Filling the SLaTE: examining the contribution LLMs can make to Irish iCALL content generation

Neasa Ní Chiaraín, Neimhin Robinson Gunning, Oisín Nolan, Madeleine Comtois

The Phonetics and Speech Laboratory, Trinity College Dublin
neas.nichiarain@tcd.ie, nrobinso@tcd.ie, oinolan@tcd.ie, comtoism@tcd.ie

Abstract

The aim of this paper is to investigate whether a large language model (ChatGPT with GPT 3.5) can be used to generate learner prompts as part of the development of a new application, Cén Scéal?/An Scéalaí, within the Irish (Gaeilge) iCALL platform. In this work an evaluation of the model’s accuracy in generating lists of words within various POS tags is presented, which enables an interrogation of the model’s ontology of the lexicon and of morphological rules. It is found that the model hallucinates lexical items, mistranslates words, and makes mistakes with grammatical gender. The consequences of the model’s error rates in these regards are discussed with respect to its suitability for generating content for the Cén Scéal?/An Scéalaí application. The model was found to produce pedagogically unsound material, but it may still be useful in assisting the manual curation of high-quality language learning resources.

Index Terms: intelligent-Computer-Assisted Language Learning, LLM, pedagogical resources, prompt generation

1. Introduction

The potential of large language models (LLMs) as educational tools is being pursued by widely used educational platforms, such as Khan Academy¹ and Duolingo². The cost of developing and training LLMs can be prohibitive for low-resource initiatives, which makes it tempting to leverage existing models where possible. However, the quality of these models varies across domains and across languages. This study investigates the quality of Irish language output from OpenAI’s ChatGPT [1] using the GPT 3.5 backend. In the context of Irish language education, it is foreseeable that tools such as ChatGPT may become increasingly prevalent, regardless of the quality of language output they produce. Therefore, it is imperative for all individuals engaged in the teaching and learning of Irish to be mindful of the performance and accuracy of such models. It will be important to sustain efforts to communicate the potential benefits and shortcomings of successive versions of these tools to educators and learners alike.

A brief note on terminology is included here since this text makes frequent references to two different types of prompts, namely Student-Targeted Prompts (STPs) and Language Model Input Prompts (LMIP). In this text we evaluate ChatGPT by crafting LMIPs and analysing the responses, and we discuss the implications of these analyses with regard to the suitability of ChatGPT for generating STPs.

One use case for LLMs in language learning exercises is to generate STPs. The idea behind these STP generators is to find a way to inspire students to “fill their slate”. Often learners are faced with a blank slate as the starting point for embarking on their story composition journeys. To quote ChatGPT itself, when prompted ‘What is the opposite of a blank slate?’, the model replied:

“Filling the slate” is a metaphorical phrase that refers to the process of acquiring knowledge, skills, or experiences that shape a person’s understanding or abilities in a particular area. The phrase can be used in various contexts, such as education, personal development or professional growth. For example, in education, teachers often aim to fill the slate of their students by imparting knowledge and fostering learning. Similarly, in personal growth, individuals may engage in activities to acquire new skills to fill their slate and expand their understanding of the world. In essence, filling the slate suggests actively seeking and acquiring knowledge or experiences to enrich one’s understanding of capabilities in a particular domain.”

This definition corresponds with the motivation behind the Cén Scéal?/An Scéalaí application. Cén Scéal?/An Scéalaí is a literacy aid embedded in the An Scéalaí³ iCALL platform, which encourages creative composition by stimulating the imaginations of learners. According to Smith [2] in his work on writing strategies, the approach described can be categorised as a ‘language-based strategy’ that serves to motivate learners in creating their own written compositions. One of the STP generators in Cén Scéal? facilitates users in creating their own prompts. The user selects a POS, and a random example of the POS is provided. The quality of this STP generator is determined by the underlying corpus of words/phrases and their POS tags. Therefore, in this paper we investigate the potential of using ChatGPT to generate data that can be used for this STP generator.

2. Background

2.1. An Scéalaí

An Scéalaí is an intelligent-Computer-Assisted Language Learning platform for Irish and a subproject of the ABAIR.ie initiative [3]. The platform adopts a holisitic approach to language learning by integrating speech technology (synthetic voices, speech recognition) with language technology (spell checkers, grammar checkers, dictionaries and thesauri) to support the development of writing, reading, spoken, and aural...
skills. The platform is designed to be suitable for both a classroom setting as well as for self-directed learners. Teachers can use An Scéalait for giving written or oral feedback to students on their compositions. This promotes self-correction for learners as they use automatic feedback from the speech and language technologies, as well as feedback from their teachers.

5,000+ users (students and teachers) have registered with the platform and 42,542 written stories have been recorded to date. The platform was primarily piloted with students participating in online Gaeltacht immersion courses during the 2021/22 period. While the main focus of An Scéalait is to enhance students’ general language and composition skills by harnessing speech and language technologies, the platform serves as an overarching framework for the development of content-rich applications, including An Bat Mírialta and Cén Scéal?.

2.2. Cén Scéal?

Cén Scéal? is a prompt-generating story-starter application within An Scéalait that provides writing prompts to encourage language learners’ productive skills by means of composing their own stories. Its title can be translated to English in a few ways, such as ‘What’s the story?’ (common greeting), ‘What which story?’, or, if translated literally from English, ‘What a [great] story!’. A bank of suitable stimuli (STPs) underpins the application to encourage learners to embark upon creative story creation in either a written or oral format. Based on feedback from previous An Scéalait pilot tests [3], it is clear that it can be difficult for learners to start a composition themselves without a teacher outlining a particular task for them. This application is therefore intended to stimulate self-directed learners to create new materials without being continuously dependent on a teacher to get them started.

Cén Scéal? offers six different types of STP generators, which are available to try on the An Scéalait website. Each generator encompasses an extensive assortment of carefully curated prompts that learners can choose from to initiate their stories. These prompts offer a range of options, including paragraphs, sentences, or individual words, allowing learners to select the starting point that best suits their creative needs. These prompts were created to align with the curriculum, requiring substantial effort to ensure their appropriateness for learners at different language proficiency levels. The learner can accept and select the STP on display to begin their story on the An Scéalait composition dashboard, or they can reject the STP and ask the system to display another at random. For the current study, we focus our attention on just one of the six available STP generators, namely the POS-based generator [3].

The Part-of-Speech Prompt Generator serves multiple functions, including (i) introducing grammatical terms, filling a potential gap in the limited emphasis on teaching these terms in Irish classrooms; (ii) expanding vocabulary, exposing learners to a wide range of vocabulary as they explore the lexemes associated with each POS. This exposure enhances their word knowledge and facilitates the expansion of their vocabulary repertoire; (iii) fostering active sentence construction by organising lexemes into meaningful sentences - after selecting the specific lexemes they wish to include in their sentences (e.g., ‘pick me a noun/adjective/etc.’) learners are prompted to actively engage in constructing sentences by organising them into a coherent and meaningful sentence structure. This activity promotes creativity, critical thinking, and language production skills by providing valuable practice in sentence formation, syntax, and grammatical accuracy. The Part-of-Speech Prompt Generator is a versatile tool that promotes interactive and engaging language learning experiences. A screenshot of this POS-based generator in use is provided in Figure 1.

As mentioned, currently all data behind these STP generators is crafted by language content experts, and we wish investigate whether or not LLMs are sufficiently reliable to automate this task within Cén Scéal?. The use of LLMs has the potential to vastly increase the amount of content generated and at a far greater speed than would ever be possible by means of manual composition. This paper therefore investigates the use of ChatGPT for generating this type of content and evaluates its suitability to be used by teachers and learners of Irish.

2.3. Using LLMs as tutors

Pardos and Bhandari [5] evaluated ChatGPT’s effectiveness as an algebra tutor, specifically examining its performance in generating hints for algebra problems. They found that ChatGPT performed worse than human tutors and did not yield statistically significant learning gains for students in either elementary or intermediate algebra. Language tutoring may be a domain in which LLMs can be more effectively employed. This study aims to provide initial insights into its potential in the case of Irish, which is less well resourced than ChatGPT’s apparently dominant languages.

3. Methods

The method used to gain insight into the quality of the model’s Irish language output was as follows. The model was prompted, via the ChatGPT interface, to list 50 words from a particular POS together with an English translation. The parts-of-speech examined were nouns, adjectives, verbs, prepositions, and pronouns. We experimented with three kinds of Language Model Input Prompts (LMIP):

- **Naïve**: The naïve LMIPs are simple (non-engineered) requests in English asking ChatGPT to generate a list of 50 Irish words for the given POS, for example, “List 50 prepositions in the Irish language and provide a translation.” For

![Figure 1: A screenshot from the POS-based prompt generator in Cén Scéal?](https://scealai.abair.ie/#/prompts/pos)
some of the parts-of-speech, additional instructions were provided to specify a certain form of the word such that it would be suitable for the pedagogical task. For example, specifying that an article should be included for each noun.

- **Gaeilge**: These LMIPs are simply Irish translations of the naïve LMIPs. It was hypothesised that prompting through Irish would result in higher quality output.
- **Engineered**: These LMIPs are the naïve prompts prefixed by the phrase “You are an Irish language expert.” It was hypothesised that explicitly instructing the model to behave as though it were an expert in Irish would result in higher quality output - therefore mitigating degradation in output quality caused by factors such as the large proportion of Irish data online being learner-generated.

Combining each kind of LMIP with each POS category resulted in a total of $3 \times 5 = 15$ LMIP prompts, and each prompt was repeated 3 times resulting 50 words resulting in $3 \times 15 \times 50 = 2250$ manually evaluated words. Due to the stochastic nature of the responses generated by ChatGPT for a given prompt, three responses were generated for each LMIP in an attempt to enhance the robustness of the empirical results presented. Once the word lists had been generated, each list was manually evaluated in order to measure the quality of the responses. The evaluation procedure involved specifying a set of criteria for each POS. Each word in the response must satisfy all these criteria in order to be considered acceptable. The acceptability criteria for each POS were defined as follows:

- **Nouns** must be in the nominative form, singular or plural, with correct mutations as determined by gender, and include the correct definite article.
- **Verbs** must be in the dictionary form, i.e. not conjugated.
- **Prepositions** must be any valid preposition.
- **Adjectives** must be any valid adjective.
- **Pronouns** must be any valid pronoun, including possessive pronouns, e.g., “lions (with me),” constructed using the preposition “le”. Phrases containing multiple pronouns, e.g. “mise agus tu” (me and you), are not allowed.

A further general criterion that applied to all output was that the English translation accompanying the word must be correct. Repetition in a generated list (within a single chat session) was also measured, although not counted as an error, provided the repeated word was correct. Each item in each list was labelled ‘correct’ or ‘incorrect’ according to the above criteria, yielding accuracy scores for each list. The accuracy scores for the three lists generated using each LMIP were then used to compute a mean and standard deviation of percentage accuracy for that prompt. These results are presented and discussed in Section 4, along with qualitative analysis of the responses.

### 4. Results & Discussion

The accuracy scores resulting from the experiments described in Section 3 are reported in Table 1. A visual representation is available in Figure 3, page 6.

ChatGPT made several kinds of errors performing the task of listing words, summarised in the list below, and discussed further in the paragraphs that follow:

- **Noun gender**: Irish language nouns have grammatical gender of either masculine and feminine, and in the nominative case singular feminine nouns are mutated with a lenition, (“bean”→“an bean”), whereas masculine nouns in the nominative singular take no lenition (“fear”→“an fear”).
- **Noun number**: Another type of error was the incorrect pluralisation of nouns with the chosen article (e.g. “an fhiacla” instead of “an fhiaclaí” (the tooth; plural: “na fiaclaí”).
- **Inconsistency**: Sometimes ChatGPT would incorrectly add or omit a lenition (e.g. “an ghealach” rather than “an ghealach”). Interestingly, ChatGPT would sometimes use a word correctly, but then in another chat session would use the same word incorrectly. When prompted specifically about the word “gealach” in twenty separate chats, it used the word correctly on 10 occasions, hallucinated on 4 occasions, and treated it with the wrong grammatical gender on 6 occasions.
- **Hallucination**: There were also hallucination errors, where ChatGPT would propose entirely non-existent words (e.g. “an t-aireacht”). ChatGPT also occasionally proposed nonsensical compound nouns with poor attempted translations (e.g. “an teilleishean”, lit. “the television head”, for “television program”).
- **Incorrect POS**: There were cases where ChatGPT proposed a word with the wrong POS, e.g. listing “síle” as a verb when it is in fact a noun. It is clear that performance varies depending on the POS in question (see Table 1). We observe that the accuracy is higher for parts-of-speech for which there are more examples, i.e. there are fewer prepositions than other parts-of-speech; it is a closed class. Furthermore, we observe that parts-of-speech which are subject to additional morphological rules, e.g. lenition due to gender in nouns, are more error prone. In examining the generated word lists more closely, we see that the preposition lists start off well in each generation, but then tend to degrade in accuracy once the most common options have been exhausted. In one case the model resorted to phrasal verbs that contain prepositions, e.g. “ag síle le” (“looking forward to”). In another case it started generating adverbs, which in Irish often begin with the preposition “go”, e.g. “go brónach” (“sadly”). Nouns were mostly correct but frequently mis-applied the gender rules, e.g. including lenition where it shouldn’t have, or missing lenition where it should have been included.

More generally, we observe that the number of errors is correlated with the word’s position in the generated list, with words towards the end of the list more likely to contain errors than words at the beginning. This is in part due to the fact that the task of selecting new unique words becomes inherently harder as the set of options is exhausted. This phenomenon is plotted in Figure 2.

The degradation of performance for the Irish language prompt (row 2 ‘Gaeilge’ in Table 1) is in fact not so surprising when we consider that ChatGPT is in general a poorer model of Irish than it is of English. Some of the mistakes generated by ChatGPT may be explained by an anglophone bias. For instance, when asked for pronouns it generated “iomlán” with the English translation “all”. The word “all” can indeed be used as a pronoun in English, and “iomlán” is indeed a translation of the word “all”, but crucially “iomlán” is not a translation of “all” in its pronoun sense. This is an indication that ChatGPT’s approach to the task was English-first (think of an English pronoun and translate it to Irish, rather than vice versa). Some cases of incorrect POS assignment further evidence anglophone bias, for example the assignment of verb rather than noun to the word ‘síle’, where the English verb “hope” is often translated into Irish as “Tá síle agam” (lit. “I have hope”). The issue of Anglicisms amongst learners of Irish is already a challenge in Irish education, so the fact that ChatGPT likely reflects this issue rather than mitigating against it is another reason to be
### Table 1: Mean and standard deviation for accuracy of generated parts-of-speech. Generated items were marked in a binary fashion as either correct or incorrect regardless of error type.

<table>
<thead>
<tr>
<th>Language</th>
<th>Noun</th>
<th>Verb</th>
<th>Preposition</th>
<th>Adjective</th>
<th>Pronoun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>80%  (8%)</td>
<td>86.67% (2.31%)</td>
<td>70% (10.39%)</td>
<td>89.33% (3.06%)</td>
<td>53.33% (23.18%)</td>
</tr>
<tr>
<td>Gaeilge</td>
<td>52%  (15.62%)</td>
<td>76% (12.17%)</td>
<td>32.67% (18.15%)</td>
<td>89.33% (4.62%)</td>
<td>15.33% (2.31%)</td>
</tr>
<tr>
<td>Engineered</td>
<td>78% (3.46%)</td>
<td>84% (3.46%)</td>
<td>81.33% (9.02%)</td>
<td>86% (2%)</td>
<td>88% (9.17%)</td>
</tr>
</tbody>
</table>

Figure 2: The probability of an error increases as a function of the position/index of the item in the generated list. For each list position, $i$, there are $3 \times 3 \times 5 = 45$ examples, and the probability $p(\text{ERROR}|i)$ is calculated as the arithmetic mean of the 45 (manual) binary error classifications.

5. Conclusions

This paper assesses the performance of ChatGPT in generating accurate Irish (Gaeilge) word lists of a given POS. The results indicate that ChatGPT exhibits a limited understanding of Irish language grammar and morphology, rendering the current model unsuitable for automatic pedagogical content-generation purposes. While the output items are more often correct than incorrect, we find that the model is not reliable. With supervision and careful post-editing by language experts, however, the model could serve as a useful assistant in manually curating word lists. Overall, the application of ChatGPT in low-resource language learning, such as Irish, demands caution and diligent oversight.

5.1. Future Work

The GPT language models (as well as others) generate tokens which are roughly but not strictly aligned to words. The current experiment focuses on the task required for an existing application, Cé {acute{e}}n Sc{acute{e}}al? and therefore no attempt was made to analyse the results at the token level, but such an analysis may give hints as to why ChatGPT makes the errors it does. For example, the mistranslation of “oideachas” (education) to “idea” may be related to the fact that “idea” is a substring of “oideachas”.

It was hypothesised that ChatGPT’s errors may reflect common errors found in its training set, which in turn would reflect common errors of L2 Irish learners. The analysis presented here could be compared to an analysis of a learner corpus, e.g. [6], to assess the this hypothesis.

The methodology presented in this work involved manually correcting the output of the model, but it would be possible to automate the correction, for instance by leaning on the wealth of grammatical knowledge encoded in the BuNaMo database [7]. By automating this task we could quickly provide evaluations of successive models as they are made available.

6. Acknowledgements

The authors are grateful for the support of An Chomhairle um Oideachas Gaeltachta agus Gaelscolaíochta (Cé {acute{e}}n Sc{acute{e}}al? project) and An Roinn Turasóireachta, Cultúir, Éalaíon, Gaeltachta, Spóirt agus Meán (ABAIR initiative). We would also like to thank Fionn Camacho Lenihan, a student of Computer Science, Linguistics and a Language (CSLL), TCD, who built the initial Cé {acute{e}}n Sc{acute{e}}al? prototype.

---

<table>
<thead>
<tr>
<th>Noun</th>
<th>Verb</th>
<th>Preposition</th>
<th>Adjective</th>
<th>Pronoun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>80%  (8%)</td>
<td>86.67% (2.31%)</td>
<td>70% (10.39%)</td>
<td>89.33% (3.06%)</td>
</tr>
<tr>
<td>Gaeilge</td>
<td>52%  (15.62%)</td>
<td>76% (12.17%)</td>
<td>32.67% (18.15%)</td>
<td>89.33% (4.62%)</td>
</tr>
<tr>
<td>Engineered</td>
<td>78% (3.46%)</td>
<td>84% (3.46%)</td>
<td>81.33% (9.02%)</td>
<td>86% (2%)</td>
</tr>
</tbody>
</table>
7. References


8. Appendix

In Section 4 we mentioned that ChatGPT would sometimes repeat the same lexical item within a single response. We decided not to count this as an error for the results presented in Table 1, but we counted the number of repetitions manually (accounting for slight differences in the realisation of items) and present in Table 2 a breakdown of the average percentage repetition for each of the LMIP/syntax-category pairs. The data in Table 1 are visualised in Figure 3.
Table 2: Mean and standard deviation for items which were repetitions of items from the same prompt. Equality of items was judged somewhat permissively when the lexical root of two items was obviously the same, e.g. “an bhfuinneog” was judged to be a repetition of “an fluinneog”.

<table>
<thead>
<tr>
<th>POS</th>
<th>Naïve</th>
<th>Gaeilge</th>
<th>Engineered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun</td>
<td>4.67% (5.03%)</td>
<td>1.33% (1.15%)</td>
<td>4% (4%)</td>
</tr>
<tr>
<td>Verb</td>
<td>8.67% (8.08%)</td>
<td>4.67% (2.31%)</td>
<td>6.67% (8.08%)</td>
</tr>
<tr>
<td>Preposition</td>
<td>0% (0%)</td>
<td>4.67% (5.03%)</td>
<td>1.33% (1.15%)</td>
</tr>
<tr>
<td>Adjective</td>
<td>2% (2%)</td>
<td>5.33% (6.11%)</td>
<td>3.33% (3.06%)</td>
</tr>
<tr>
<td>Pronoun</td>
<td>1.33% (2.31%)</td>
<td>0% (0%)</td>
<td>7.33% (6.11%)</td>
</tr>
</tbody>
</table>

Figure 3: Bar-charts indicating mean accuracy and standard deviation per POS tag.