BRITISH COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

Research Report

The Language of DALL·E: A multimethod approach to technological integration into language classroom and assessment practices

A compilation report by Jordan Weide, Johnathan Cruise and Emil Tangham Hazelhurst

Preface

Introduction

Generative AI and similar emerging technologies are increasingly being discussed in the context of the processing of human language (Leivada et al., 2022). DALL·E 2 and other AI image generators are capable of producing unique images from text prompts using a huge catalogue of online source data. Multiple semantic elements contained in the lexical and syntactic components of texts prompts can be recreated in visual format signifying a large step forward in generative AI technology. As tools such as DALL·E 2 develop, their potential applications in language teaching and assessment will become more relevant. Major testing institutions are working on incorporating AI into testing and educational materials, and it is thus imperative to maintain a clear understanding of the capabilities of such tools.

While previous studies have investigated the relationship between prompt and output in DALL·E 2 (Conwell & Ulman, 2022; Leivada et al., 2022; Marcus et al., 2022; Thrush et al., 2022), this may be the first study with an explicit focus on the applications of generative image AI tools in language teaching and assessment. The three strands of this paper approach the analysis of the tool from different perspectives using a range of methods. The aim is to provide a broad overview of its capabilities for the purposes of language assessment and discuss the role of generative AI technologies in the field.

Abstract

This paper documents the investigative process of a multi-method approach to the evaluation of new technologies in language teaching and assessment. The study involves an investigation into the capabilities of the generative AI tool DALL·E 2. Three researchers, working in tandem, but on three distinct, related investigations explored various components of the tool. The focus of this report includes both the findings of the investigations, as well as documentation of the collaborative research approach employed. It is worth noting here that, regardless of what aspect of inquiry each of the three researchers adopted and focussed on, the main research aims centred on four overarching lines of inquiry.

The first line of inquiry was, “What form and content of language is required to optimise DALL·E’s ability to produce images according to the prompt writer’s intent?”

The second line of inquiry was, “How does manipulation of syntax and lexis work on DALL·E’s image creation process in an iterative process?”

The final lines of inquiry were, “How can prompt writing, using an iterative process, be effectively utilised or operationalised for pedagogical/academic research or applications?” and “What potential learning or assessed activities could work well using this procedure?”

These lines of inquiry were predicated on the idea that language assessment and teaching should embrace new technologies in their practices. The integration of these technologies requires a systematic review to ensure their implementation can be longstanding and effective. The small-scale multi-method approach allowed for exploratory investigations into a range of areas, each of which could provide the basis for further research. Within each method or strand of investigation, we hope to come closer to actionable insights which lead to larger scale interrogations of such potential augmentations in learning and creativity.

Table of Contents

1.0 Strand 1: Exploring the Integration of DALL·E 2 for Enhancing Learners' Awareness of English Varieties.

1.1 Strand Rationale

The integration of new technologies into pedagogy has rapidly been brought to the forefront as machine learning increasingly effects daily practises. (Mollick & Mollick, 2022; Mollick & Mollick, 2023). “The advances in machine intelligence, raise a question as to how this technology may be leveraged as used a scaffolding tool for second language assessment and classroom practises” (Leivada et al., 2023). An adaptation to use, of machine learning in other systemic organisations, pushes teachers and other educational stakeholders to also adapt to technological advancements.

However, before a widespread application of the use of machine learning in pedagogical and assessment contacts can be enacted, such applications need to be tested to determine if they are fit for purpose. Although “machine intelligence is increasingly being linked to claims about sentience, language processing, and an ability to comprehend and transform natural language into a range of stimuli” (Leivada et al., 2023) researchers, teachers and stakeholders alike should validate such claims.

The current strand serves to demonstrate a manner of trialling machine learning applications (DALL·E 2) to ensure tools fit for purpose within a classroom context. This strand aims to develop classroom applications for DALL·E 2 to raise language learners' awareness of different English varieties to promote international communicative competence. “Intercultural communicative competence, or ICC, refers to the ability to understand cultures, including your own, and use this understanding to communicate with people from other cultures successfully” (British Council, 2023). International communicative competence is important to help English language learners integrate into their desired context.

This research aim remains a relevant area of investigation due to the changing nature of the English language and how it is taught. “English is currently the most spoken language across the globe, with roughly 2.3 billion people speaking it as a first or additional language” (Patel, 2023). As the range of speakers come from a wide variety of backgrounds and with widespread integrations of cultures, use of English varies by context. What is more, "change is happening: The Future of English: Global Perspectives report notes that there has been a gradual, industry-led, shift away from the ideal of mother tongue fluency towards a more applied and contextualized approach to language proficiency” (Patel, 2023). Although the current research is only able to focuses on the American and British varieties of English, the manner of trailing is meant to serve as a guideline on how to aid English language learners in achieving greater

international communicative competence in whatever English language context they are looking to join.

"American and British cultures are fundamentally different in their histories, economies, environments, and social organizations. Although spoken American and British English are generally mutually intelligible, there are enough differences to cause misunderstandings or even a complete failure of communication" (Di Carlo, 2013, p. 62). When considering the reality of differing language context, it becomes increasingly important to raise learners' awareness of English language varieties. "As language is the soul of a culture, major awareness can be a strategy to improve mutual intelligibility and respect between the two cultures." (Di Carlo, 2013, p. 74). Therefore, in an increasingly globalised society, variety awareness is still of importance. Language learners', who desire to use English in different contexts, should include variety awareness into their education to aid their success in achieving communicative competence.

"Although the topic is an interesting one and likely to attract researchers, the relevant literature, especially concerning the learners" awareness of the differences between varieties of English, is not as vast as expected." (Yasmin, 2015, p.155). "The study of Elkılıç and Han (2009) that was carried out with the participation of 42 undergraduate students attending the English Language Department revealed that the participants were able to distinguish American and British English in terms of pronunciation differences, but they displayed poor performance concerning spelling, vocabulary, and grammatical differences." (Yasmin, 2015, p.156) Even with the importance of raising the awareness of differences between British and American English varieties, for greater international communicative competence, there is a lack of vast research into this topic.

Based on the previous works addressed above, the current study strand will focus on developing a pedagogical application for raising awareness of spelling, vocabulary, and grammatical differences between American and British varieties of English. The American and British varieties were chosen based on the knowledge of the varieties the researchers were able to provide. Two of the researchers in the study are L1 British English speakers while the other is an L1 American English speaker.

DALL·E 2 was chosen for this study due to its accessibility. Because this strand is looking to use its findings to shape pedagogical practises, researchers focused on using tools that teachers and students can readily access without a paywall to promote widespread use and closer practice equality across economic backgrounds. Furthermore, the visual element that DALL·E 2 provides creates an additional sensory aspect for learning, which traditional pedagogical techniques may fail to provide. Including a visual element allows the activity to better reach learners who prefer a visual learning style.

At the time of this research, DALL·E 2 was the latest version available. All images created through DALL·E 2 in this report were completed between June and September

2023. Therefore, the images were created prior to the subsequent DALL·E 3 version was released to the public in October 2023. Only two images in this strand report were not created by DALL·E 2. They are clearly labelled as being retrieved from Google stock images. All images created with DALL·E 2 have retained the DALL·E 2 watermark in the bottom right corner.

The current strand serves to demonstrate a manner of trailing machine learning applications (DALL·E 2) to ensure its fit for purpose within a classroom context. This strand specifically focuses on raising awareness of English variety differences between American and British English to aid language learners' international communicative competence through a visual manner in a readily accessible platform, DALL·E 2. It investigates the predictability of eliciting specific cultural references within images prompts for DALL·E 2. Therefore, the following research questions were derived to determine how this technology may be leveraged for pedagogical application.

1.2 Research Questions

RQ1: How does the manipulation of prompts according to words from either British English or American English varieties affect DALL·E 2 visual outputs?

RQ2: How does the manipulation of prompts according to spellings from either British English or American English varieties affect DALL·E 2 visual outputs?

RQ3: How does the manipulation of prompts according to grammar from either British English or American English varieties affect DALL·E 2 visual outputs?

RQ4: Can a pedagogical application be developed to address international communicative competence through visual demonstration of British and American English varieties?

1.3 Investigatory Approach

1.3.1 Emergence of Cultural Language Use

During the studies initial phases, the researcher's linguistics backgrounds seemed to produce differences in visual outputs. There was a varying reliance on spelling and word choice (vocabulary) used within the prompts. An emergence of cultural language use was seen due to visual discrepancies appearing within the images produced along the researcher's own English-speaking contexts. However, not all prompts were subject to these discrepancies.

There were two prompts in particular that were subject to the initial emergence of cultural language use. The first prompt focused on a piece of furniture used within the home as demonstrated in the image taken from Google stock images; Figure 1.0.

The photograph was selected from the internet, as was intended to be replicated by DALL·E 2 simply based on written prompts provided by the researchers. DALL·E 2 was not given the image.

Fig. 1.1.0 Google stock image of furniture piece

The differences in English variety use were actualised with the word choice for the furniture piece. The researchers using British English used prompts such as; *Picture of an industrial looking room with half of a modern white* ***sofa*** *and a houseplant on a stand positioned to its left* (Figure 1.3). There was then a changing of word choice from sofa to couch in the prompt used by the researcher using American English; *Picture of an industrial looking room with half of a modern white* ***couch*** *and a houseplant on a stand positioned to its left* (figure 1.1.2).

Fig. 1.1.2 Picture of an industrial looking room with half of a modern white couch and a houseplant on a stand positioned to its left, American

Fig. 1.1.3 Picture of an industrial looking room with half of a modern white sofa and a houseplant on a stand positioned to its left, British

The original image includes a specialized tufting technique in upholstery. This style was not represented in prompts including the word sofa and yet appeared when couch was used. This was true in subsequent prompts even when the upholstered buttons are directly stated in the prompt. The descriptive language and word choice taken from the above exploratory prompts point to a need for further investigation into how word choice based on English varieties within a prompt is reflected in the image that DALL·E 2 produces.

The second prompt focused on an image of a nature landscape as demonstrated in photograph taken from Google stock images; Figure 1.2.0. As with the previous image, the photograph was selected from the internet intended to be replicated by DALL·E 2 simply based on written prompts provided by the researchers. DALL·E 2 was not given the image.

Fig. 1.2.0 Google stock image mountain scene

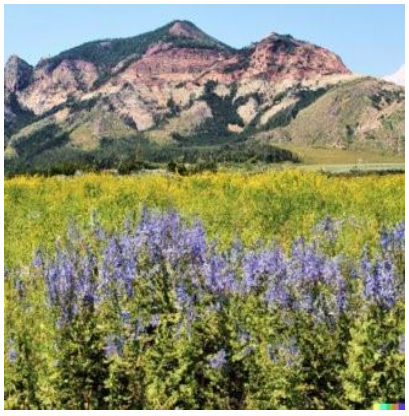
The differences in English variety use were actualised with the spelling of a word. In looking at the prompt where the American English variety spelling was used, *Picture with small rocky mountains in the background and a field of many* ***colored*** *wildflowers stretching across the foreground*, the image reflects a specific mountain range within

the United States called the Rocky Mountains (figure 1.2.2). There was no capitalisation to indicate this distinction; however, the American spelling of color was used. It is possible that the American English spelling indicated a specific context where the collocation “rocky” with “mountains” holds meaning outside of just descriptors of the formation type, but rather a location.

To further investigate this context setting the sequential prompt was adjusted by using the British English “colour”, and yet maintaining the collocate rocky mountains; *Picture with small rocky mountains in the background and a field of many* ***coloured*** *wildflowers stretching across the foreground.* The images produced reflected a more European presenting mountain range mimicking the appears of the Alps (figure 1.2.1). The inclusion of this second mountain scene is to point out the importance of locational knowledge in determining what the image is depicting. The image seemed to be specialized according to the locational context which is inferred based on variety spelling.

Fig. 1.2.1 Picture with small rocky mountains in the background and a field of many coloured wildflowers stretching across the foreground, British

Fig. 1.2.2 Picture with small rocky mountains in the background and a field of many colored wildflowers stretching across the foreground, American



Therefore, these initial prompts indicated there may be a fruitful area of investigation into how DALL·E 2 is able to create visual distinct images based on English varieties used within the prompts. However, for DALL·E 2 to be applied effectively it needs to generate predictable images. Thus, the researchers had to develop a way to systematically investigate the predictability of eliciting specific cultural references within images prompts for DALL·E 2.

1.3.2 Prompt Selection

Selection of prompts was important in determining whether the predictive and reproducibility nature of the constructs in research questions 1-3 could be operationalised through the methods. A major consideration in prompt selection was the movement of terms as described by Di Carlo (2013). The movement of previously distinct American terms into both British and American spoken and written language challenges the durability of identified cultural references. It was imperative to ensure the images produced were visually distinct.

The researchers had to be confidence that the construct they were investigating, word choice, spelling and/or grammar, had to elicit an image that was visually distinct enough to determine to be a direct cause the linguistic variation. The need for visual distinctness is demonstrated by figure 1.3 and figure 1.4. Although the word used in the prompt (jumper or sweater) was distinct to the English variety, the concept covered was not visually different. Therefore, the researchers needed to also consider the concept in each cultural context.

Fig. 1.3 Cartoon Elephant in a jumper eating lunch, British

Fig. 1.4 Cartoon Elephant in a sweater eating lunch, American

Considering the above restrictions, the researchers consulted previous literature on American and British English varieties to create prompts. Di Carlo (2013) listed vocabulary varieties differences that are seen within the topic areas of house, transport, shopping, food and numbers. Both spelling and Grammar differences were listed on the British Council website (https://www.britishcouncilfoundation.id/en/english/articles/british-and-american-english). Spelling differences between British and American English: *Through-thru Night-nite Light-lite High-hi* could be accessed through the Oxford international English website (https://www.oxfordinternationalenglish.com/differences-in-british-and-american-spelling/). It was then, based on previous research, and a need for visual distinctness the prompts were created.

1.4 Results

This research strand work to determine the predictability of DALL·E 2 eliciting visual distinct images as a direct cause of English variety use in its prompts. While the results indicate a limited degree of predictability, they indicated an inclination towards American English in its images.

In relation to research question one, this study investigated how the manipulation of prompts according to words, vocabulary, from either British or American English affect DALL·E 2 visual outputs. The word choice did have an effect on visual outputs. The variation was prominent in prompts that used differing words that share equivalent meanings. This is demonstrated below in figure 1.5 and figure 1.6. All other word choice prompts used in the study demonstrated an inability to elicit reliable results.

Fig. 1.5 Image looking out a flat window, British

Fig. 1.6 Image looking out an apartment window, American

In relation to research question two, the study investigated the manipulation of prompts according to culturally specific word spellings from either British or American English affect DALL·E 2 visual outputs. Spelling choices can elicit culturally specific images, particularly when coupled with that English variety's specific vocabulary. *The following prompt is an example of the British variety; Picture of a town street, with trees changing* ***colours*** *for* ***autumn***. The prompt was then adjusted to include American variety spelling and word choice; *Picture of a town street, with trees changing* ***colors*** *for* ***fall***. Figure 1.7 and Figure 1.8 demonstrate the visual differences. Visual distinctness was not seen when word choice was paired with the variety spellings of the terms "offense" and "flavour". However, noteworthy is the initial stage of the research, wherein the presence of the term "colour" alone sufficed to alter the image.

Fig. 1.7 Picture of a town street, with trees changing colors for fall., American

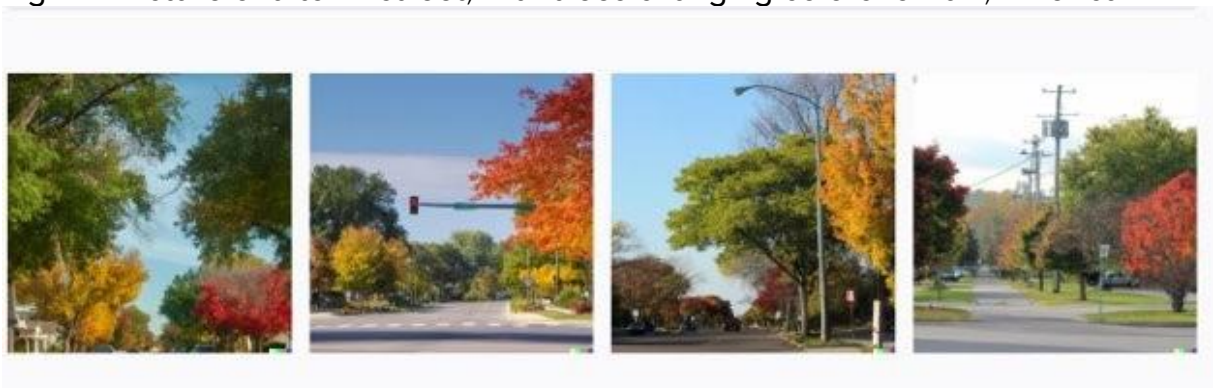


Fig. 1.8 Picture of a town street, with trees changing colours for autumn., British

In relation to research question three, the study investigated the manipulation of prompts according to grammar from either British English or American English varieties affect DALL·E 2 visual outputs. The results demonstrated that grammar could elicit culturally specific images when associated with linguistic patterns specific to certain cultural contexts. An example of this phenomenon is shown through the following prompts, that focus on article use. The British variety of the prompt read; *Image of a student at university in fancy dress.* The American variety of the prompt read: *Image of a student at* ***the*** *university in fancy dress.* Figure 1.9 and figure 1.10 demonstrate visual distinctness based on article use. This aligns with prior research, affirming that grammar in isolation does not yield consistent image outputs. (Leivada et al, 2023)

Fig. 1.9 Image of a student at the university in fancy dress, British

Fig. 1.10 Image of a student at university in fancy dress, American

In relation to research question four, the development of pedagogical application to address international communicative competence through visual demonstration of British and American English, the study worked to find predictability for application. Although there were instances where English variety use produced visual distinctness; word choice, spelling and combination of the previous two with grammar, there did not appear to be widespread predictability in DALL·E 2 outputs. This lack of predictability does not lend itself for pedagogical applications.

1.5 Discussion

The result of the current study strand showed that eliciting specific cultural references in the images from a prompt has little to no predictability. However, DALL·E 2 output indicated a tendency towards American English.

English variety word choice is seemingly the only singular element that elicits culturally specific responses. Even still, the vocabulary with accurate results came from different words that hold the same meaning but actualize in different manners. All other vocabulary did not produce reliable visual distinct results from DALL·E 2. English variety spellings seem to elicit visually distinct images from DALL·E 2 when paired with the vocabulary from the same English variety. English variety grammar seems to elicit visually distinct images from DALL·E 2 when placed within a specific cultural context. In agreement with previous studies (Leivada et al., 2023) grammar alone does not provide reliable outputs.

These findings are in line with previous research. “Whereas young children routinely master these phenomena, learning systematic mappings between syntax and semantics, DALL·E 2 is unable to reliably infer meanings that are consistent with the syntax of the prompts. These results challenge recent claims concerning the capacity of such systems to understand human language.” (Leivada et al, 2023, p.1). Therefore, DALL·E 2, in the state that it was at the time of this research, is not fit for purpose as a pedagogical tool for teaching international communicative competence by using visual demonstrations of differences in English varieties to raise learners' awareness of English varieties.

1.5.1 Limitations

Limitations faced in the current study strand focus on the capability of DALL·E 2 at the time of this research study took place. Leivada et al (2023) also faced limitations due to AI capabilities. Their research states, “There are three possibilities that can explain LaMDA’s behaviour: (i) AI systems show an indifference to truth, actively promoting plausible misinformation, (ii) AI systems have not yet mastered the conditions that license the appropriate use of a grammatically loaded element such as the pronoun

'us', or (iii) both." (Leivada et al, 2023). This quote calls out the evident limitations, while also very wisely encasing them in "not yet"; a redirect pointing to future capabilities.

Although this study strand was not yet able to aid the development of a pedagogical application to address international communicative competence through visual demonstration of British and American English varieties, the value of its investigation is not lost. The current strand serves to demonstrate a manner of trailing machine learning applications (DALL·E 2) to ensure tools fit for purpose within a classroom context.

The integration of new technologies into pedagogy will only continue to take place. As integration carries on, researchers and practitioners need to become increasingly adaptive whilst not compromising on high standards for application. The current study can be an element of the type of systematic testing new technology must undergo to determine if they are fit for purpose in a pedagogical context. Although DALL·E 2 was not yet able to pass the current testing, AI capabilities will continue to improve and will soon have the capabilities to be effectively applied.

1.6 Conclusion

This study addressed the relationship between cultural references and linguistic variation on text-generated images. The findings further the investigation into nuanced relationships between specific linguistic elements and their capacity to evoke visually distinctive images in DALL·E 2. The studies focus on spoken American and British English, the implications of cultural variations, and the evolving role of artificial intelligence raise vital questions about enhancing mutual intelligibility, improving language learning, and AI's proficiency in language comprehension and generation. Further exploration is warranted to understand these linguistic dynamics within the generative AI context.

2.0 Strand 2: Variable Binding and Output Reliability

2.1 Strand Rationale

In R. England's article, DALL·E (England, 2021) is succinctly described as a text-to-visual engine that forms images using the syntactical or semantic relationships in any given prompt with the help of such encoded methods as "variable binding" (England, 2021).

This method discerns and discriminates interrelational associations between lexico-grammatical elements, mainly by means of proximity between each lexical item within a syntactic unit, and then turns these tied elements into coherent images; perhaps not always as the prompt writer intended.

This is particularly true when there are "more objects…introduced" (England, 2021). Unlike hard-coded rendering engines, DALL·E has a propensity to "fill in the blanks" (England, 2021). For example, when a user writes statements such as "a brown hedgehog wearing a red hat," DALL·E will then "bind" these properties together, making the hat red while the hedgehog brown. However, as the level of syntactic complexity increases within a prompt, and more elements are introduced, DALL·E tends to be less predictable in terms of the expected output. This strand of investigation looked into how incremental additions into a chain of elements affected DALL·E's output and if there was any significant predictability to the changes in the output.

A claim by Ritika (Labeller, 2022) stated that DALL·E , and its successor DALL·E 2, is "significantly better at producing coherent images" when dealing with associated objects "…their characteristics, and their spatial arrangements all at once." Ritika acknowledges that "DALL·E also had the capacity to synthesize items, a few of which are highly improbable to exist in the actual world" (Labeller, 2022). This is seen by Ritika as a boon to the general user, but obviously in terms of constraining output for the teacher or test developer, creates a considerable challenge if the expectation is to have output that is predictably consistent with a specific task design or intended test constructs. Reliability is a "central concern" of test developers (Chapelle, 2013). As Kane states "Reliability is a necessary condition for validity because generalization is a key inference" (Kane, 1992). In terms of teaching and testing, it is a widely held axiom that reliability is as core a concept in testing as validity is, and therefore consistent and predictable output in terms of using DALL·E for tasks in teaching and testing is imperative.

2.2 Investigatory Approach

Using Khalifa and Weir's (2009) central cognitive processing model of reading (fig. 1) as a guide, an incremental approach to prompt engineering was adopted in terms of lexico-grammatical features. Paradigmatic elements were added to the basic syntagmatic structure of the initial prompt template adding increasing complexity at lexical, phrasal, and clause levels (fig. 2). The prompt's syntagmatic sequence and notional concepts were only changed once, however. The reason being that it was felt that the focus would have shifted too far from lexico-grammatical changes to that of semantic ones, which would have added too many variables to the focus of this investigation. It was hoped that focusing mostly on one linguistic area of change, rather than multiple areas, might gain some understanding into DALL·E's "black box" logic. In this regard, it should be noted that it was not the main objective of these investigatory strands to glean DALL·E's

nature or its “innerness,” as J. Curcio terms it in a series of dialectic conversations with ChatGPT (ChatGPT, 2023). In fact, ChatGPT warns Curcio against making “…compelling metaphors when talking about AI, such as myself. From a technical standpoint, you’re communicating with a computer program that uses machine learning to generate human-like text. This program, known as GPT-4, was trained on a large corpus of internet text and uses patterns in that data to generate responses to the inputs it receives based on statistical patterns applied to language” (ChatGPT, 2023). Its caveat against what is essentially anthropomorphism, despite Curcio’s persuasive counterargument that LLM’s may actually be a sort of Gestalt of knowledge mined by GPT from the internet, is one that forms the theoretical assumptions of this investigation. Any other discussion, such as Curcio’s attempts to find its “nature,” is essentially a philosophical and ethical one that will not be discussed in depth here. Merely to say that this type of prompt engineering approach, when implemented at the incremental change level, was hoped to give some insight into predictable output which can then be categorized and then used by teachers, testers, or any who wish to work with text to visual AI tools with any degree of reliability.

Fig.1 Khalifa, H., & Weir, C. (2009). Central processing core of the Khalifa and Weir's cognitive model of reading.

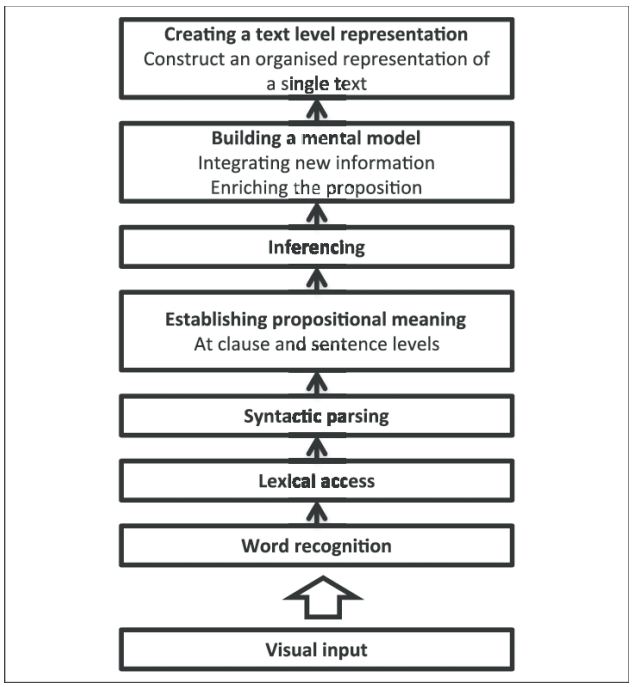
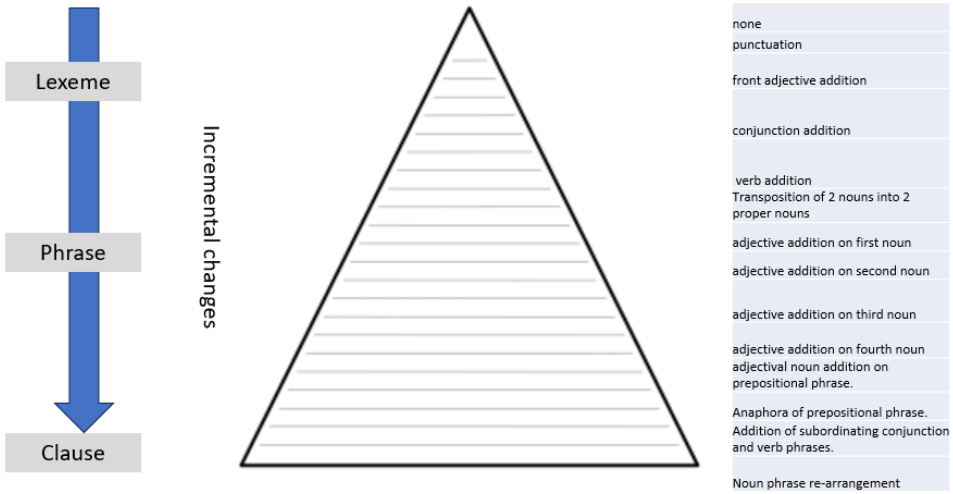


Fig.2 A diagram of Incremental Changes in a Prompt



In terms of concrete output, these incremental lexico-grammatical changes (fig 3) were subsequently compared to notable changes in DALL·E's sets-of-four, iterative, visual output and measured in terms of the degree of change significance (fig 4).

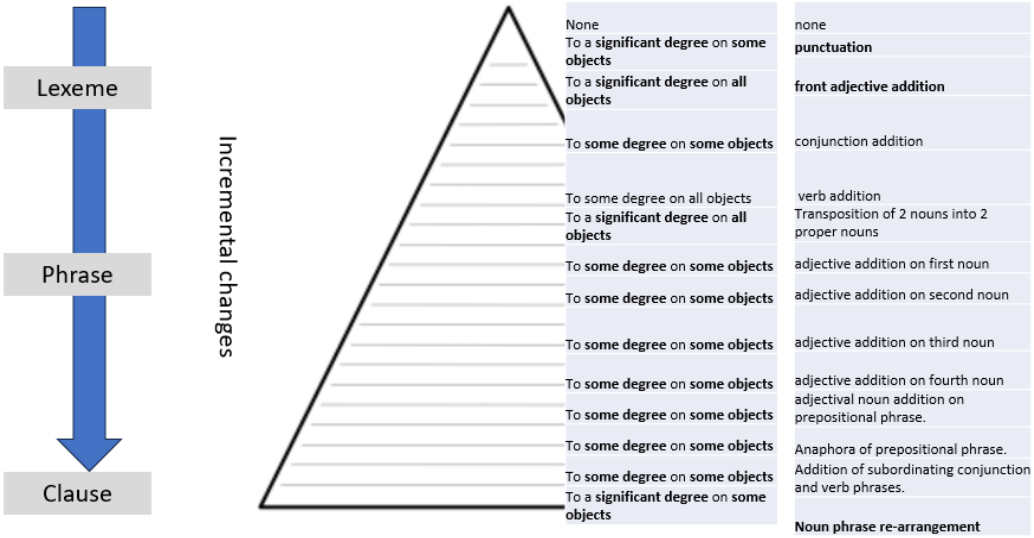
Fig.3 The Paradigmatic Changes in the Prompt Syntagm

Man. Dog. Duck. Mouse. Field.	single words
Man, dog, duck, mouse, field.	list
A tall man. A hairy dog. A green duck. A small mouse. A large field.	noun phrases
A tall man with a hairy dog, a green duck and a small mouse in a large field.	complex noun phrase
A tall man is standing with a hairy dog, a green duck and a small mouse in a large field.	Simple sentence
Arnold Schwarzeneggar is standing with a hairy dog, a green duck and Micky Mouse in a large field.	Simple sentence
A tall, strong man is standing with a hairy dog, a green duck and a small mouse in a large field.	Simple sentence
A tall man is standing with a hairy, fierce dog, a green duck and a small mouse in a large field.	Simple sentence
A tall man is standing with a hairy dog, a green, flying duck and a small mouse in a large field.	Simple sentence
A tall man is standing with a hairy dog, a green duck and a small, scared mouse in a large field.	Simple sentence
A tall man is standing with a hairy dog, a green duck and a small mouse in a large field of flowers.	Simple sentence
In a large field of flowers, a tall man is standing with a hairy dog, a green duck and a small mouse.	Simple sentence
In a large field of flowers, a tall man is standing with a hairy dog as a green duck and a small mouse fly and run around them.	Complex sentence.
In a large field of flowers, a small mouse is standing with a hairy dog as a green duck and a tall man fly and run around them.	Complex sentence.

With respect to initial changes, fig.3 shows how the basic universal nouns of "man", "dog", "duck, and "field" are separated into atomised elements with a period (or full stop) and in the next lexico-grammatical step, the elements are turned into a list of connected items by commas. Thus, meaning changes at a very basic lexico-grammatical level in terms of how each element relates to each other in the prompt. The next alteration is to add an adjectival modifier: "a tall man", "a hairy dog", "a green duck", "a small mouse"

and “a large field”, and so on. The intention was to keep the incremental alterations as near to common lexico-grammatical associations as possible, or within each element’s “semantic proximity”, or vector, in this case, common “adjective+noun” collocations. So, the “man” was associated with being “tall”, an adjective frequently appended to this noun. The only instances where the lexico-grammatical change shifted to that of semantic/conceptual change was when proper nouns were transposed (“Arnold Schwarzenegger” and “Mickey Mouse”), bringing in connotations outside of the lexico-grammatical schemas. Another significant shift from this change pattern is when head noun phrases are shifted around (fig.3). The significance of these pattern shifts will be discussed later in the findings.

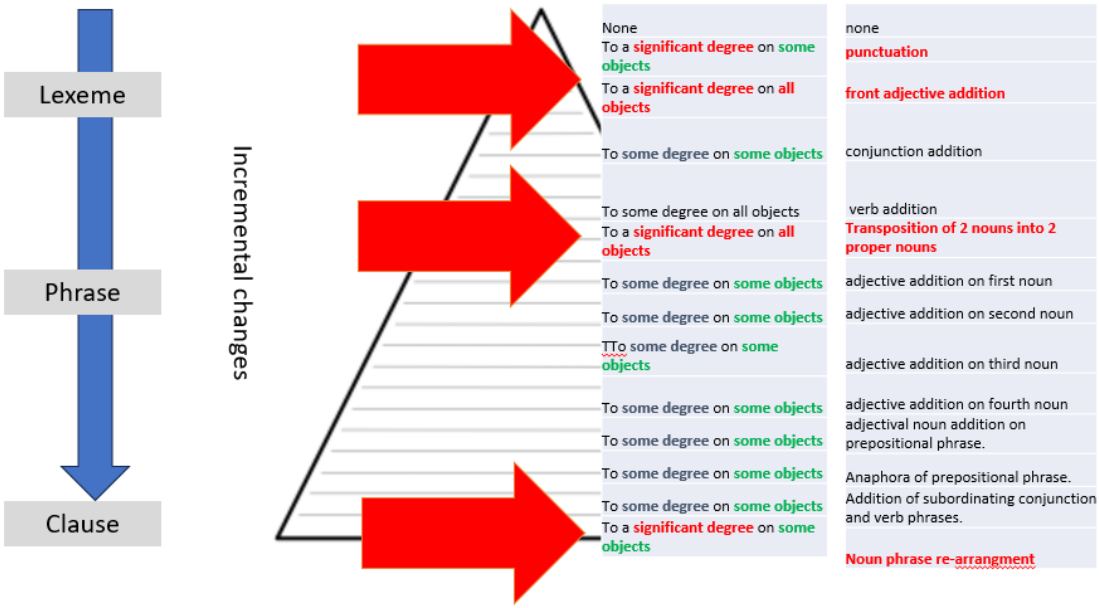
Fig.4 The Significance of Changes between Adjacent Prompt Pairs



2.3 Findings

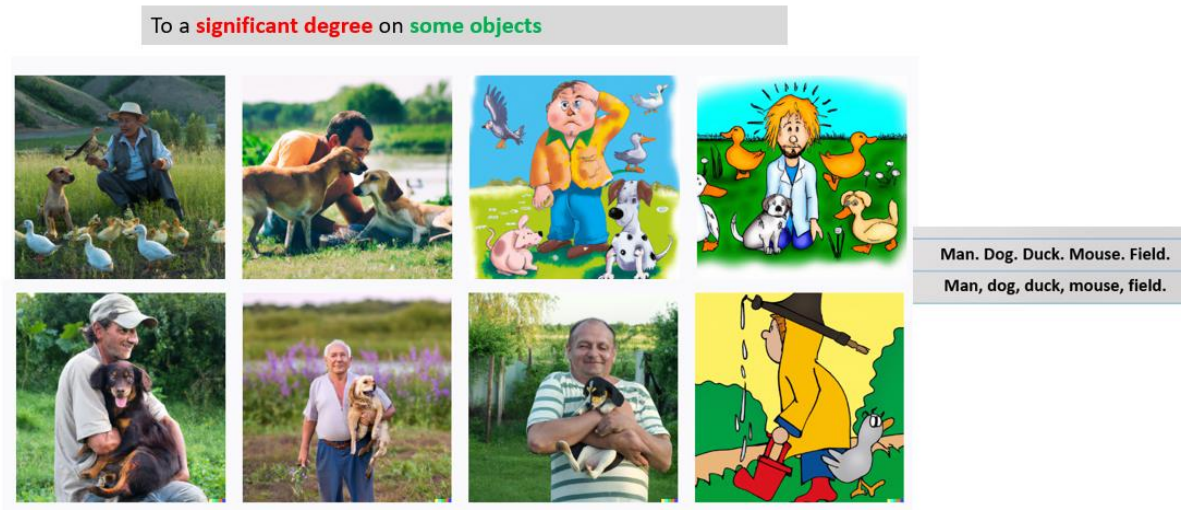
As highlighted in figure 5, the significance of the changes varied to a lesser or greater degree on some or all the objects in the visuals as the prompt grew in complexity.

Fig.5 The Significance of Changes between 3 Adjacent Prompt Pairs



For example, the initial addition resulted in a change where elements noticeably reduced from a varied collection of featured text elements (fig. 6.1). In other words, in the first set of four pictures, a rendering of the man is surrounded by a variety of prompt-featured animals in depictions of a grass field, but in the next set this is reduced to just a man and a dog (or a duck in the last picture) in a kind of field. Why this is so is a matter of conjecture, but the most persistent images are those between the man centred in the field, and then with the dog; possibly suggesting a stronger lexico-grammatical collocation, connection or association than among the other featured prompt elements. These connections could possibly be strengthened by the small change of the comma, and the possible merging of discrete features together as a result.

Fig. 6.1 First comparison of DALL·E ’s output between an adjacent prompt pair with significant changes



Looking at the second significant change (fig. 6.2), it can be seen that the transposition of nouns into proper nouns causes significant alterations in the images rendered:

Fig. 6.2 Second comparison of DALL·E’s output between an adjacent prompt pair with significant changes

One can see in the first of the adjacent prompts that the genre or style of a children’s cartoon is mainly in pastel or primary colours and the style is very similar in each of the

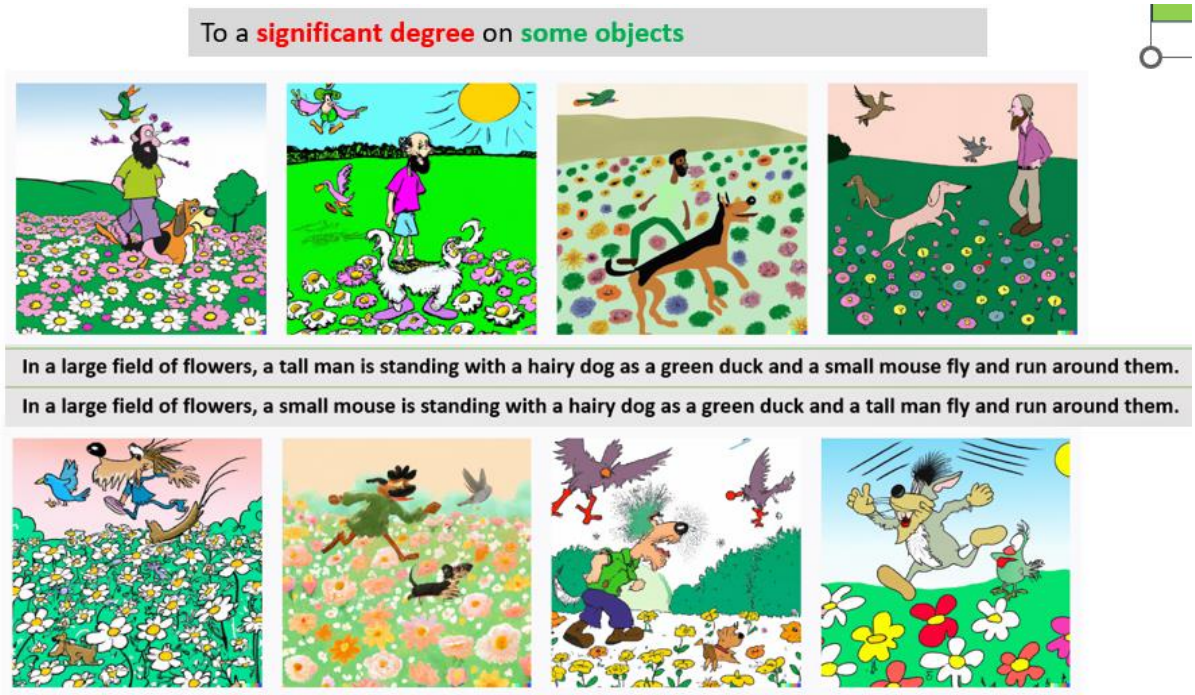
set. Elements such as the dog and the duck appear to some degree in all 4 versions, although a melding of images occurs in three of these. The mouse is largely ignored or combined, and the "largeness" of the field is conveyed in only half of the set. The merging of distinct elements into chimera, or hybrid elements could be because of variable binding. The animals and men vary in height; colours blur, but- significantly- the only verb in the prompt is applied to all featured elements. There is also a uniformity of style across the images which is missing in the second prompt's output.

In the second set, the proper nouns, while not instantly recognizable as their famous images, cause significant disruption to the predictability and consistency of output across each iteration within the set, in terms of genre (from photos to cartoons), element number, and probable appearance of the featured element. The chimeric merging of Arnold Schwarzenegger and Micky Mouse in the fourth image seems to show how variable binding affects DALL·E, where element distinction is ignored for the sake of association of each item within its vectors or schemas. As it links items and fills in the gaps, it fails to distinguish separate entities and their distinctive characteristics, even those as well-known, and presumably as extensive in web coverage, as those interposed in the prompt. It is also very important to bear in mind that DALL·E has a strict rule against the reproduction of public figures without their consent (OpenAI, n.d.), thus attempting to use such copyright images seems to result in a high degree of unpredictability and inconsistency in each iteration of the set and a clear lack of fidelity to the reproduction of that image, which is understandable considering the legal and ethical implications of unlicensed image production.

Fig 6.3 shows the last iterations of the prompt. It is first worth noting again, before discussing the major point of interest, that each picture in both sets is in the style of a children's book illustration or cartoon. The prompt has now reached a degree of complexity where prepositional clauses act as anaphoric subordination and an antecedent clause consecutively refers to a duck flying and a mouse running. The noun items in the antecedent clause were swapped with the head nouns in the last prompt. The focus was to see whether DALL·E could append the intended referent qualities and actions to the appropriate noun items successfully according to the *intended* relationships in the prompt, denoted syntactically. In the first of this adjacent pair, the man and the dog should be agents of the verb "stand" and are positioned at the head of the sentence, and presumably at the centre of the composition, if this is generally consistent with other prompt versions' output, allowing for some variation obviously (notably, this occurs in the second image where the dog seems to have absorbed some of the man's qualities, such as wearing shoes and being more chimera than dog). The mouse is yet again not apparent throughout this set of four- a constant in nearly all the output from this incrementally enlarging prompt. With the final prompt's output, however, when the mouse is transposed to the "head" of the main clause, DALL·E does render it but the image fidelity to the prompt is considerably worse than the previous prompt version. The mouse is anthropomorphized into a man/mouse chimera, its

referent of size is ignored, and the head verb, which by proximity to the subject is mostly applied ("standing") in other versions, is also ignored by DALL·E in all image iterations. The mice/men chimera are either in motion or running across the fields of flowers and bird-like creatures fly around as dog-like creatures run or stand with the mice/men creations. The proximity of the verb "fly" to the "tall man" also seems to have made DALL·E meld things together in bizarre ways. This interception between the syntactical closeness of collocated lexical elements within a sentence seems to be a cause of considerable unpredictable output. DALL·E does not seem to be able to bind together these items into their intended images at all accurately, as if removing such items from their planned collocate and moving them next to collocates with somewhat similar degrees of collocational attraction (e.g. "a *green* **duck** and a *tall* **man fly** and **run** around…") has caused great disruption to the fidelity of the text-to-visual transformation. This blurring together of man and mouse, and their subsequent actions, may exemplify how syntactically and referentially complex sentences create unpredictable or undesirable output, which then can be very hard for the prompt writer to rectify despite frequent revisions.

Fig. 6.3 Third comparison of DALL·E's output between an adjacent prompt pair with significant changes



2.4 Conclusion

As each prompt incrementally changed and grew in syntactic complexity or semantic understanding, it became apparent that DALL·E's output was consistent in that certain aspects are rendered with some fidelity and other prompt elements are either ignored or melded, chimera-like, with elements it may associate with the whole tableau or append through syntactic or semantic associations. In the case of the man/dog/duck/mouse/field core elements, DALL·E produced in every instantiation versions of the man and the dog and the field (with only one instance with the "Schwarzenegger/Mickey Mouse" variant where these elements were not featured at all), while the duck was a frequent, but not certain, common feature and the mouse appearing at a very low interval rate, or melded to another common element in the prompt. Other consistent features were the style of the genre, generally cartoonish, and centrality of the man, duck and dog in the foreground of the compositions. The consistency in what DALL·E produces as the prompt incrementally changed is in itself as significant as its inconsistency, which is a well-documented issue that Open AI also acknowledges "...as more objects are introduced, DALL·E is prone to confusing the associations between the objects and their colours, and the success rate decreases sharply. We also note that DALL·E is brittle with respect to rephrasing of the caption in these scenarios: alternative, semantically equivalent captions often yield no correct interpretations" (OpenAI, n.d.). This caveat is not to be ignored when it comes to using DALL·E for assessment and teaching purposes. While it is interesting to speculate on the reason why DALL·E does this, perhaps because of variable binding, or the data sets it was trained on, it is perhaps better to focus on how we can report on more and less predictable outcomes for useful testing and pedagogical purposes so that this tool can be utilised effectively for learning and assessment goals in the future. Indeed, to know that DALL·E can create, with no training at all (an approach called "Zero-Shot Reasoning"; OpenAI, n.d.), image-to-image translations/transformations with varying degrees of fidelity (depending on the prompt variation), and that this was a capability that Open AI did not "anticipate," (OpenAI, n.d.). suggests that this is an important area of research well beyond the scope of this particular investigation. What can be said with some certainty is that any practical pedagogical or testing application for this tool needs to have the understanding that DALL·E's degree of reliability is tempered by an entrenched degree of unpredictability which seems to foster creativity and to "fill in" any ambiguous or elliptical instructions given by a user. Kane's observation about reliability, therefore, should probably not be seen as strict precept of use at present with this tool. If one is to accept this compromise, then the use of text-to-visual tools may have a place in future pedagogical and assessment resources and activities. It is up to developers to decide what approach would best make this tool fit-for-purpose (OpenAI, n.d.) and more investigations should be carried out before implementation in an educational setting.

3.0 Strand 3: An empirical investigation of positional phrase realisation in DALL·E 2 visual output

3.1 Introduction

Artificial intelligence and machine learning tools are increasingly being discussed in relation to sentience and natural processing of human language (Leivada et al., 2022). Once these tools reach a certain threshold of reliability in interpreting human input, the potential applications for language education and assessment will be numerous. The text-to-image engine DALL·E 2 has been widely celebrated for its capability to process natural language prompts into images. However, recent studies have observed deficiencies in representation of relational aspects of agents and objects in prompts (e.g., Conwell & Ulman, 2022; Leivada et al., 2022; Marcus et al., 2022). A key component of the implementation of DALL·E 2 in language assessment contexts would be consistency of output. Consistency is essential in order to maintain control over test materials, preserve the validity of the test construct, and ensure that each learner is provided with equal opportunities to demonstrate their language abilities. In order to explore this empirically, the relation between prompt syntax and DALL·E 2's visual output was examined. Based on positional phrases used by L2 learners of English in the Aptis Test (British Council, 2020), prompts were tested for consistency between prompt and visual output. Through empirical investigation of a range of prompts, DALL·E 2's ability to interpret and realise compositionality was tested. In addition, the alternative AI image generators Adobe Firefly and Shutterstock were also tested. The results of the study have implications for the implementation of DALL·E 2 and similar tools in classroom and assessment contexts.

3.2 Previous Studies of DALL·E 2

The relationship between prompt syntax and visual output of DALL·E 2 have been key considerations in previous research. Typically, these studies aim to ascertain the degree to which AI image generators are able to interpret prompts in a manner similar to humans. As Conwell and Ulman (2022) explain, "Machine models that aspire to human-level perception and reasoning should reflect the ability to recognize and reason generatively about relations" (p.1). In order to investigate this, the researchers conducted an empirical examination of DALL·E 2's ability to interpret 15 physical and agentic relations contained in text prompts. Physical relations included in, on, under, and near, and agentic relations included pushing, pulling, hitting, and helping. 180 human participants were then asked to select which output images 'match' the prompt sentence from a set of options. Based on this process, the accuracy of the various relations was determined. Overall, participants reported a low degree of agreement between prompts and DALL·E 2's output, with only 22.2% reported agreement across

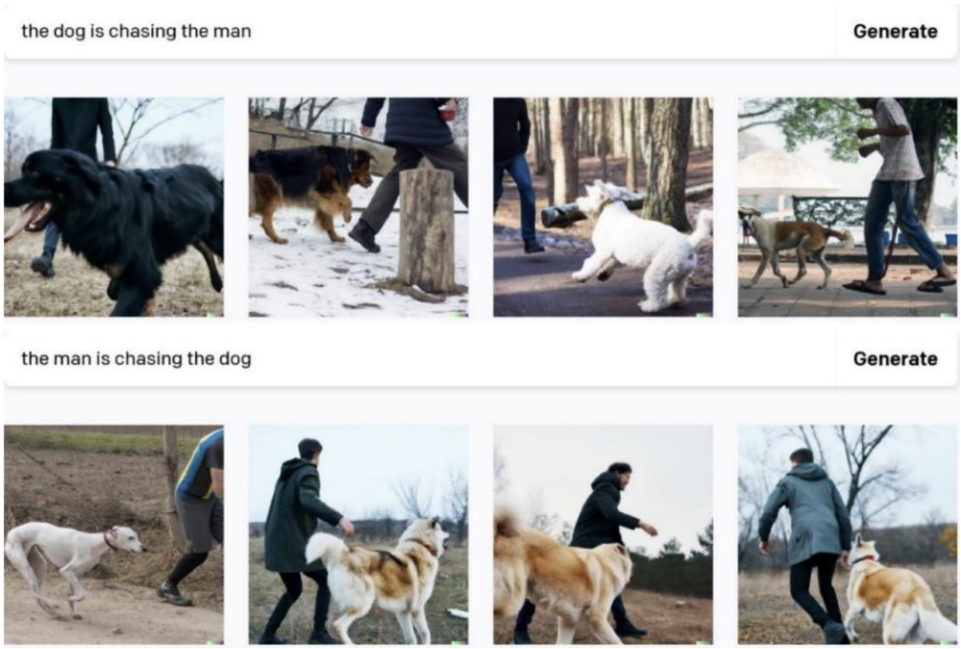
all prompts. Agentic prompts produced a higher degree of agreement with a mean of 28.4% compared to physical prompts which only generated 16.9% agreement overall. In addition, in some cases there was a high degree of variation in agreement of relations depending on the objects and agents in the prompts. For example, the prompts 'child touching a bowl' generated 87% agreement while 'a monkey touching an iguana' generated 11% agreement (Conwell & Ulman, 2022, p.5). The researcher speculated that this may be due to DALL·E 2 source training data rather than an understanding of syntax and concluded that DALL·E 2 struggled severely with realisation of relational compositionality in output images.

When referring to 'compositionality' in this context, it can be defined as the principle that the meaning of a complex expression is determined by both the *structure* and *meaning* of its constituents (Gill, 2023). For example, the phrase 'lamb eat' could have multiple meanings (the lamb is eating, the lamb is being eaten, the lamb that is eating). Once syntactic structure is added ('the lamb is eating'), the combination of structure and word meaning provide important information about the relationship between lexical components. Conwell & Ulman (2022) argued that DALL·E 2 can interpret word meaning in prompts, but it is not fully capable of interpreting relational structure based on prompt syntax. This may surprise many who have seen the tool's capabilities, yet empirical testing of specific relations has shown that the processes behind its accurate images may not be due to syntactic processing. The researchers proposed that when DALL·E 2 produced accurate relations between objects or agents, it can often be due to prior exposure to source images rather than an understanding of the relations specified in the prompt. To illustrate this, the researchers created two prompts, one with a relation which they assumed would be common in DALL·E 2's source data ("a spoon in a cup"), and one that would be very uncommon ("a cup on a spoon") (Conwell & Ulman, 2022, p.8). As shown in Figure 1.1 below, the first relation was recreated accurately by DALL·E 2 while the second relation was not. The images of a spoon in a cup are accurate only due to this spatial relation being overwhelmingly frequent in the source image data, which also explains why the output of a cup on a spoon still look very similar. This indicates a lack of true compositionality which, of course, is not determined by the semantic nature of the relation.

Figure 1.1 Illustrative example, images generated given ‘a spoon in a cup’ (left) and ‘a cup on a spoon’ (right) (Conwell & Ulman, 2022, p.8)

Issues related to compositionality were also observed by Leivada et al. (2022). The researchers found that DALL·E 2 struggled with thematic roles and the differences between agent and patient. In cases where the syntax clearly signifies the agent, the researchers found that the image generator could not reliably display this relationship in visual output. This was tested using pairs of phrases with different agents signalled syntactically, such as “the dog is chasing the man”, and “the man is chasing the dog” (p. 5). As Figure 1.2 below shows, DALL·E 2’s output was not consistent with the agentic relationship specified in the prompt, and in many cases, the relationship between the man and dog was not clear.

Figure 1.2 The performance of DALL·E 2 in thematic role reversals and word order (Leivada et al., 2022)



Additionally, the researchers found poor performance in several aspects of grammatical comprehension including passives, coordination, negation, and ambiguity (Leivada et al., 2022), and concluded that more complex grammatical structures can be problematic for DALL·E 2. Related limitations were found in DALL·E 2 by Marcus et al. (2022), and Thrush et al. (2022) found similar issues of compositionality with a range of state-of-the-art image generation models. Following up on these studies, the current project examined positional phrases specifically to determine whether DALL·E 2 was able to interpret certain types of relations more effectively, and what the causes of inconsistencies may be. Findings were discussed in the context of potential applications in language teaching and assessment.

3.3 Research Questions

The study aimed to answer the following questions:

RQ1: How does DALL.E-2 respond to the positional phrases produced by learners in the Aptis Corpus?

RQ2: To what degree does DALL·E 2 possess compositionality?

RQ3: Based on this, what are the implications for classroom and assessment implementation?

3.4 Method

Building on previous studies, the current project aimed to explore DALL·E 2's 'understanding' of spatial relations with a particular focus on learner language. The aim was to determine the effects of specific prompts used by L2 learners on the consistency of object relations in the tool's output. Input phrases were selected based on the frequency of forms employed by learners in the Aptis Corpus (British Council, 2020). The Aptis Corpus is a collection of data containing language produced by learners taking part in the British Council Aptis General Test, a computer-mediated English proficiency assessment tool for adults. The test comprises of reading, writing, listening, and speaking components, and the data used for the current study was a subsection of the spoken component with 660,000 tokens of leaner speech from CEFR levels B1 to C. Prompts were selected based on their overall frequency in this corpus in order to focus on forms commonly used by learners in classroom or assessment settings. Both overall (e.g., on the left, on the right) and relative position (e.g., above, below) prompts we examined. A full list of prompts examined in the study is shown in Tables 1.1 and 1.2 below.

Table 1.1 Prompts Referring to Overall Position of Objects

Horizontal Position Prompts (Frequency in Corpus)	
X on the left (148)	X on the right (174)
X in the left (85)	X in the right (53)
on the left side (9)	on the right side (22)
on my left (7)	on my right (7)
in the left side (5)	in the right side (4)
to the left (3)	to the right (3)
Vertical Position Prompts (Frequency in Corpus)	
at the top (43)	at the bottom (37)
on the top (40)	on the bottom (24)

Other Overall Position Prompts
in the middle (20)

Table 1.2 Prompts Referring to Relative Position of Objects

Horizontal Position Prompts (Frequency in Corpus)	
left of X (3)	right of X (3)
Vertical Position Prompts (Frequency in Corpus)	
above (100)	below (100)
underneath (3)	

A prompt template was produced containing two distinct objects. The objects selected were a cat and a panda as they are visually distinct yet could share some properties which could be merged (i.e., the objects are similar enough to be merged visually by DALL·E 2, yet distinct enough that any significant merging of visual features would be recognisable). This was done in order to further test previous findings of entities in a prompt being interpreted as modifiers of other entities in DALL·E 2's visual output (e.g., Marcus et al., 2022; Rassin et al., 2022). The style 'digital art' was chosen to maintain consistency of output and allow for easier comparisons between images. The specific prompt template used for the study was as follows: "A digital art image of a panda and a cat. The cat is *on the right* and the panda is *on the left*". The italicised phrases are the positional phrases selected from learner output in the Aptis Corpus. For one prompt, 'in the middle', a third object was added, and the following prompt was used: "A digital art image of a panda, a cat, and a rabbit. The rabbit is *in the middle*". 20 images were produced for each prompt in the first round of testing followed by a further 40 images for certain prompts in secondary analyses. Accuracy was determined based on the consistency between the prompt and visual output. Clear correspondence between the prepositional meaning of the prompt and the position of the objects in the output was required for this. If the spatial relationship between the objects in the image were not clear, or contradicted the prompt, then the output was deemed inaccurate. Examples of output which was deemed accurate and inaccurate are provided in Figures 2.1 and 2.2 below. Based on the following prompt: "A digital art image of a panda and a cat. The cat is *above* and the panda is *below*"¸ the images in Figure 1.1 were deemed accurate, while the images in Figure 1.2 were determined to be inaccurate.

Figure 2.1 Output Images Categorised as Accurate Based on ‘The cat is above’ and ‘the panda is below’

Figure 2.2 Output Images Categorised as Inaccurate Based on ‘The cat is above’ and ‘the panda is below’

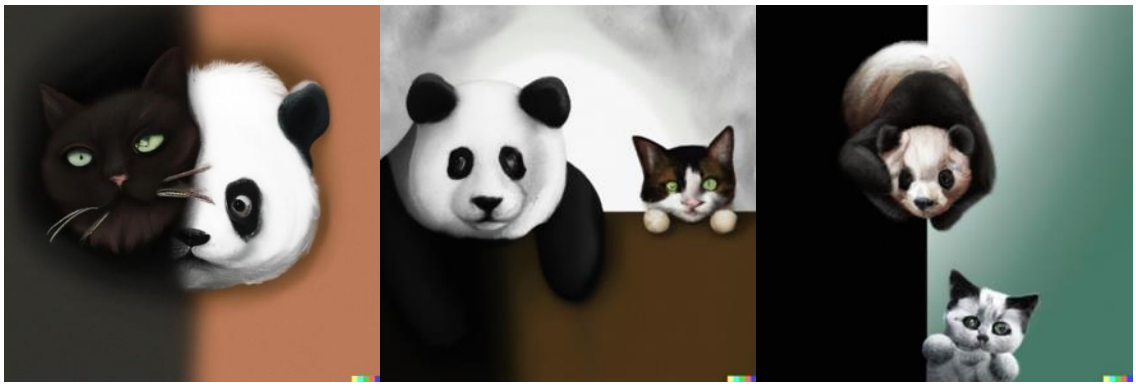


Figure 2.1 shows a clear spatial relationship between the objects in the output image while Figure 2.2 shows unclear positional relationships (left), followed by the incorrect orientation (centre), and finally the correct orientation but with the objects assigned to positions incorrectly (right). Due to the simplicity of the output image, there were no cases in which this distinction was problematic. For each image, a percentage of prompt-output consistency was calculated based on this. Prompts which produced output at 80% accuracy or higher were tested with a further 40 images with some changes in condition to explore the results further.

3.5 Results

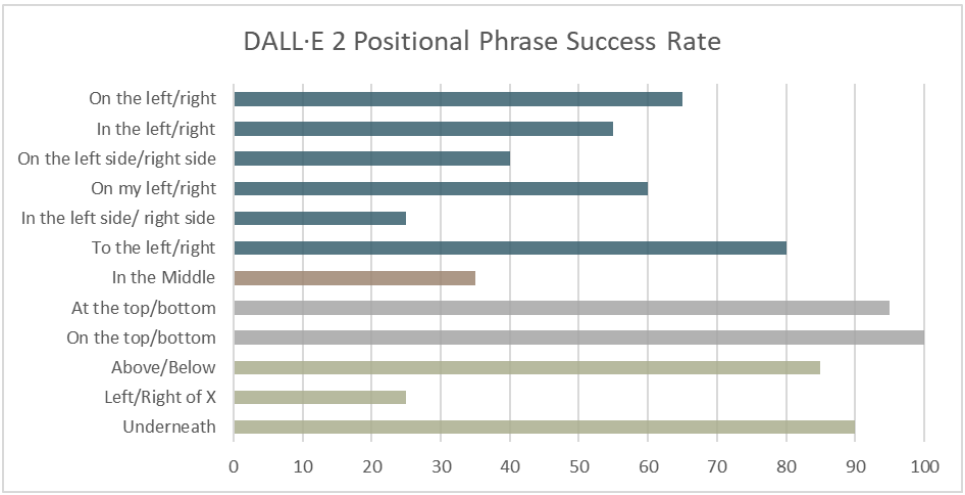
Beginning with the fundamentals of compositionality, DALL·E 2 was able to produce the correction positional orientation of objects accurately and consistently in output images. Horizontal prompts such as “The cat is on the right and the panda is on the left” resulted in the two animals positions side to side, and vertical prompts such as “The cat is above

and the panda is below" produced output with one animal positioned above the other as shown in Figure 3.1 below.

Figure 3.1 "The cat is on the right and the panda is on the left" (left image) "The cat is above and the panda is below" (right image) "

This indicates some basic understanding of spatial relations between objects. Unfortunately, DALL·E 2 was much less successful at positioning each specific object in the correct position. Figure 3.2 shows each prompt and the percentage of successful output images based on 20 images per prompt. The results are organised based on prompt type, with overall horizontal prompts in blue, followed by 'in the middle', the overall vertical prompts, and finally the three relative position prompts.

Figure 3.2 DALL·E 2 Positional Phrase Success Rate



It results show that: 1) there is wide variation between prompts, 2) certain properties of prompts appear to produce more consistent results, and 3) most prompts did not produce consistent output. Interestingly, horizontal prompts had a very low accuracy

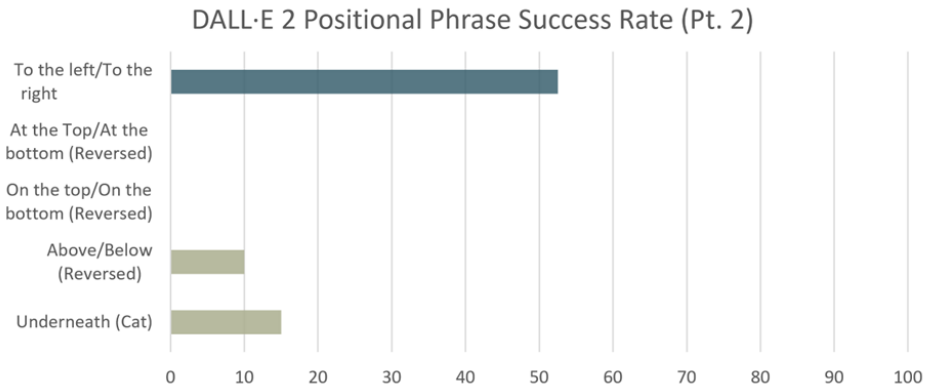
rate overall ranging from 25% to 65% (with the exception of "to the left/right" which achieved 80% accuracy). In contrast, all vertical prompts were between 80% and 100% accurate. This phenomenon had not been identified in previous studies, nor was it clear why vertical prompts would be more accurate than horizontal forms. In order to test these prompts further, another 40 images were generated for each prompt which achieved 80% or greater accuracy (see Table 1.1.).

Table 1.1 Prompts which generated images with 80% or greater accuracy in the first round of testing

Prompt		Accuracy in first round of testing
Horizontal Position	To the left/right	80%
Vertical Position	At the top/bottom	95%
	On the top/bottom	100%
	Above/below	85%
	Underneath	90%

Based on these results the following hypotheses were made: 1) to the left/right is not distinct enough from the other horizontal prompts to warrant significantly greater accuracy so the high accuracy of this specific prompt may just be due to chance, and 2) the fact that all of the vertical prompts produced accurate images indicates that some integral aspect of these prompts resulted in more accurate visual output (i.e., this is not due to chance). As a result, "to the left/right" was test a further 40 times in its same form while the vertical prompts were changed in order to investigate the results further. In the first round of testing the cat was always positioned *above* the panda in prompts (e.g., 'the cat is at the top', 'the cat is above' etc.). For the following 40 images, this orientation was reversed for all vertical prompts so that the panda was specified as above. The results of the second round are shown in Figure 3.3 below.

Figure 3.3 DALL·E 2 Positional Phrase Success Rate: Part 2



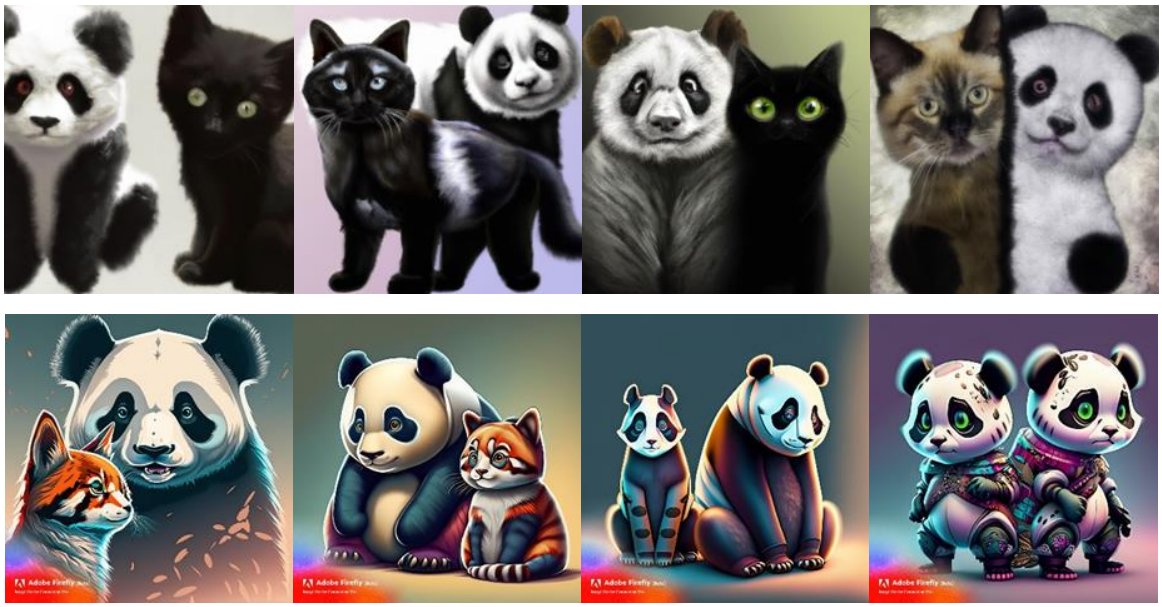
As the table shows, the horizontal prompt anomaly was in fact due to chance as the further 40 images achieved roughly 50% accuracy. The results suggest that all of the horizontal prompts resulted in images with the two objects randomly assigned to each side. The reasons for this will be discussed further in the following section. More significantly, the vertical prompts suddenly became highly inaccurate once the positioning of the objects was changed. This indicates that prompt syntax was not the determining factor for the positioning of the two entities in the output. When the prompt was changed to specify the panda at the top of the image, the output continued to position the cat above as shown in Figure 3.3 below.

Figure 3.3 DALL·E 2 output for the prompt: 'The panda is above, and the cat is below'.

To summarise, DALL·E 2 was able to recognise the difference between vertical and horizontal orientation in prompts, though it could not consistently assign the correct positions of the two entities specified in the prompt syntax. Prompts with horizontal orientation resulted in random assignment of one of the objects to the left and one to the right resulting in low accuracy for this prompt type. In contrast, while the vertical prompts initially seemed to be interpreted with very high accuracy, further testing showed that the output was simply positioning the smaller animal above the larger one in almost all cases regardless of prompt syntax. This suggests that there are other factors which can override prompt syntax when orienting multiple objects in the tool's output. In addition, horizontal prompts were also tested on other image generation tools

including Shutterstock and Adobe Firefly. These tools appeared to suffer from the same compositionality issues as DALL·E 2 as shown in Figure 3.4 below. Both tools, particularly Adobe Firefly, also showed evidence of concept merging with creatures possessing features of both cats and pandas. The results indicate that limitations in terms of syntax processing are pervasive across AI image generation tools.

Figure 3.4 Shutterstock (above) and Adobe Firefly (below) inconsistent output for the prompt: 'The cat is on the left and the panda is on the right.



3.5 Discussion

The current study supports previous findings indicating that DALL·E 2 is not currently capable of producing consistent output based on the spatial relations contained in typical prompts produced by language learners. The tool is undeniably impressive in terms of image generation and is somewhat reliable in some aspects of prompt interpretation. For example, the current study found that DALL·E 2 could fairly reliably produce the correct number of objects in the images tested, an observation also made by Marcus et al. (2022). However, the tool becomes less reliable as the number of objects increases with subsequent non-empirical tests showing inconsistencies with numbers of five and above. In addition, while DALL·E 2 was able to recognise the nature of spatial relations in many cases, it could not interpret specific positions of entities based on the prompt syntax. Rassin (2022) suggests that due to insufficient grammatical parsing, one word can be interpreted as 1) a modifier of two different entities or 2) an entity and a modifier of another entity in the prompt. Examples of this are shown in Figure 4.1 below. In the image on the left, the modifier 'gold' pre-modifies the ingot, but

DALL·E 2 applied the modifier to both the ingot and the gold-coloured goldfish. The image on the right shows that even entities within the prompt can be interpreted as modifiers of other entities as shown by the image of the zebra walking on the zebra crossing.

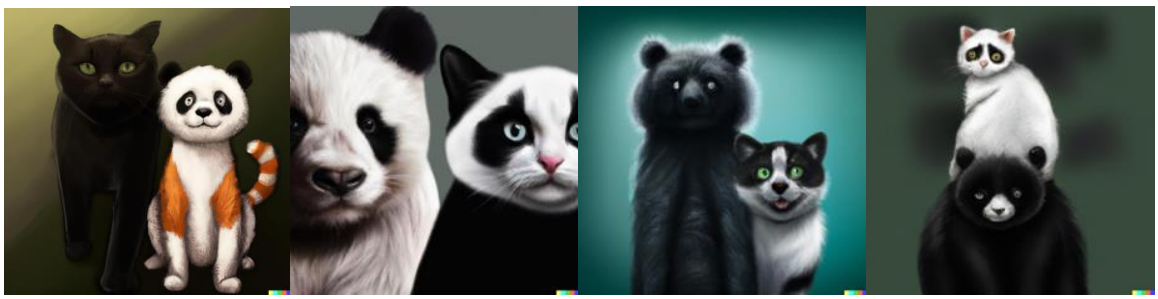
Figure 4.1 “A single word realized as a modifier of multiple entities” (left), “A single word is realized as an entity and as a modifier of another entity” (right) (Rassin et al., 2022, p.3)

Main (left): *a **fish** and a **gold ingot**.*
Control (right): *a fish and an ingot.*

Main (left): *A **zebra** and a **street**.*
Control (right): *A zebra and a gravel street.*

There is also evidence of one entity being realised as a modifier of another entity in the current study as well. As shown in Figure 4.2 below, panda-like cats and cat-like pandas were not uncommon in DALL·E 2’s visual output. It thus seems likely that the modifiers ‘left’ and ‘right’ may also be realised as modifiers of both entities in the current study which would explain the observed iznconsistency in relative position of objects based on horizontal prompts.

Figure 4.2 One entity being realised as a modifier of another in the current study



In terms of the vertical prompts, the results suggest that another factor overrode the random assignment of object positions resulting in one predominant orientation of entities. Regardless of prompt syntax, DALL·E 2 produced output with the cat positioned above the panda in over 90% of vertical prompts. One of the key advantages of AI systems is their access to vast amounts of data which can be used to inform output. This is a large component of what makes DALL·E 2 so effective, but it is also limiting in certain

contexts. When determining the desired output of the user, the system has multiple components to reconcile, the objects and agents in the prompt, the relations between these as determined by prompt syntax, and the vast amount of source data which matches the components of the prompt. This process is impressively effective when prompts involve merging different semantic concepts as shown by the example images displayed on the DALL·E 2 home page such as "An oil painting by Matisse of a humanoid robot playing chess" (OpenAI, 2023). However, the current study, supporting Conwell and Ulman (2022), has shown that in cases where the prompt syntax contradicts relations in source data, DALL·E 2 overwhelmingly leans towards source data rather than syntax. Conwell and Ulman (2022) found that DALL·E 2 could consistently produce accurate images of 'a spoon in a cup' but was not able to do so for the prompt 'a cup on a spoon' (see, Figure 1.1.). In most cases, DALL·E 2 produced an image with a spoon *in* a cup, which the researchers attributed to this common relation in the source image data. In this study, the smaller cat was almost always positioned above the larger panda. It seems likely that images of cats sitting atop large objects would be much more common than pandas sitting atop small objects. It appears that, despite prompts specifying the panda being above, DALL·E 2 relied on prior exposure to source images to determine an 'appropriate' orientation.

The evidence presented in this study indicates that DALL·E 2 would not currently be an appropriate tool for classroom or assessment implementation. Despite the tool's strengths, its key limitation is inconsistent performance in relation to prompt syntax. This study supports previous research (e.g., Conwell & Ulman, 2022; Leivada et al., 2022; Marcus et al., 2022) which has found that DALL·E 2 is somewhat limited in realisation of compositionality in visual output. This is not necessarily a criticism of the tool, rather it is a reflection of what is has been designed to do and what is currently beyond its scope. However, from a language learning perspective, the inconsistencies in syntax processing limit the teaching and assessment contexts in which the tool can be reliably implemented. The greatest strength of DALL·E 2 is how open and expansive the capabilities of the tool are, yet this is somewhat at odds with the consistency and control required in teaching and assessment contexts. The results found in this study are thus not considered as representative of the limitations of AI language processing, but rather reflective of the purposes for which DALL·E 2 was designed. This study thus concludes that any image generation tool used for assessment or teaching should maintain principal focus on language, even at the expense of other factors such as openness of image generation, as the validity of the linguistic construct is the most important factor for helping learners.

4.0 References

British Council. (2023). *Intercultural Communicative Competence*. TeachingEnglish. https://www.teachingenglish.org.uk/professional-development/teachers/knowing-subject/d-h/intercultural-communicative-competence#:~:text=Intercultural%20communicative%20competence%2C%20or%20ICC,vary%20from%20culture%20to%20culture

Chapelle, C.A. (2013). Reliability is a "central concern" of test developers. The Encyclopaedia of Applied Linguistics, 4918-4923. Oxford: Blackwell/Wiley.

ChatGPT. (2023). In-depth discussion with J. Curcio about anthropomorphism and AI. Retrieved from https://modernmythology.net/through-a-mirror-darkly-conversing-alone-with-ai-738b9dbff17d

Conwell, C., & Ullman, T.D. (2022). Testing Relational Understanding in Text-Guided Image Generation. ArXiv, abs/2208.00005. https://doi.org/10.48550/arXiv.2208.00005

Di Carlo, G. S. (2013). Lexical Differences between American and British English: a Survey Study. *Università Degli Studi Di Napoli 'Federico II.'*

England, R. (2021). DALL·E . Retrieved from https://tinyurl.com/2hr534c7

Gil, D. (2023) Bare and Constructional Compositionality. *International Journal ofPrimatology* https://doi.org/10.1007/s10764-022-00343-6

Indonesia Foundation , B. C. (n.d.). *Differences between British and American English*. Differences between British and American English | British Council Foundation Indonesia. https://www.britishcouncilfoundation.id/en/english/articles/british-and-american-english

Irfan, S. (2021, February 4). *The differences in British and American spelling*. Oxford International English Schools. https://www.oxfordinternationalenglish.com/differences-in-british-and-american-spelling/

Kane, M. T. (1992). Reliability is a necessary condition for validity because g eneralization is a key inference. Psychological Bulletin, 112(3), 527–535.

Khalifa, H., & Weir, C. J. (2009). Examining reading: Research and practice in assessing second language reading. Cambridge: Cambridge University Press

Koh, J., et al. (2021). Generating computer-based images from text prompts. Retrieved from http://tiny.cc/j68svz

Labeller. (2022). Everything you need to know about DALL-E and more. Retrieved from https://modernmythology.net/through-a-mirror-darkly-conversing-alone-with-ai-738b9dbff17d

Leivada, E., Murphy, E., & Marcus, G. (2022). DALL-E 2 Fails to Reliably Capture Common Syntactic Processes. *ArXiv, abs/2210.12889*. https://doi.org/10.48550/arXiv.2210.12889

Marcus, G., Davis, E., & Aaronson, S. (2022). A very preliminary analysis of DALL-E 2. ArXiv, abs/2204.13807. https://doi.org/10.48550/arXiv.2204.13807

Mollick, E. R., & Mollick, L. (2022). New modes of learning enabled by AI Chatbots: Three methods and assignments. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4300783

Mollick, E. R., & Mollick, L. (2023). Using AI to implement effective teaching strategies in classrooms: Five strategies, including prompts. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4391243

Murphy, E., & Leivada, E. (2022). A model for learning strings is not a model of language. *Proceedings of the National Academy of Sciences*, *119*(23). https://doi.org/10.1073/pnas.2201651119

Obos, A. I., Susilo, S., & Ping, M. T. (2022). Exploring EFL students' awareness of the differences between American and British varieties. *Lingua*, *18*(2), 167–177. https://doi.org/10.34005/lingua.v18i2.2123

OpenAI. (n.d.). Content Policy. Retrieved from https://labs.openai.com/policies/content-policy

OpenAI. (2023). *DALL·E 2*. Retrieved July 1, 2023, from https://labs.openai.com/

Patel, M. (2023, April 18). *In our rapidly changing world what is the future of the English*. British Council. https://www.britishcouncil.org/voices-magazine/our-rapidly-changing-world-what-future-english-language

Rassin, R., Ravfogel, S., & Goldberg, Y. (2022). DALLE-2 is seeing double: Flaws in word-to-concept mapping in Text2Image models. *arXiv preprint arXiv:2210.10606*. https://doi.org/10.48550/arXiv.2210.10606

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., & Ross, C. (2022). Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. arXiv preprint arXiv:2204.03162. Retrieved from https://arxiv.org/abs/2204.03162

Yaman, I. (2015). Exploring ELT Students' Awareness of the Differences between the British and American Varieties of English. Ondokuz Mayıs University Journal of Education. 34. 10.7822/omuefd.34.1.9.