

Tamil Co-Writer: Towards inclusive use of generative AI for writing support

Antonette Shibani¹, Faerie Mattins^{2,3}, Srivarshan Selvaraj^{2,3}, Ratnavel Rajalakshmi² and Gnana Bharathy¹

¹ University of Technology Sydney

² Vellore Institute of Technology, Chennai

³ University of Southern California

Abstract

The increasing use of generative AI in education highlights its potential for enriching learning experiences. One application for utilising the capabilities of Large Language Models (LLMs) such as the Generative Pre-trained Transformer (GPT) is the creation of writing support tools. In particular, tools that can work in partnership with humans to co-write with AI hold great promise and have been tested out for English language writing. However, the adaptation of such tools to languages other than English is limited, presenting a disadvantage for learners from linguistically diverse backgrounds. In the current study, we extend previous works in English to develop a writing aid prototype for the low-resource Indian regional language Tamil for co-writing with AI called Tamil Co-Writer. The tool additionally provides a visual summary of user interaction and co-authorship metrics for each writing session for users to reflect on their usage of AI in their own writing. We posit that such interactive tools using the latest generative AI technologies can help writers improve their writing skills and productivity in their own regional languages supporting inclusive AI for education.

Keywords

Large language model, generative AI, artificial intelligence, LLM, Tamil, writing, GPT, keystroke analysis, CoAuthorViz, Tamil Co-Writer, inclusive AI, equity


1. Introduction


Large Language Models (LLM) are sophisticated artificial intelligence (AI) systems trained on massive amounts of textual data to generate text. LLMs such as Generative Pre-trained Transformer (GPT) are capable of producing language that is grammatically correct and appears human-written. They are seen to be performing well in a variety of language-related tasks such as translating, summarizing, responding to questions, and creating new content, and are increasingly employed across many sectors. Although relatively new, LLMs are starting to be deployed in intelligent support systems for learners and writers [1, 2]. For writing support, a tool that uses GPT-3 called CoAuthor was used to collect a dataset of collaborative writing between humans and AI [3]. The technology allowed users to write freely, solicit suggestions from GPT-3, accept or reject those suggestions, and alter previously written texts or accepted suggestions in any sequence they desired. Past research also presented a visual representation (CoAuthorViz) of user interactions with the tool to study AI-dependency behaviours of users [5]. This demonstrates the usefulness of AI-based tools in enhancing learner capabilities; however, they only support English language users and do not address the needs of diverse groups of learners to promote inclusivity in education.

The objective of the current study is to introduce a working prototype of an AI co-authoring tool for the regional language ‘Tamil’ [4]. Tamil is a Dravidian language primarily spoken in the Indian state of Tamil Nadu and northeast Sri Lanka. While estimated to be spoken by over 80 million native speakers worldwide, Tamil falls under the category of low-resource languages for Natural Language Processing research, characterized by limited tools and datasets. Recent studies in under-resourced languages aim to bridge this gap by specifically targeting the creation of additional resources [6, 7]. Our prototype tool called ‘Tamil Co-Writer’ can help Tamil language learners engage in interactive writing sessions with AI support in the form of auto-generated suggestions. The tool incorporates speech to text input using Automatic Speech Recognition (ASR) for increased accessibility. The learner also receives a summary statistic and a visual graph at the end of each writing session using the tool to reflect on their dependence

Joint Proceedings of LAK 2024 Workshops, co-located with 14th International Conference on Learning Analytics and Knowledge (LAK 2024), Kyoto, Japan, March 18-22, 2024.

✉ antonette.shibani@uts.edu.au (A. Shibani); rajalakshmi.r@vit.ac.in (R. Rajalakshmi)

 0000-0003-4619-8684 (A. Shibani)

 © 2024 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

















on AI suggestions in their writing. Our study demonstrates how tools such as this can cater to linguistically and culturally diverse groups of users to aid their writing for equitable use of AI. The contributions of this paper are as follows:

- Development of a novel writing aid prototype for Tamil using the open-source GPT-2 model called Tamil Co-Writer, incorporating an ASR model for text-to-speech input and visual statistics of AI usage.
- The conceptualisation and evaluation of how LLMs can be effectively embedded in writing tools for improved accessibility and language support for diverse users whose first language is not English.

2. Background

Automated writing analysis, feedback, and digital tools have long been used to support writers in their writing process [8, 9]. In learning analytics, the sub-field of writing analytics has examined how tools and analytics techniques can be used to help learners with their writing processes and products [10, 11]. With the advent of advanced technologies such as LLMs, more sophisticated tools are now being developed that invite writers to co-write with AI as an active companion. Most well-documented among AI co-creation tools is CoAuthor [12], which is a combination of an interface, dataset, and experiment all in one. The tool was evaluated for use by over 60 people for creative and argumentative writing tasks where they co-wrote with GPT-3 generated text suggestions. The interactions between the writer and the GPT-3 suggestions were also captured using keystroke logging, which has further led to the study of human-AI interaction behaviours [5]. Another such tool produced 'sparks', which were new sentences generated by the AI to inspire users to write scientific content [1]. Here, the inspirations helped writers with elaborating sentences with detail, providing interesting angles to engage readers, and showing potential reader perspectives. A web application called Wordcraft provided multimodal machine intelligence to help writers make integrative leaps in creative writing as they wrote stories [13]. A similar co-creative story authoring tool utilizes large language models and story grammars, allowing authors to easily engineer text generation to meet their expectations [14]. While most tools support the drafting process of writing, other tools that support the revision process not fully handing over the creative process to AI are also being developed. This includes a human-in-the-loop iterative text revision system called Read, Revise, Repeat (R3) where writers interacted with model-generated revisions for deeper edits [15].

Commercial tools such as Grammarly, ProWritingAid, Quillbot, and Scribbr that previously focused on grammar and style also appear to have been transformed by LLMs into co-writing tools in English, providing feedback on content relevance, tone, cohesion. They enhance contents upon request, and correct errors [16]. Being commercial tools, their exact capabilities and limitations have not been well documented in the research space and constantly evolve over time. In the new world, plagiarism checkers are also being incorporated into tools such as Quillbot [19], which, while suggesting improvements or providing feedback, also check for potential plagiarism in content.

A plethora of products facilitate co-authoring, including content generators, programming automation, and AI-based virtual assistants propelled by the increasing availability of generative AI. Co-pilots and assistants where writers can obtain suggestions "as-they-type" are touted as the future of writing tools. However, all tools discussed above have been developed for writing in English. There are no such authoring tools created for low-resource languages such as the regional Indian languages, e.g. Tamil. By not focusing on languages other than English, we miss the rich cultural context and nuances in regional languages that are not well-resourced for NLP tasks. Fortunately, this is beginning to change for Indian languages as fine-tuned LLMs are starting to emerge [17]. Our research aims to contribute to the growing space for Tamil learning technologies and data to cater to linguistically diverse user groups and research areas. This paper explains the technical components in building such an AI co-writing system using open-source technology that can be generalized for other languages for writing support.

3. Development and evaluation of Tamil Co-Writer

Our novel collaborative writing tool ‘Tamil Co-Writer’ consists of a simple user interface for writing Tamil text and several underlying technical components that facilitate AI support and speech recognition. Figure 1 shows a flow chart of the individual components, which are explained below.

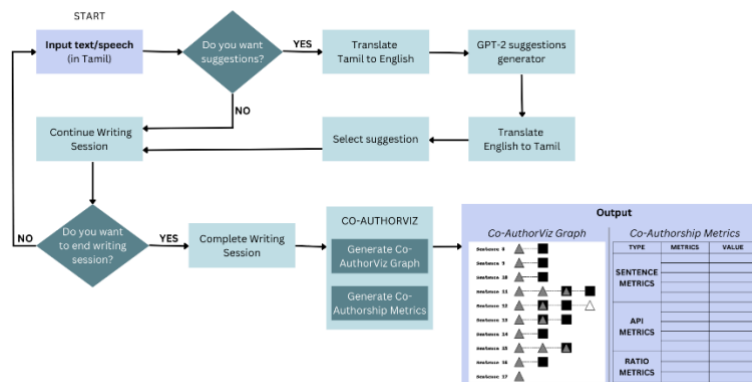


Figure 1: Workflow and components of Tamil Co-Writer

3.1. Tool functionality

The front end of Tamil Co-Writer contains an input text field that allows users to start writing their text in a writing session. The user has an additional option to provide input in speech format, which is processed by a fine-tuned ASR model to convert speech-to-text. This option can improve the accessibility of writing support tools by providing users with the option to write text as they speak their language. The use of non-standard vocabulary, accent differences, and background noise make ASR models difficult to accurately identify speech [18], with increased challenges in low- resource languages such as Tamil having insufficient training data. When using the tool, the user first enters the text in Tamil or speaks through the ASR system in Tamil. Upon asking for a GPT-2 suggestion (by clicking the ‘Get suggestion’ button), this Tamil text is translated to English and sent out for text generation using GPT-2. Five suggestions automatically generated using GPT-2 are displayed. The writer can accept a suggestion as is, reject a suggestion, or accept the suggestion and then modify it. At the end of the writing session, the writer obtains summary statistics and graphs that show their level of collaborative writing with the aid of GPT-2. The website is built using Django 2 and locally tested for prototyping. The tool thus consists of the three main components below at the back end, described next:

1. The input module processes the written text or speech from the user.
2. The automated text generation module then suggests new text using the GPT-2 model.
3. The metrics generation module populates the visual graph and key metrics of AI usage.

3.2. Input module

The input module processes the written text entered by the user directly in Tamil or the entered speech for conversion to text. A screenshot of the user interface is in Figure 2.

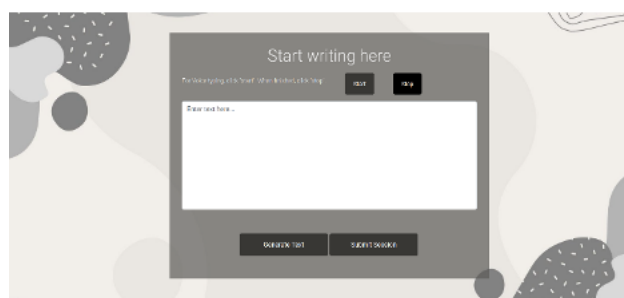


Figure 2: Tamil Co-Writer user interface for input

For ASR, we use a fine-tuned XLSR Wav2Vec2.0 model using the Connectionist Temporal Classification (CTC) algorithm [20, 21]. CTC operates by introducing a unique 'blank' symbol that may be introduced in the target sequence between any two consecutive output symbols. The method then converts each potential output symbol, including the blank symbol, into a sequence of probabilities. Blank symbols were employed during decoding to calculate the most likely output sequence given the input sequence. The ability to accommodate the input and output sequences of varying lengths is one of the CTC's main features, making it ideal for voice recognition jobs where the input and output sequences' lengths might change. Moreover, CTC is capable of handling situations where the alignment between the input and output sequences is not known beforehand, such as when the input voice signal could have pauses or other disturbances.

For fine-tuning the XLSR Wav2Vec2.0 model for better performance for Tamil, we use the common voice open dataset [22]. This dataset has over 850 voices and 7.83GB worth of data. For training data that is this big, a powerful machine having high GB and RAM for running non-stop for a minimum of 40 hours is required. In this research, we employ Google Colab Pro with 100 computer units to train the data. The dataset was cleaned by removing special characters. The audio data used to train the XLSR-Wav2Vec2 model was captured at 48 kHz for Babel, Multilingual LibriSpeech (MLS), and Common Voice and then down-sampled to 16 kHz. Common Speech data needed to be down-sampled to 16 kHz for training as it was initially captured at 48 kHz. The Wav2Vec2FeatureExtractor feature extractor was used with a feature size of 1, sampling rate of 16000, and a padding value of zero. Batchwise padding was performed on the training samples to obtain the longest sample. The feature vector size, which is the combined dimension of all the features derived from the speech representations, is 1. Table 1 shows the hyperparameters of the model.

Table 1
Hyperparameters for the fine-tuned Tamil ASR model

Parameters	Value
Epochs	20
Batch size	16
Gradient accumulation steps	2
Evaluation strategy	Steps
Half-precision floating point format (FP16)	True
Save strategy	Epoch
Evaluation steps	100
Logging steps	10
Learning rate	0.0001
Attention dropout	0.1
Hidden dropout	0.1

The input Tamil text is then converted to English for processing using the existing Google translate API [23], google trans of version 4.0. A web service enables programmers to include translation capabilities into their programs by querying the API built by Google. This Google Cloud Platform-provided API offers a straightforward REST-based user interface for translating text between languages. The query to the Google Translate API contains the text to be translated and the destination language, and the translated text in the intended language is returned by the API. The API offers access to both neural and statistical machine translation models and covers over 100 languages, including some less widely spoken languages. By selecting glossary words or offering their own translation models, developers may also alter the translation results. After translating the input Tamil text to English, Tamil Co-Writer sends it to the GPT-2 text generator to generate suggestions for the writer.

The accuracy of ASR or machine translation was gauged using the word error rate (WER) measure. As compared to the total number of words in the reference text, WER reflects the proportion of words that the system erroneously recognizes or translates [24]. A lower WER score generally denotes a higher degree of system accuracy with respect to speech recognition or machine translation. However, it is important to note that it is not always a true indicator of the system's performance as there might be other contextual factors in the language that are better assessed by a human native speaker. During the manual evaluation of some examples of our Tamil ASR, we observed errors in a few characters and in some cases, the affix and suffix. Many of the predictions are close to ground truth phonetically even

though there are errors in the written script, and Tamil is a phonetic language. An additional metric could be introduced in the future to measure the phonetic proximity, that is comparing the phonetic content in addition to than written content. In the fine-tuned ASR model we used, we achieve a WER of 60%, which is good for basic applications in low-resource languages such as Tamil that lack huge training sets. While noting the limitations, we proceed to use this model for translation for the current prototypical version of our tool, as the main objective of this research is not to enhance ASR in Tamil, but rather to showcase how it can be used in an AI-based writing aid.

3.3. Automated text generation module

The automated text generation module in Tamil Co-Writer generates Tamil text suggestions to the writer using the open-source large language model GPT-2. The large-scale neural language model Generative Pre-trained Transformer 2 (GPT-2) was created by OpenAI as an extension of the original GPT model and performed better across a range of NLP tasks due to its substantially greater size [25]. The Transformer architecture, a kind of neural network that processes input sequences via self-attention techniques serves as the foundation for GPT-2. The model is pretrained on a sizable corpus of text data using an unsupervised learning technique, allowing it to pick up on the linguistic patterns and structures without direct supervision. Once trained, GPT-2 produces writing that resembles that of a human being by predicting the following word in a series based on the preceding ones using probabilities. It can also be applied for specialized jobs such as text categorization, question-answering, and language translation. The capability of GPT-2 to produce cohesive fluid language that is difficult to distinguish from human written material is one of its standout characteristics. Table 2 shows the parameters used for the GPT-2 model to evaluate and generate text suggestions for the writer. While newer models can provide much better performance (and are discussed as part of future work), GPT-2 provides a baseline model for testing that is free to access and deploy.

Table 2
Parameter set for text generation in GPT-2 model

Parameters	Value
Maximum length	30
Number of return sequence	5
Temperature	0.3
Top-p	1
No repeat ngram size	2

3.4. Metrics generation module

The tool additionally logs the keystroke-level action of users to analyse and provide summaries of their interactions with AI suggestions using the metrics generation module. This kind of learning analytics from user logs of AI interaction helps users reflect on their process of writing and reliance on AI, which is deemed to be the future of assessment [25]. The visualization follows CoAuthorViz, which is a graphical representation introduced by recent work [5], to represent co-authorship behaviors during a writing session when users interact with AI suggestions at the sentence level.

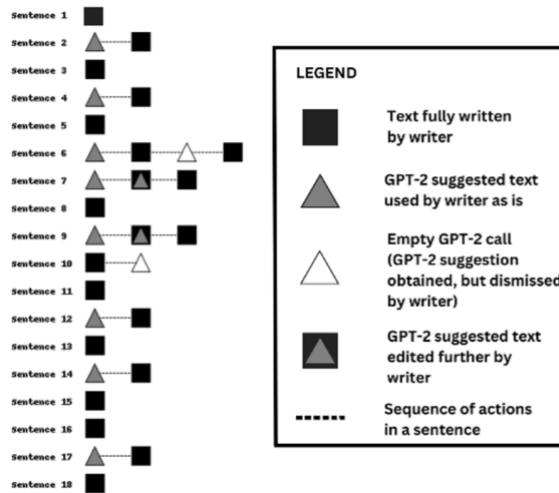


Figure 3: Example of a CoAuthorViz writing session

The CoAuthorViz visual representation and its interaction metrics provide insight into how writers utilize AI writing assistants and can help writers reflect on their dependence on AI suggestions in their writing practices. This was used to analyze human-AI collaborative writing for English, and how the graph can be read is explained in detail in the previous paper with examples [5]. Here we illustrate CoAuthorViz for our Tamil Co-Writer tool in Figure 3 using an example from a test writer. In this case, majority of the writing was done independently by the writer (sentences 1, 3, 5, 8, 10, 11, 13, 15, 16, and 18 with black squares), and even when text from GPT-2 was obtained (sentences 2, 4, 6, 7, 9, 12, and 18 with GPT-2 written text), the writer still added more text. Additionally, we can identify instances when a GPT-2 call was made, but the writer disregarded it (white triangles in sentences 6 and 10 depict empty GPT-2 calls). We also see instances where the writer made changes to the phrases recommended by GPT-2, suggesting partial satisfaction with the suggestions and further editing (squares enclosing gray triangles in sentences 7 and 9).

In addition to the creation of CoAuthorViz for Tamil, we present a summary of the most important events noted in each writing session, which offers concrete metrics that authors may use with visuals to learn more about their writing patterns, as demonstrated in a previous work [5]. This summary includes three metrics: sentence, API, and ratio. The sentence metrics define co-authorship across all sentences in the writing session, the API metrics keep track of the number of calls made to obtain GPT-2 suggestions and the Ratio metrics consolidate them to study co-authorship behaviours in relation to the total number of sentences generated in a writing session. Table 3 shows key co-authorship metrics for the example writing session discussed in Figure 3. The author has written independently for most of the writing session as indicated by their Autonomous Writing Indicator ($RB = 0.61$). The total usage of GPT-2 in a sentence being low ($RC = 0.38$) indicates that the user has a more independent writing style; or is not too satisfied with the AI-offered suggestions. In none of the sentences was the user completely dependent on the GPT-2 suggestion ($SC = 0$). These metrics can help writers understand how much of their writing is produced by the model and how much their own work is at the end of a writing session.

Table 3
Co-authorship metrics for the sample writing session in Figure 3

Parameters	Value
Total number of sentences (SA)	21
Number of sentences completely authored by the writer (SB)	13
Number of sentences completely authored by GPT-3 (SC)	0
Number of sentences co-authored by GPT-3 and writer (SD)	8
GPT-3 dependence indicator (RA)	0
Autonomous writing indicator (RB)	0.61
Total GPT-3 usage in sentences (RC)	0.38

4. Discussion and future work

Despite the increasing popularity of large language models in NLP based applications, low-resource languages such as the regional Indian language Tamil have a lower adoption rate. In this work, we introduced a novel human-AI collaborative writing tool prototype for Tamil writing called Tamil Co-Writer that used GPT-2 to auto-generate text suggestions to the user when they require help in their writing session. Additionally, we introduced a fine-tuned ASR model for Tamil with 60% WER, which enables the user to provide Tamil speech as an input for improved accessibility. We also demonstrated the use of a visual representation called CoAuthorViz and its co-authorship metrics applied through our tool for writers to improve their understanding of AI usage when co-writing with AI. We posit that AI-generated suggestions can not only help language writers obtain new ideas when writing in Tamil, for instance, in coming up with characters and plots for creative writing outputs such as stories, but they can also provide them with exposure to new vocabulary and different ways of structuring sentences.

In the current prototypical version of our tool, we employed the GPT-2 model for equitable access, because it is an available open-source for deployment. However, newer models created with larger training data, such as higher versions of GPT, can provide better text suggestions for the user (with the downside of costs). We plan to use open-source versions of large language models such as LLaMA or Falcon LLM in later versions of the tool when deploying to users. Exciting new developments such as the introduction of an early-stage LLM for Tamil (fine-tuned from LLaMA with 16,000 Tamil tokens using the LoRA methodology) are starting to occur [17]. LoRA introduces trainable low-rank matrices into specific layers of a pre-trained model, reducing the need for pre-training a large number of weight parameters directly and achieving higher training efficiency. We see this as a synergistic development and a potential opening for an NLP ecosystem in Tamil. If open-source LLMs such as these are developed for global regional languages, tools such as Tamil Co-Writer would be further augmented for supporting learners. Together, they could add to a more inclusive and diverse linguistic AI landscape and ecosystem.

Currently, the Tamil Co-Writer tool also uses Google API for translation purposes with trial access; in future versions, we would like to use better and open-source translation models specifically tailored for Tamil to deploy the tool without paid access to API. This would also improve the user experience for writers with faster processing times, as there would not be any time lost in the back-and-forth translation through APIs. There is also scope to create a better- fine-tuned model for ASR by increasing the run time, input data, and training on heavy resources. Future work will involve additional support for code-switching between languages [28] for multi-lingual usage and improved user interaction features.

Learning analytics offers new opportunities to support effective human-AI collaboration by helping educators and students become aware of the processes involved in learning, in addition to the final products. In current work, in addition to using AI to support their writing, users can also reflect on their usage of AI-generated suggestions using the CoAuthorViz visualisation and metrics. The creation of visual representations to study co-authorship behaviours and associated metrics opens up new avenues to investigate writing processes, such as studying user characteristics and collaboration dynamics among writers [5]. Analytics from generative AI can thus help close the LA cycle by facilitating personalised and adaptive interventions [27]. Writers may use this data to spot patterns and trends in their writing processes, such as if they frequently employ AI for particular kinds of material. This language-agnostic approach can help improve students' writing practices and feedback-seeking behaviors when engaging with AI by understanding their own and AI's respective roles in the writing process, and aids researchers to study these processes.

Future work can also inform feedback mechanisms to provide feedback to the user when over-reliance on AI is observed in their writing and develop effective models for optimal human-AI collaboration for writing. The research thus provides useful insights from our prototype evaluation that can be extended to other languages to cater to diverse groups of audiences and their writing needs for language development and can pave the way for more inclusive AI tools for education in the future.

References

- [1] K. I. Gero, V. Liu, L. Chilton, Sparks: Inspiration for Science Writing using Language Models, in: Proceedings of the 2022 ACM Designing Interactive Systems Conference, DIS '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 1002–1019. doi:10.1145/3532106.3533533

- [2] J. Kim, S. Suh, L. Chilton, H. Xia, Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing, in: Proceedings of the 2023 ACM Designing Interactive Systems Conference, DIS '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 115–135. doi:10.1145/3563657.3595996
- [3] M. Lee, P. Liang, Q. Yang, Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities, in: Proceedings of the 2022 CHI conference on human factors in computing systems, 2022, pp. 1-19.
- [4] Wikipedia.org, Tamil language, 2023. URL: https://en.wikipedia.org/wiki/Tamil_language
- [5] A. Shibani, R. Rajalakshmi, F. Mattins, S. Selvaraj, S. Knight, Visual representation of co-authorship with GPT-3: Studying human-machine interaction for effective writing, in: M. Feng, T. K\"aser, and P. Talukdar 16th International Conference on Educational Data Mining, Bengaluru, India, 2023
- [6] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, J. P. McCrae, Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text, in: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL) Marseille, France, 2020
- [7] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, S. Gupta, Muril: Multilingual representations for indian languages. arXiv preprint arXiv:2103.10730, 2021
- [8] S. Knight, A. Shibani, S. Abel, A. Gibson, P. Ryan, N. Sutton, R. Wight, C. Lucas, A. Sandor, K. Kitto, M. Liu, AcaWriter: A learning analytics tool for formative feedback on academic writing, 2020
- [9] O. Kruse, C. Rapp, C. M. Anson, K. Benetos, E. Cotos, A. Devitt, A. Shibani, Digital writing technologies in higher education: theory, research, and practice, in: Springer Nature, 2023
- [10] A. Gibson, A. Shibani, Natural Language Processing-Writing Analytics, by Charles Lang, George Siemens, Alyssa Friend Wise, Dragan Gašević, and Agathe Merceron. 2nd ed. Vancouver, Canada: SoLAR, 2022, pp. 96-104. URL: https://solaresearch.org/wp-content/uploads/hla22/HLA22_Chapter_10_Gibson.pdf
- [11] A. Shibani, Analytic Techniques for Automated Analysis of Writing, Digital Writing Technologies in Higher Education: Theory, Research, and Practice, 2023, pp. 317-331.
- [12] M. Lee, P. Liang, Q. Yang, Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities, in: Proceedings of the 2022 CHI conference on human factors in computing systems, 2022.
- [13] N. Singh, G. Bernal, D. Savchenko, E. L. Glassman, Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence, in: ACM Transactions on Computer-Human Interaction, 30(5), 2023, pp. 1-57.
- [14] A. Riddle, A hybrid approach to co-creative story authoring using grammars and language models, in: Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, 2022.
- [15] W. Du, Z. M. Kim, V. Raheja, D. Kumar, D. Kang, Read, revise, repeat: A system demonstration for human-in-the-loop iterative text revision, arXiv preprint arXiv:2204.03685, 2022.
- [16] Grammarly, How we use ai to enhance your writing, Grammarly Spotlight, 2019. URL: <https://www.grammarly.com/blog/how-grammarly-uses-ai/>
- [17] A. Balachandran, Tamil-Llama: A New Tamil Language Model Based on Llama 2, arXiv preprint arXiv:2311.05845, 2023.
- [18] R. Errattahi, A. E. Hannani, H. Ouahmane, Automatic speech recognition errors detection and correction: A review, in: Procedia Computer Science, 128, 2018, pp. 32-37
- [19] Quillbot.com, URL: <https://quillbot.com/>
- [20] A. Baeovski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems, 33, 2020, pp. 12449-12460
- [21] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: Proceedings of the 23rd international conference on Machine learning, 2006.
- [22] Common Voice Mozilla, URL: <https://commonvoice.mozilla.org/en>
- [23] Google Cloud Translation AI, URL: <https://cloud.google.com/translate?hl=en>

- [24] A. Ali, S. Renals, Word error rate estimation for speech recognition: e-WER, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018
- [25] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI blog, 1(8), 9, 2019.
- [26] J. M. Lodge, S. Howard, M. Bearman, P. Dawson and Associates, Assessment reform for the age of artificial intelligence, A. G. Tertiary Education Quality and Standards Agency, 2023. URL: <https://www.teqsa.gov.au/guides-resources/resources/corporate-publications/assessment-reform-age-artificial-intelligence>
- [27] L. Yan, R. Martinez-Maldonado, D. Gašević, Generative Artificial Intelligence in Learning Analytics: Contextualising Opportunities and Challenges through the Learning Analytics Cycle, arXiv preprint arXiv:2312.00087, 2023
- [28] N. Jose, B. R. Chakravarthi, S. Suryawanshi, E. Sherly, J. P. McCrae, A survey of current datasets for code-switching research, in: 2020 6th international conference on advanced computing and communication systems (ICACCS), 2020