that providing training in post-editing machine-translated texts can be helpful in correcting raw MT output and in gaining insight into MT limitations. Developing the skills to post-edit requires focused attention and advanced reading ability, valuable both in language learning and in professional translation.

One of the potential downsides of encouraging the use of Google Translate by students is that it could “lead

to reductionist perceptions of language” (Pellet & Myers, 2022, p. 57) among both learners and teachers. That is, it may encourage the notion that human language is a discrete code that can be quite simply reencoded based on a one-to-one correspondence between languages (Hellmich & Vinall, 2021). Ryu et al. (2022) lament that students may see Google Translate as an “answer key,” pointing to a simplistic view of language. Such an instrumentalist view of language “fails to acknowledge the richness and complexity of human interaction, identity, and culture” (Urlaub & Dessein, 2022, p. 57). MT captures the semantic dimension of language but misses the nuances and the context-dependent, contingent essence of human communication. For students, accuracy in language use (as is offered by the use of text editors’ CF or machine translations) may be seen as the primary goal of language learning. That is, after all, how they have been conditioned by the nature of most L2 formal assessment.

The implication of the use of MT and other intelligent language writing assistance for views on the nature of language may lead to a reconsideration of the goals of language education generally (Vinall & Hellmich, 2022). Given the availability of advanced writing and translation tools, linguistic accuracy “can no longer be viewed as a synonym of learning and excellence” (Klekovkina & Denié-Higney, 2022, p. 107). If MT were capable of capturing the total essence of a language, that will reduce language to the instrumental role laid out in Crossley (2018). He asserts that MT, available synchronously in mobile devices, may spell the end of FL instruction, as “society may come to see learning a new language in an FL environment as an antiquated endeavor akin to using a horse for transportation” (p. 547). In his view, tourists using MT to have a “direct conversation with local citizens” will have a “more authentic experience” in the target culture “that will allow for a greater understanding of cultural, economic, and social norms” (p. 545). Beyond the interpersonal awkwardness and technological difficulty of using a mobile translator as an intermediary, there is a whole dimension of nonverbal and affective communication that is missing in this kind of “authentic encounter.” Machine Translation will suffice for transactional language needs for tourists (ordering food, checking into a hotel), but will hardly be a substitute for genuine person-to-person encounters (Godwin-Jones, 2019).

One of the potential positive aspects of introducing MT into the classroom is the benefit of leveraging students’ L1 in the learning process and contrasting its usage patterns with the L2. This aligns with the multilingual turn in SLA studies (Ortega, 2013) and with the goal of having learners be able to function comfortably among multiple languages (MLA, 2007). In examining how differently lexico-grammatical constructions work in different languages, students are led away from a conventional separation of grammar and vocabulary towards a usage-based language model emphasizing patterns and frequently combined chunks of language (Tomasello, 2003). In that way, examination of machine-translated texts can counter the idea that language is based on a set of grammar rules (N. C. Ellis, 2017). There is the potential in working with MT for students to gain insights into the formulaic aspects of language in its emphasis on statistical probability. These insights are not automatic gains, particularly if students are following the common usage pattern of using MT for lexical assistance. Teacher mediation of the kind illustrated in Pellet and Myers (2022) will likely be needed. Helpful as well in moving in this direction would be working with corpora, as that can be instrumental in developing a usage-based understanding of language (N. C. Ellis, 2017; Godwin-Jones, 2017).

Automatic Text Generation

One of the primary reasons for improved performance in writing and translation tools is the growing power of *large language models*, used as a backend to natural language understanding in virtually all language technology today. To understand the ability of AI systems to write on their own, it is useful to know how predictive text technology is made possible through the emergence of a new generation of language models. Language modeling at its most basic involves predicting the next word in a text given the previous words (Ruder, 2018). Large language models are AI systems built from huge collections of data analyzed by machine learning and resulting in an ability to deal with human language in increasingly effective ways. Language models used in NLP are not based on grammatical knowledge of the language, but rather are built upon artificial and mathematical modeling of language data. Models based on statistical analysis go

back to the 1940s and the N-gram model (M. Zhang & Li, 2021). Machine learning began to arrive in the 1980s and neural networks were first used in the word2vec model in 2013. In that model, and in current implementations, words are represented by a string of numbers called a vector, obtained through determining how a word co-occurs with other words in context (Ferrone & Zanzotto, 2020). Vectors for phrases and sentences are obtained by combining vectors for words. Building on these developments, current language models use deep machine learning consisting of multiple layers of mathematical computations operating in parallel. Crunching the numbers requires arrays of powerful computers running complex algorithms. The process results in more reliable vector representation and better probability calculations of words, phrases, sentences, and paragraphs (M. Zhang & Li, 2021). Crucially, the resulting model has superior contextual awareness. In 2017, transformer language models were introduced (M. Zhang & Li, 2021), with the best-known being BERT, developed by Google. This approach uses non-linear processing of text chunks, based on the concept of “attention,” using weighted semantic importance over syntactic order. This has proven to be a highly effective architecture for NLP (Ferrone & Zanzotto, 2020). A recent study used BERT to assign Common European Framework of Reference for Languages (CEFR) levels automatically to L2 learner texts (Schmalz & Brutti, 2021).

Developing language models has traditionally involved the application of supervised learning, meaning that the systems were trained on sets of labeled data. Those trained models could not generalize to other tasks. A goal in AI development has been to create a language model based on unsupervised learning, that is, using raw data rather than having to take the slower and more costly step of analyzing and tagging data (Ruder, 2018). Such a system is considered to be “pre-trained,” so can be used in a variety of domains without needing to be supplied beforehand with tagged data for that specific domain. OpenAI developed such a system with its GPT series, which stands for generative pre-trained language model. The latest version of this model is GPT-3, released in 2021. The improved performance of the GPT model is based on scale effect, that is, by increasing massively the amount of data collected and analyzed (M. Zhang & Li, 2021).

GPT-3 represents a giant leap forward in NLP, but in terms of AI-enabled writing assistance, it follows trends already underway. Writing tools have in recent years moved in the direction of automatic text generation, based on advances in language modeling. Intelligent text generators herald perhaps the “biggest transformation of writing since word processors” (Floridi & Chiriatti, 2020, p. 691). It holds true at any rate that they provide writing assistance not at all available a few years ago. This is a feature Google calls *smart compose* that was added to its writing and search products in 2018. It features sentence completion suggestions and the ability to tailor autocompletion options somewhat to context. In its email client, for example, wording and auto-complete suggestions are based not only on what you have just typed, but also on the content and sender of the message to which you are replying. Microsoft Editor, available built into Microsoft Office products and as a web browser extension, incorporates the familiar spelling and grammar checking, but adds text prediction and rewriting options. Grammarly now also has predictive-text capabilities (Dizon & Gayed, 2021). According to Dale (2021), auto-completion has become a must-have feature in writing tools. The fact that in these products rewriting prompts are provided whether or not errors are detected means that they effectively provide automatic rewriting, at least at the sentence level.

Until recently, autocompletion options were limited to words and phrases and hence stood a good chance of being on target. Dale (2020) calls this “short leash” (p. 485) text generation. GPT-3, and its likely more powerful successors have changed the game, however. GPT-3 based writing tools are capable of generating extended discourse in a variety of genres in virtually any content area (Godwin-Jones, 2021). Generated texts are often indistinguishable from those written by humans, particularly in terms of flow and textual coherence. In the words of Dale and Viethen (2021), the generated texts are “eerily convincing” (p. 516). GPT-3 text generation does not require training in either content or genre, but just a short prompt and it is off and running. The way one generates text is to describe the task briefly in English or to supply an example. Even though the overwhelming bulk of the data used by GPT-3 is in English, enough data was collected in other languages that it is capable of generating texts in a variety of languages. After being supplied with the first few lines of a Dante sonnet, for example, GPT-3 continued the poem with stanzas in Italian using

the appropriate sonnet form (Floridi & Chiriatti, 2020).

GPT-3 can compose poetry, write computer code, translate text, do summarizations, correct grammar, power chatbots and much more (Dale, 2021). Rather than helping you to write, it writes for you. This raises a whole series of issues, particularly if such text generators are used in educational environments. One can assume that, as is the case with MT, intelligent text generators, when widely available, will be used by students across all disciplines (Eaton et al., 2021). This raises troubling questions of authenticity, creativity, and attribution. Authorship is shared between humans and machines, making them in essence co-creators. Language educators will face the challenge of assessing such written work, finding appropriate ways to assign credit. As is the case with MT, writing teachers will need to be creative in finding tasks which combine automatic text generation and student effort.

Automatic writing evaluation systems are likely to be of little help in evaluating texts generated automatically by AI systems, as grammatical accuracy and text cohesiveness are likely to be scored positively. On the other hand, although GPT-3 texts may read well, the text, or parts of it, may be strung-together words with relatively little real content. M. Zhang and Li (2021) comment that the GPT-3 model “has not met expectations in natural language reasoning” (p. 832). It does not have a real knowledge of the text it is generating. Argumentation is likely to be weak. A further concern with GPT-3 texts is that, as the system is data-driven, the dataset on which texts are based is anything and everything found on the Internet, including biased views and hateful language (Godwin-Jones, 2021). That kind of unwanted language will find its way into generated texts.

Classroom Integration

The new frontier in writing assistance, represented by the emerging availability of text generators, as well as the widespread use of MT and AWE/AWCF tools, clearly poses both opportunities and challenges to writing teachers and to L2 educators generally. Rejecting out of hand the use of advanced writing tools by students—or just ignoring their existence—is unacceptable at a time when such tools have become a “naturalized part of the modern, globalized world” (Hellmich & Vinall, 2021, p. 4). What is needed is thoughtful, informed differentiation in the use and the advocacy of AI-enabled tools, based on situated practice, established goals, and desired outcomes. The use of those tools should not be determined by convenience, but rather by how they fit into pedagogical and curricular objectives (Z. Li, 2021). Unfortunately, administrative fiat, institutional guidelines, marketing hype, or pressure from stakeholders may not provide school systems, language programs, or individual faculty choice in the matter. Even if a specific tool, such as an AWE system, is mandated, there is still room for writing experiences that are varied. A balanced use of AI writing tools might involve assigning a variety of writing tasks, some using the system and others not. Targeted reading audiences, whenever possible, could extend beyond both AI system and teacher (Grimes & Warschauer, 2010). Artificial intelligence writing tools should not “distract students from the communicative purpose of writing” (Grimes & Warschauer, 2010, p. 34). That entails integrating tool use “into a broader writing program emphasizing authentic communication” (Grimes & Warschauer, 2010, p. 34).

That holds true for MT use as well as it does for AWE and AWCF tools. That might mean having students use MT in everyday communicative activities in the classroom. Even simple tasks or elementary topics can be revelatory in terms of signaling the importance of context and word choice. This is particularly the case in areas such as register, genre, or style. A common greeting like “Hello,” for example, could be translated in a variety of ways depending on context, such as the level of formality or familiarity in the exchange. Such pragmatic and contingent considerations will likely not be reflected in machine-translated texts. Pellet and Myers (2022) discuss a variety of ordinary L2 learning tasks, which could be used to demonstrate both language pragmatics and the limitations of MT. Similarly, Ranalli (2021) advocates having students perform a review of AWE feedback, critically examining its effectiveness and usefulness. He suggests that after such an activity, students be supplied a text which they are asked to proof, identifying errors and providing corrections. Such hands-on tasks are helpful in guiding students to become informed consumers

of language tools.

John and Woll (2020) recommend using Grammarly in focus-on-form activities to target particular grammar points, based on Grammarly's categorization functionality. They also recommend a two-step process, in which teachers first have learners examine their own compositions for a particular type of error and then let them run their writing through an AWCf tool to see if it flags errors in that grammatical category. Another approach to integrating AI tools into L2 instruction is to assign tasks which incorporate specific materials currently or previously studied in the course. Writing assignments might, for example, require that students include a checklist of specific grammar or vocabulary, as outlined in Knowles (2022). The author also stipulates that students identify and label the examples used. Grading rubrics used in that study are made with the use of Google Translate in mind, as they include identification of grammar and vocabulary. Pellet and Myers (2022) also describe classroom activities with Google Translate that lead students to make explicit connection with recently studied topics. In that case, students are asked to discuss the sociopragmatic issues involved in the translation of a given text (i.e., "what's your name?"). Additionally, in a writing class, students might be asked to keep a learning diary outlining their experiences with the AI tool used.

In addition to such practical classroom experiences and purposeful teacher mediation, a number of studies have shown the utility of explicit instruction and guidance in the use of AI tools. That can involve learning how AI-based systems work, what kind of writing they are best used for, and what kind of performance one is likely to expect. That builds not only familiarity but also confidence, which are important factors in leading to "calibrated trust" (Ranalli, 2021, p. 14) in their use; students should develop realistic expectations of utility in tool choice and use. Ideally, those pedagogical practices align with other training, providing overall writing strategies training (Ling et al., 2021). Training and usage modeling of AI tools can lead to greater metalinguistic awareness. That can result in those tasks becoming *language-related episodes* (Swain & Lapkin, 1995); that is, activities in which students explicitly talk about language they are using. The analysis of language-related episodes is also discussed in the study of Grammarly use in Koltovskaia (2020). Pellet and Myers (2022) provide a series of examples built around MT, as do Woodworth and Barkaoui (2020) for AWE. Such experiences hold promise to help with the appropriate use of advanced language tools. They are also likely to result in greater learner autonomy.

Student views of technology tools can have a determining effect on their effectiveness, and that is also the case for teacher beliefs (Jiang et al., 2020). External factors come into play for teachers as they do for students. Teachers' use of technology is likely shaped by administrative factors as well as by curricular concerns. An important factor is the degree of comfort a teacher has with the technology tool, which is likely dependent on training and support (Kessler, 2021). In the case of advanced language tools such as intelligent text editors and MT, a complicating factor is the speed of development and change. Teachers who rejected Google Translate as unacceptable because its translations were unreliable may not subsequently have learned how improved that service has become. Automatic writing evaluation and AWCf systems likewise are in a constant state of quality improvement and feature growth.

Ecological Perspectives

Teachers using AWE or MT as pedagogical tools for writing improvement and language development are likely to be using a variety of other strategies as well for providing feedback. Automatic writing evaluation and AWCf studies stress the importance of continuing to use teacher feedback on writing tasks, not to rely solely on automated feedback (Y. Huang & Wilson, 2021; Z. Li, 2021). Link et al. (2020) describe an ideal hybrid situation in which an AWE tool provides sentence level feedback while the teacher attends to higher order writing issues. This is indeed one of the promises inherent in AWE use, that teachers' time and effort be directed not at fixing language mechanics but used to deal with stylistic and content-related features. Combining AWE with peer review is an additional option (Hockly, 2019), as can be done with MT as well (S.-M. Lee, 2020). Y. Huang & Wilson (2021) recommend a three-stage revising process, from AWE to peer evaluation to teacher corrective feedback. In fact, AWE systems, such as Criterion and [MI Write](#), have

mechanisms for including peer review. The Google Translate activities described in Pellet and Myers (2022) also entail substantial peer to peer interactions.

A good deal of the negative views of AI tools such as AWE and MT can be traced to ineffective integration into local instructional environments. Cotos (2014) points out in the discussion of the benefits of AWE that little consideration is typically accorded to contextual factors. Yet, how such a tool is integrated makes all the difference in how successful it is likely to be; Grimes & Warschauer (2010) provide illuminating examples. Wholeheartedly embracing AWE as a “silver bullet” (Warschauer & Ware, 2006, p. 175) for writing instruction is as misguided as banning Google Translate from use. Artificial Intelligence tools can play a role in developing student writing skills, but that needs to be, as Y. Huang & Wilson (2021) state, a supporting, not leading role. The “ecology of implementation” (Cotos, 2014, p. 59) of advanced language tools calls for thoughtful and deliberate use in contextually appropriate ways.

That larger, ecological perspective is often missing in discussion of AI tools. There are, however, several studies which point in that direction. Social informatics is evoked in Grimes and Warschauer (2010) as one way to break down the separation between people, technology, and organizations, involving a “more complex and locally situated process of technology diffusion” (p. 10). This is in contrast to a “tool” focus which implies a technological determinism undervaluing the roles played by people and organizations. Jiang et al. (2020) draw on mediated learning experience theory (I. Lee, 2014), in which AWE systems are considered sociocultural artifacts mediated through teacher and student use. That perspective helps to show that the use of AWE necessarily impacts not only student writing but also the nature of teacher CF. Woodworth and Barkaoui (2020) and Nunes et al. (2021) evoke sociocultural theory to categorize the scaffolding process used in the mix of AWE and teacher CF. Hellmich and Vinall (2021) use an ecological and complex systems approach to characterize the emerging outcomes from the different components of tools and people interacting across different scales (individual, classroom, department, and institution). Another important dimension to consider is time. Most studies of digital tools are short term. The few longitudinal studies published, such as S.-M. Lee (2020), are important in showing the presence or absence of long-term benefits.

Z. Li (2021) points out that teachers’ roles in technology-enhanced classrooms are under-investigated compared to the focus on tool use. He points out how the use of a tool like AWE inevitably changes the ecology of a learning and instructional system. Understanding that dynamic involves taking into consideration individual teacher differences that can play a major role (Link et al., 2020) as well as student characteristics (Y. Huang & Wilson, 2021; Ranalli, 2021). Patout and Cordy (2019) envision an AWE system able to take into account local settings as well as factors such as intended audience’s knowledge of the topic, the assigned writing task’s appropriate register, genre conventions, and lexical considerations based on the intended readers. Viewing writing assistance in classroom settings from an ecological setting can lead one to anticipate that the same tool used in different environments is likely to have widely different results. Complexity theory tells us that complex systems, involving the interplay of individuals, non-human entities, and institutions, as is the case in the tools discussed here, will have emergent outcomes that vary (Larsen-Freeman, 2018). The outcomes depend on initial conditions, evolving nonlinear and potentially unexpected processes/encounters, and shifting layered relationships among system components (Hellmich & Vinall, 2021).

Given that variability, it is important to trace individual case histories to illuminate what factors might be contributing to success or failure in working with AI writing tools. This aligns with the person-centered perspective increasingly adopted in SLA research (Larsen-Freeman, 2018). Studies such as Z. V. Zhang (2020) or Z. V. Zhang and Hyland (2018) on the use of AWE systems show widely different patterns of emergence through the individual students followed. In their examination of the use of Google Translate for learning Dutch vocabulary, van Lieshout and Cardoso (2022) used questionnaires to identify individual variables such as prior use of Google Translate, languages spoken, educational backgrounds, and experiences with autonomous language learning. Ranalli’s (2021) study of individual students using Grammarly demonstrates how different learning orientations can have a substantial impact on how digital

writing tools are used. Learner identities played a major role in that orientation, with students' confidence in their language abilities and self-image as learners being important factors. Interestingly, in one learner's case another factor was the "process-related awareness of Grammarly's workings" (p. 13) that the individual had as a premium user of the service. Ranalli (2021) concludes that the degree of users' engagement in such a system is "complex and multi-faceted" (p. 13) and therefore can vary considerably. The study by Koltovskaia (2020) likewise follows closely how individual students used Grammarly and looks at patterns of engagement (and disengagement) that have revealing consequences for the effectiveness in the use of the tool's feedback. Qualitative research featuring individual student learning pathways can be quite helpful in sorting out learner differences that can affect the use of digital tools.

Among the variables that can affect the dynamics of AI tool use is the human-machine relationship. In any given implementation of technology and instruction there is likely to be a range of attitudes and reactions to the technology, from enthusiastic acceptance to utter rejection and many stages in between. That can bring learner emotions into play, a factor that can have repercussions on both technology use and learning effectiveness. Studies have emphasized the important role emotions such as anxiety or mistrust can have on learner motivation in the use of technology tools (Nazari et al., 2021; Nunes et al., 2021; Sun & Fan, 2022). The issue of trust is raised in Ranalli (2021) as a central factor in student reception of AWE feedback. The degree of trust can be affected by a variety of variables including subjective perceptions about individual workload, the nature of the interaction with the software, and other considerations having little to do with the actual capabilities of the software in question. Ranalli (2021) suggests using human-automation trust theory (J. Lee & See, 2004) to illuminate the dynamics at play. He points out that trust may play a central role in the degree of engagement a user has with the technology tool. In the case of the use of AWE by L2 students, there is a high degree of vulnerability in using an unmastered language with a tool about which they have little understanding and whose feedback they are expected to both understand and implement.

Conclusion: Coexistence with AI

Developments in automated writing call for greater awareness on the part of L2 and writing instructors of what AI tools are capable of providing in terms of writing assistance or authoring (Carvalho et al., 2022). Considering the media hype around AI-based writing, it is likely that regardless of views on their usefulness or ethics, text generators and other emerging writing tools will be used by many students. Given that likelihood, it is incumbent on teachers to find ways for students to use the tools appropriately and whenever possible to integrate their use into instruction (Otsuki, 2020). Learning to use AI writing tools not only is important for L2 learners but also will figure largely for many of them in their lives after graduation. Facility with such tools has become a major component in digital literacy in both the classroom and the workplace.

One can but hope that the tools geared towards the educational community, especially AWE, will follow suggestions from researchers to add features that would improve their usefulness to both teachers and students. One of the principal areas of improvement would be adding flexibility of use. Helpful, for example, would be a mode in which errors are highlighted only but not labeled or corrected (Ranalli, 2021). In fact, in general the ability to turn features on and off would be very useful. Frankenberg-Garcia (2020) provides a helpful example in a writing tool for helping L2 writers with using collocations appropriately. The system (ColloCaid) features an incremental display of information on collocations, with writers able to display different levels of detail, along with metalinguistic information and examples. Additionally, users can customize collocation cues. Along with feedback, the tool also aims to 'feed forward,' raising awareness of collocations writers may not remember or know how to look up. Such scaffolding support can provide the kind of flexibility needed for learners with different language backgrounds and learning goals.

Also needed for systems to be more accommodating to learner context is a greater built-in awareness of process writing, so that the feedback could be calibrated for the different stages of writing and revision (Ranalli, 2021). Better adjustments in CF for different writing genres are also needed. In general, a greater awareness of how AWE systems are used in actual practice (as is reflected in AWE research) would allow

for better compatibility of tool and teaching/learning environments. For example, in an L2 context, the feedback might be different in nature and in formulation that it is in L1 settings. Adding L1 glosses to AWE feedback, for instance, could be helpful, as one study has shown (Wilken, 2018). That could be especially welcome for novice learners, particularly if there are options for displaying either L1 or L2 feedback. Such a feature is built into an AI writing tool for English as a Second Language described in Gayed et al. (2022), which uses a reverse translation function to offer a simultaneous translation in the user's L1, displayed under their writing.

For CALL and L2 writing researchers, it would be useful to retrieve maximum data from tools for intelligent writing assistance. Helpful as well would be information about backend services used to drive the software, that is, AI and datasets, so that researchers have a better understanding of how the systems work. Unfortunately, that is likely not anything companies will be eager to reveal. Greater transparency would be helpful in allaying concerns among teachers as well as building a higher level of trust with users. Remaining a black box does not help develop a better human-automation relationship. It seems inevitable that AI systems will to some degree follow developments in robotics and chatbots to make the user interface as human-seeming as possible (Godwin-Jones, 2019). That is likely to influence trust and usage patterns. This development will be important to follow from an ecological perspective, so that we look beyond tool features to the larger environment of teaching, learning, and automation.

An ecological perspective on the use of AI tools and education invites a wider, societal consideration of issues like learner agency and equity. Carvalho et al. (2022) address those issues, focusing on the question of designing for learning in an AI world. The authors point out that the growing incorporation of AI systems into both everyday life and education is likely to lead to significant changes and disruptions, bringing in its wake a growing sense of uncertainty. The article argues that “both educators and learners should be involved not only in learning but also in co-designing for learning in an AI world. Further, they together should explore the knowledge, goals and actions that could help people shape future AI scenarios” (p. 1). As an underpinning philosophy, the authors argue for the use of a “capability approach” (p. 8) originally conceived by Sen (1985), which emphasizes the centrality of human wellbeing and individual freedom of choice within societal and institutional constraints. As AI plays an ever-increasing role in education, it is likely that both learners and teachers will be co-creating with algorithmic systems. To do that equitably, designing for learning will need to view AI tool from a broader, social perspective and consider the impact on individual lives. This aligns with a proposal in Lütge et al. (in press) to teach global citizenship through FL education so as to “empower educational actors with the capacity to orient themselves when confronted with unknowns” (cited in Kramsch, 2022, p. 2). Faced with AI-enabled writing tools that provide extensive assistance in L2 writing, language educators will need to find ways to value learner freedom and reward creativity.

References

- Arroyo, D. C., & Yilmaz, Y. (2018). An open for replication study: The role of feedback timing in synchronous computer-mediated communication. *Language Learning*, 68(4), 942–972. <https://doi.org/10.1111/lang.12300>
- Bitchener, J., & Ferris, D. R. (2012). *Written corrective feedback in second language acquisition and writing*. Routledge.
- Bridgeman, B., & Ramineni, C. (2017). Design and evaluation of automated writing evaluation models: Relationships with writing in naturalistic settings. *Assessing Writing*, 34, 62–71. <https://doi.org/10.1016/j.asw.2017.10.001>
- Burstein, J., Elliot, N., & Molloy, H. (2016). Informing automated writing evaluation using the lens of genre: Two studies. *CALICO Journal*, 33(1), 117–141. <https://doi.org/10.1558/cj.v33i1.26374>

- Camacho, A., Alves, R. A., & Boscolo, P. (2021). Writing motivation in school: A systematic review of empirical research in the early twenty-first century. *Educational Psychology Review*, 33(1), 213–247. <https://doi.org/10.1007/s10648-020-09530-4>
- Carvalho, L., Martinez-Maldonado, R., Tsai, Y.-S., Markauskaite, L., & De Laat, M. (2022). How can we design for learning in an AI world?. *Computers and Education: Artificial Intelligence*, 3, 1–9. <https://doi.org/10.1016/j.caeai.2022.100053>
- Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing*, 32(3), 385–405. <https://doi.org/10.1177/0265532214565386>
- Conijn, R., Zaanen, M. V., & Waes, L. V. (2019). Don't wait until it is too late: The effect of timing of automated feedback on revision in ESL writing. In M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, & J. Schneider (Eds.), *Transforming learning with meaningful technologies* (pp. 577–581). Springer. https://doi.org/10.1007/978-3-030-29736-7_43
- Cotos, E. (2014). *Genre-based automated writing evaluation for L2 research writing*. Palgrave Macmillan.
- Crossley, S. A. (2018). Technological disruption in foreign language teaching: The rise of simultaneous machine translation. *Language Teaching*, 51(4), 541–552. <https://doi.org/10.1017/S0261444818000253>
- Dale, R. (2020). Natural language generation: The commercial state of the art in 2020. *Natural Language Engineering*, 26, 481–487. <https://doi.org/10.1017/s135132492000025x>
- Dale, R. (2021). GPT-3: What's it good for?. *Natural Language Engineering*, 27(1), 113–118. <https://doi.org/10.1017/s1351324920000601>
- Dale, R., & Viethen, J. (2021). The automated writing assistance landscape in 2021. *Natural Language Engineering*, 27(4), 511–518. <https://doi.org/10.1017/s1351324921000164>
- Dembsey, J. M. (2017). Closing the Grammarly gaps: A study of claims and feedback from an online grammar program. *The Writing Center Journal*, 36(1), 63–100.
- Dizon, G., & Gayed, J. M. (2021). Examining the impact of Grammarly on the quality of mobile L2 writing. *JALT CALL Journal*, 17(2), 74–92. <https://doi.org/10.29140/jaltcall.v17n2.336>
- Dodigovic, M., & Tovmasyan, A. (2021). Automated writing evaluation: The accuracy of Grammarly's feedback on form. *International Journal of TESOL Studies*, 3(2), 71–88. <https://doi.org/10.46451/ijts.2021.06.06>
- Dodigovic, M., Mlynarski, J., & Wei, R. (2016). Vocabulary size assessment as a predictor of plagiarism. In C. Gitsak, & C. Coombe (Eds.), *Current issues in language evaluation, assessment and testing: Research and practice* (pp. 222–235). Cambridge Scholars Publishing.
- Eaton, S. E., Mindzak, M., & Morrison, R. (2021, May 29–June 3). *Artificial intelligence, algorithmic writing & educational ethics* [Paper presentation]. Canadian Society for the Study of Education [Société canadienne pour l'étude de l'éducation] (CSSE) 2021, Edmonton, AB, Canada. <http://hdl.handle.net/1880/113569>
- Ellis, N. C. (2017). Cognition, corpora, and computing: Triangulating research in usage-based language learning. *Language Learning*, 67(S1), 40–65. <https://doi.org/10.1111/lang.12215>
- Ellis, R. (2016). Focus on form: A critical review. *Language Teaching Research*, 20(3), 405–428. <https://doi.org/10.1177/1362168816628627>

- Enriquez Raído, V., & Sánchez Torrón, M. (2020) Machine translation, language learning and the 'knowledge economy.' In M. Filimowicz & V. Tzankova (Eds.), *Reimagining communication: Action* (pp. 155–171). Taylor and Francis/Routledge.
- Ferrone, L., & Zanzotto, F. M. (2020). Symbolic, distributed, and distributional representations for natural language processing in the era of deep learning: A survey. *Frontiers in Robotics and AI*, 153. <https://doi.org/10.3389/frobt.2019.00153>
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4), 681–694. <https://doi.org/10.2139/ssrn.3827044>
- Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, 32(4), 365–387. <https://doi.org/10.2307/356600>
- Frankenberg-Garcia, A. (2020). Combining user needs, lexicographic data and digital writing environments. *Language Teaching*, 53(1), 29–43. <https://doi.org/10.1017/s0261444818000277>
- Fredholm, K. (2014). Effects of online translation on morphosyntactic and lexical-pragmatic accuracy in essay writing in Spanish as a foreign language. In S. Jager, L. Bradley, E. J. Meima, & S. Thouëсны (Eds.), *CALL design: Principles and practice. Proceedings of the 2014 EUROCALL Conference. Groningen, The Netherlands* (pp. 96–101). Research-publishing net. <https://doi.org/10.14705/rpnet.2014.000201>
- Fredholm, K. (2015). Online translation use in Spanish as a foreign language essay writing: Effects on fluency, complexity and accuracy. *Revista Nebrija de Lingüística Aplicada a la Enseñanza de las Lenguas*, 18, 1–18.
- Fredholm, K. (2019). Effects of Google translate on lexical diversity: Vocabulary development among learners of Spanish as a foreign language. *Revista Nebrija de Lingüística Aplicada a la Enseñanza de las Lenguas*, 13(26), 98–117.
- Gayed, J. M., Carlon, M. K. J., Oriola, A. M., & Cross, J. S. (2022). Exploring an AI-based writing assistant's impact on English language learners. *Computers and Education: Artificial Intelligence*, 100055. <https://doi.org/10.1016/j.caeai.2022.100055>
- Ghufron, M. A. (2019, April). *Exploring an automated feedback program 'Grammarly' and teacher corrective feedback in EFL writing assessment: Modern vs. traditional assessment* [Paper presentation]. The 3rd English Language and Literature International Conference, Semarang, Indonesia. <https://doi.org/10.4108/eai.27-4-2019.2285308>
- Godwin-Jones, R. (2017). Data-informed language learning. *Language Learning & Technology*, 21(3), 9–27. <https://doi.org/10125/44629>
- Godwin-Jones, R. (2018). Second language writing online: An update. *Language Learning & Technology*, 22(1), 1–15. <https://dx.doi.org/10125/44574>
- Godwin-Jones, R. (2019). In a world of SMART technology, why learn another language?. *Journal of Educational Technology & Society*, 22(2), 4–13.
- Godwin-Jones, R. (2021). Big data and language learning: Opportunities and challenges. *Language Learning & Technology*, 25(1), 4–19. <http://hdl.handle.net/10125/44747>
- Grimes, D. & Warschauer, M. (2010) Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Language, and Assessment*, 8(6), 1–43.
- Guo, Q., Feng, R., & Hua, Y. (2021). How effectively can EFL students use automated written corrective feedback (AWCF) in research writing?. *Computer Assisted Language Learning*, 1–20. <https://doi.org/10.1080/09588221.2021.1879161>

- Hegelheimer, V., & Lee, J. (2012). The role of technology in teaching and researching writing. In M. Thomas, H. Reinders, & M. Warschauer (Eds.), *Contemporary computer-assisted language learning* (pp. 287–302). Bloomsbury
- Hellmich, E., & Vinall, K. (2021). FL instructor beliefs about machine translation: Ecological insights to guide research and practice. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, 11(4), 1–18. <https://doi.org/10.4018/ijcallt.2021100101>
- Hibert, A. I. (2019). Systematic literature review of automated writing evaluation as a formative learning tool. In M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, & J. Schneider (Eds.), *Transforming learning with meaningful technologies* (pp. 199–212). Springer. https://doi.org/10.1007/978-3-030-29736-7_15
- Hockly, N. (2019). Automated writing evaluation. *ELT Journal*, 73(1), 82–88. <https://doi.org/10.1093/elt/ccy044>
- Huang, H. W., Li, Z., & Taylor, L. (2020, May). The effectiveness of using Grammarly to improve students' writing skills. In *Proceedings of the 5th International Conference on Distance Education and Learning* (pp. 122–127). Association for Computing Machinery. <https://doi.org/10.1145/3402569.3402594>
- Huang, Y., & Wilson, J. (2021). Using automated feedback to develop writing proficiency. *Computers and Composition*, 62, 102675. <https://doi.org/10.1016/j.compcom.2021.102675>
- Hussein M. A., Hassan H., & Nassef M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, 5(208). <https://doi.org/10.7717/peerj-cs.208>
- Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching*, 39(2), 83–101. <https://doi.org/10.1017/s0261444806003399>
- Jiang, L., Yu, S., & Wang, C. (2020). Second language writing instructors' feedback practice in response to automated writing evaluation: A sociocultural perspective. *System*, 93, 1–12. <https://doi.org/10.1016/j.system.2020.102302>
- John, P., & Woll, N. (2020). Using grammar checkers in an ESL context: An investigation of automatic corrective feedback. *CALICO Journal*, 37(2), 169–172. <https://doi.org/10.1558/cj.36523>
- Jolley, J. R., & Maimone, L. (2022). Thirty years of machine translation in language teaching and learning: A review of the literature. *L2 Journal*, 14(1), 26–44. <https://doi.org/10.5070/l214151760>
- Kang, E., & Han, Z. (2015). The efficacy of written corrective feedback in improving L2 written accuracy: A meta-analysis. *The Modern Language Journal*, 99(1), 1–18. <https://doi.org/10.1111/modl.12189>
- Karyuatry, L., Rizqan, M. D., & Darayani, N. A. (2018). Grammarly as a tool to improve students' writing quality. *Journal Sains Sosial dan Humaniora*, 2(1), 83–89. <https://doi.org/10.30595/jssh.v2i1.2297>
- Kessler, G. (2021). Current realities and future challenges for CALL teacher preparation. *CALICO Journal*, 38(3), i–xx. <https://doi.org/10.1558/cj.21231>
- Klekovkina, V., & Denié-Higney, L. (2022). Machine translation: Friend or foe in the language classroom?. *L2 Journal*, 14(1), 105–135. <https://doi.org/10.5070/l214151723>
- Knowles, C. L. (2022). Using an ADAPT approach to integrate Google Translate into the second language classroom. *L2 Journal*, 14(1), 195–236. <https://doi.org/10.5070/l214151690>

- Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing*, 44, 1–12. <https://doi.org/10.1016/j.asw.2020.100450>
- Kramersch, C. (2022). General editor's preface. *L2 Journal*, 14(1), 1–3. <https://doi.org/10.5070/l214256399>
- Larsen-Freeman, D. (2018). Looking ahead: Future directions in, and future research into, second language acquisition. *Foreign Language Annals*, 51(1), 55–72. <https://doi.org/10.1111/flan.12314>
- Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2014). Automated grammatical error detection for language learners. *Synthesis Lectures on Human Language Technologies*, 7(1), 1–170. <https://doi.org/10.2200/s00275ed1v01y201006hlt009>
- Lee, I. (2014). Revisiting teacher feedback in EFL writing from sociocultural perspectives. *TESOL Quarterly*, 48, 201–213. <https://doi.org/10.1002/tesq.153>
- Lee, J., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80.
- Lee, S.-M. (2020) The impact of using machine translation on EFL students' writing. *Computer Assisted Language Learning*, 33(3), 157–175. <https://doi.org/10.1080/09588221.2018.1553186>
- Lewis-Kraus, G. (2016, December 14). The great A.I. awakening. *The New York Times*. <https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html>
- Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, 60, 309–365. <https://doi.org/10.1111/j.1467-9922.2010.00561.x>
- Li, Z. (2021). Teachers in automated writing evaluation (AWE) system-supported ESL writing classes: Perception, implementation, and influence. *System*, 99, 102505. <https://doi.org/10.1016/j.system.2021.102505>
- Ling, G., Elliot, N., Burstein, J. C., McCaffrey, D. F., MacArthur, C. A., & Holtzman, S. (2021). Writing motivation: A validation study of self-judgment and performance. *Assessing Writing*, 48, 1–15. <https://doi.org/10.1016/j.asw.2020.100509>
- Link, S., Mehrzad, M., & Rahimi, M. (2020). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*, 1–30. <https://doi.org/10.1080/09588221.2020.1743323>
- Lütge, C., Merse, T. & Rauschert, P. (Eds.) (in press). *Global citizenship in foreign language education: Concepts, practices, connections*. Routledge.
- MLA Ad Hoc Committee on Foreign Languages. (2007). Foreign languages and higher education: New structures for a changed world. *Profession*, 2007, 234–245.
- Nazari, N., Shabbir, M. S., & Setiawan, R. (2021). Application of Artificial Intelligence powered digital writing assistant in higher education: Randomized controlled trial. *Heliyon*, 7(5), e07014. <https://doi.org/10.1016/j.heliyon.2021.e07014>
- Nova, M. (2018). Utilizing Grammarly in evaluating academic writing: A narrative research on EFL students' experience. *Premise: Journal of English Education and Applied Linguistics*, 7(1), 80–96. <https://doi.org/10.24127/pj.v7i1.1332>
- Nunes, A., Cordeiro, C., Limpo, T., & Castro, S. L. (2021). Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020. *Journal of Computer Assisted Learning*, 38(2), 599–620. <https://doi.org/10.1111/jcal.12635>
- O'Neill, E. M. (2016). Measuring the impact of online translation on FL writing scores. *The IALLT Journal of Language Learning Technologies*, 46(2), 1–39. <https://doi.org/10.17161/iallt.v46i2.8560>

- O'Neill, E. M. (2019). Training students to use online translators and dictionaries: The impact on second language writing scores. *International Journal of Research Studies in Language Learning*, 8(2), 47–65. <https://doi.org/10.5861/ijrsl.2019.4002>
- O'Neill, R., & Russell, A. (2019). Stop! Grammar time: University students' perceptions of the automated feedback program Grammarly. *Australasian Journal of Educational Technology*, 35(1), 42–56. <https://doi.org/10.14742/ajet.3795>
- O'Neill, R., & Russell, A. M. (2020). Grammarly: Help or hindrance? Academic learning advisors' perceptions of an online grammar checker. *Journal of Academic Language and Learning*, 13(1), A88–A107.
- Ortega, L. (2013). SLA for the 21st century: Disciplinary progress, transdisciplinary relevance, and the bi/multilingual turn. *Language Learning*, 63, 1–24. <https://doi.org/10.1111/j.1467-9922.2012.00735.x>
- Ortega, L. (2017). New CALL-SLA research interfaces for the 21st century: Towards equitable multilingualism. *CALICO Journal*, 34(3), 283–316. <https://doi.org/10.1558/cj.33855>
- Otsuki, G. J. (2020, January 23). OK computer: To prevent students cheating with AI text-generators, we should bring them into the classroom. *The Conversation*. <https://theconversation.com/ok-computer-to-prevent-students-cheating-with-ai-text-generators-we-should-bring-them-into-the-classroom-129905>
- Palermo, C., & Wilson, J. (2020). Implementing automated writing evaluation in different instructional contexts: A mixed-methods study. *Journal of Writing Research*, 12(1), 63–108. <https://doi.org/10.17239/jowr-2020.12.01.04>
- Patout, P. A., & Cordy, M. (2019). Towards context-aware automated writing evaluation systems. In B. Vanderose (Ed.), *Proceedings of the 1st ACM SIGSOFT international workshop on education through advanced software engineering and Artificial Intelligence* (pp. 17–20). Association for Computing Machinery. <https://doi.org/10.1145/3340435.3342722>
- Pellet, S., & Myers, L. (2022). What's wrong with “What is your name?” > “Quel est votre nom?”: Teaching responsible use of MT through discursive competence and metalanguage awareness. *L2 Journal*, 14(1), 166–194. <https://doi.org/10.5070/l214151739>
- Peng, H., Jager, S., & Lowie, W. (2020). A person-centred approach to L2 learners' informal mobile language learning. *Computer Assisted Language Learning*. <https://doi.org/10.1080/09588221.2020.1868532>
- Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it?. *Computer Assisted Language Learning*, 31(7), 653–674. <https://doi.org/10.1080/09588221.2018.1428994>
- Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing*, 52, 100816. <https://doi.org/10.1016/j.jslw.2021.100816>
- Ranalli, J., & Yamashita, T. (2022). Automated written corrective feedback: Error-correction performance and timing of delivery. *Language Learning & Technology*, 26(1), 1–25. <http://hdl.handle.net/10125/73465>
- Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*, 37(1), 8–25. <https://doi.org/10.1080/01443410.2015.1136407>

- Ruder, S. (2018, October 1). A Review of the neural history of natural language processing. *AYLIEN*. <https://aylien.com/blog/a-review-of-the-recent-history-of-natural-language-processing#nonneuralmilestones>
- Ryu, J., Kim, Y., Park, S., Eum, S., Chun, S., & Yang, S. (2022). Exploring foreign language students' perceptions of the guided use of machine translation (GUMT) model for Korean writing. *L2 Journal*, 14(1), 136–165. <https://doi.org/10.5070/1214151759>
- Saricaoglu, A. (2019). The impact of automated feedback on L2 learners' written causal explanations. *ReCALL*, 31(2), 189–203. <https://doi.org/10.1017/s095834401800006x>
- Schmalz, V. J., & Brutti, A. (2021, July 1). *Automatic assessment of English CEFR levels using BERT embeddings* [Paper presentation]. Italian Conference on Computational Linguistics 2021, Bologna, Italy. <http://ceur-ws.org/Vol-3033/paper14.pdf>
- Sen, A. (1985). Well-being, agency and freedom: The Dewey lectures 1984. *The Journal of Philosophy*, 82(4), 169–221. <https://doi.org/10.4324/9781315251240-1>
- Stevenson, M. (2016). A critical interpretative synthesis: The integration of automated writing evaluation into classroom writing instruction. *Computers and Composition*, 42, 1–16. <https://doi.org/10.1016/j.compcom.2016.05.001>
- Sukkhwan, A. (2014). *Students' attitudes and behaviors towards the use of Google Translate* [Unpublished master's thesis]. Prince of Songkla University.
- Sun, B., & Fan, T. (2022). The effects of an AWE-aided assessment approach on business English writing performance and writing anxiety: A contextual consideration. *Studies in Educational Evaluation*, 72, 101123. <https://doi.org/10.1016/j.stueduc.2021.101123>
- Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics*, 16(3), 371–391. <https://doi.org/10.1093/applin/16.3.371>
- Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.
- Truscott, J. (2007). The effect of error correction on learners' ability to write accurately. *Journal of Second Language Writing*, 16(4), 255–272. <https://doi.org/10.1016/j.jslw.2007.06.003>
- Urlaub, P., & Dessein, E. (2022). From disrupted classrooms to human-machine collaboration? The pocket calculator, Google Translate, and the future of language education. *L2 Journal*, 14(1), 45–59. <https://doi.org/10.5070/1214151790>
- van Lieshout, C., & Cardoso, W. (2022). Google Translate as a tool for self-directed language learning. *Language Learning & Technology*, 26(1), 1–19. <http://hdl.handle.net/10125/73460>
- Vinall, K., & Hellmich, E. A. (2021). Down the rabbit hole: Machine translation, metaphor, and instructor identity and agency. *Second Language Research & Practice*, 2(1), 99–118. <http://hdl.handle.net/10125/69860>
- Vinall, K., & Hellmich, E. A. (2022). Do you speak translate?: Reflections on the nature and role of translation. *L2 Journal*, 14(1), 4–25. <https://doi.org/10.5070/1214156150>
- Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, 10(2), 157–180. <https://doi.org/10.1191/1362168806lr190oa>
- Warschauer, M., Yim, S., Lee, H., & Zheng, B. (2019). Recent contributions of data mining to language learning research. *Annual Review of Applied Linguistics*, 39, 93–112. <https://doi.org/10.1017/s0267190519000023>

- Wilken, J. L. (2018). Perceptions of L1 glossed feedback in automated writing evaluation: A case study. *CALICO Journal*, 35(1), 30–48. <https://doi.org/10.1558/cj.26383>
- Woodworth, J., & Barkaoui, K. (2020). Perspectives on using automated writing evaluation systems to provide written corrective feedback in the ESL classroom. *TESL Canada Journal*, 37(2), 234–247. <https://doi.org/10.18806/tesl.v37i2.1340>
- Zhang, H., & Torres-Hostench, O. (2022). Training in machine translation post-editing for foreign language students. *Language Learning & Technology*, 26(1), 1–17. <http://hdl.handle.net/10125/73466>
- Zhang, M., & Li, J. (2021). A commentary of GPT-3 in MIT Technology Review 2021. *Fundamental Research*, 1(6), 831–833. <https://doi.org/10.1016/j.fmre.2021.11.011>
- Zhang, Z. V. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing*, 43, 100439. <https://doi.org/10.1016/j.asw.2019.100439>
- Zhang, Z. V., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing*, 36, 90–102. <https://doi.org/10.1016/j.asw.2018.02.004>
- Zheng, Y., & Yu, S. (2018). Student engagement with teacher written corrective feedback in EFL writing: A case study of Chinese lower-proficiency students. *Assessing Writing*, 37, 13–24. <https://doi.org/10.1016/j.asw.2018.03.001>
- Zomer, G., & Frankenberg-Garcia, A. (2021). Beyond grammatical error correction: Improving L1-influenced research writing in English using pre-trained encoder-decoder models. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 2534–2540). 2021 Association for Computational Linguistics. <https://aclanthology.org/2021.findings-emnlp.216/>