Building a dual dataset of text- and image-grounded conversations and summarisation in Gàidhlig (Scottish Gaelic)

David M. Howcroft[*] **Will Lamb**[†] **Anna Groundwater**[‡] **Dimitra Gkatzia**[*]
[*]Edinburgh Napier University {D.Howcroft,D.Gkatzia}@napier.ac.uk
[†]University of Edinburgh W.Lamb@ed.ac.uk
[‡]National Museum of Scotland a.groundwater@nms.ac.uk

Abstract

Gàidhlig (Scottish Gaelic; gd) is spoken by about 57k people in Scotland,[1] but remains an under-resourced language with respect to natural language processing in general and natural language generation (NLG) in particular. To address this gap, we developed the first datasets for Scottish Gaelic NLG, collecting both conversational and summarisation data in a single setting. Our task setup involves *dialogues* between a pair of proficient speakers discussing museum exhibits, grounding the conversation in images and texts. Then, each interlocutor summarises the dialogue resulting in a secondary *dialogue summarisation* dataset. This paper presents the dialogue and summarisation corpora, as well as the software used for data collection. The dialogue dataset consists of 43 conversations (13.7k words) and 61 summaries (2.0k words).[2]

1 Introduction

The preservation of minority languages and the development of Natural Language Processing (NLP) systems in low-resource settings have gained increasing attention in recent years (Howcroft and Gkatzia, 2022; Castro Ferreira et al., 2020; Zhao et al., 2022), fueled by efforts to safeguard linguistic diversity and cultural heritage (Bella et al., 2020) as well as efforts to create inclusive and fairer systems (Nee et al., 2021; Joshi et al., 2020). One such minority language is Scottish Gaelic, which despite being a recognised national language under the European Charter for Regional or Minority Languages and a rise in Gaelic-medium education, faces challenges in terms of linguistic resources for the development of natural language generation (NLG) systems. To bridge this gap, we present a novel dialogue and dialogue summarization corpus for Scottish Gaelic, laying the foundation for further developments in NLG in this language.

To construct the dataset, we adopted a task setup centered around dialogues between proficient speakers, engaging them in conversations about museum exhibits. By grounding the conversations in images and texts, we aimed to create a contextually rich conversational dataset. Subsequently, each participant summarised the dialogue, resulting in a secondary dialogue summarisation dataset.

The primary contributions of this work are twofold: the dual-corpus, comprising the dialogue and dialogue summarisation datasets; and the software employed for data collection. These contributions are pivotal in advancing research in NLG for Scottish Gaelic and hold significant potential for future developments in the field. Moreover, this paper sheds light on the challenges and complexities encountered when gathering high-quality dialogue datasets involving native speakers of low-resource languages. By addressing these challenges and presenting a robust corpus and data collection methodology, this work enables further progress in low-resource NLG, within and beyond Scottish Gaelic.

2 Corpus Collection

We ground our data collection in multi-modal sources about exhibits found at the National Museum of Scotland. Participants discuss each exhibit through a chat interface based on slurk (Götze et al., 2022).

2.1 Task Description

We adopt a conversational question-answering task grounded in short texts combined with images. Pairs of participants engage in conversation about museum exhibits, with each acting as a museum visitor (the QUESTIONER) or a museum guide (the RESPONDER). Both participants see the same heading and image for the exhibit, but in addition to

[1] https://www.scotlandscensus.gov.uk/census-results/at-a-glance/languages/

[2] The datasets, along with code for the interface, are available at https://github.com/NapierNLP/sgge.

this, the QUESTIONER sees also a list of keywords and definitions relating to the exhibit while the RESPONDER has access to a text (mean length: 405 words, std. dev. 55) providing more information about the exhibit. These textual grounding materials are provided in Gàidhlig to avoid additional influence from English and to avoid breaking immersion in the task. Participants always have access to a short version of the experiment instructions for their current role. Each conversation focuses on a single exhibit, and after each conversation, each participant summarises the discussion. For a more detailed discussion of the task and the motivation behind it, refer to (Chandu et al., 2023).

2.2 Interface

We extend the `slurk` server (Götze et al., 2022) as the basis for our experimental interface, depicted in Figure 1. To begin, participants are assigned a role, a login token and receive a 'username' for the duration of the experiment session. This username includes an integer assigning them to a particular list of exhibits to be discussed, ensuring that each participant in a pair (with matching integers in their usernames) sees the same exhibit at the same time as their partner.

In addition to the interface itself, we made several modifications to the underlying server. To facilitate participants reconnecting when poor internet connectivity interrupted their session, we added support for returning users. We also modified the code for handling user commands (e.g. `/tòiseachadh` to begin the experiment) to account for whitespace errors.

2.3 Participant Recruitment

We set participant compensation at ∼£15/hour, with each experimental session booked as a 2-hour timeslot. Recruitment posts mentioned this rate (£30 for participation in a 2-hour study).

Our initial recruitment took place through social media (Twitter, Facebook, & Discord) and a blog post on the NLP Research Group page for Edinburgh Napier University. Potential participants completed a short (3-question) comprehension quiz based on a passage about a museum exhibit as part of the pre-screening process. Those with 100% accuracy were immediately invited to complete the consent form and scheduling poll, and those with a single error were contacted over email to assess whether the error was inadvertent or actual. The initial wave of recruitment over the course of 3 months resulted in 43 people completing the screening quiz, of whom 40 were invited to join the study. Scheduling via pre-designated timeslots[3] proved to be too challenging given the relative scarcity of participants, so we switched to a general availability model. This manual process did require more correspondence compared to the pre-designated timeslots; however, this extra workload for the research team greatly increased the ease of scheduling.

With this change, we began the second three-month phase of recruitment, adding Mastodon to our set of recruitment channels. In the final month of this phase, we also published a notice on Facebook via the Edinburgh Napier University page and paid 100 GBP to promote the post throughout the Scottish Highlands and Islands. Our second wave of recruitment yielded another 26 potential participants, of whom 19 were invited to join the study.

Once scheduled, participants received an email with the full instructions for the experiment along with a copy of the articles for the exhibits for which they would play the RESPONDER role. This way each participant could be familiar with the exhibits about which they would answer questions, making it easier for them to know where to find the answers they needed. The instruction email also provided two links, one for each half of the experiment session, so that the participants could switch roles after completing a number of dialogues. Each pair of participants was assigned 12 exhibits in total, 6 in each role, although depending on the session participants might not get to every exhibit.

Our study received standard institutional ethical and data management oversight.

Challenges with Recruitment We include a detailed breakdown of the attrition rate during recruitment to highlight the biggest challenge we encountered in our work: reaching the relatively small number of speakers of Scottish Gaelic. From the 59 participants invited to join the study, 42 completed the consent form and 19 were successfully paired and scheduled. This problem is pronounced despite the language being spoken in a highly developed country with reasonably good internet connectivity across the region. Researchers working with limited groups of speakers will need to spend considerably more time, effort, and funds on recruitment compared to what they may be used to

[3] fixed times and dates in a Doodle poll

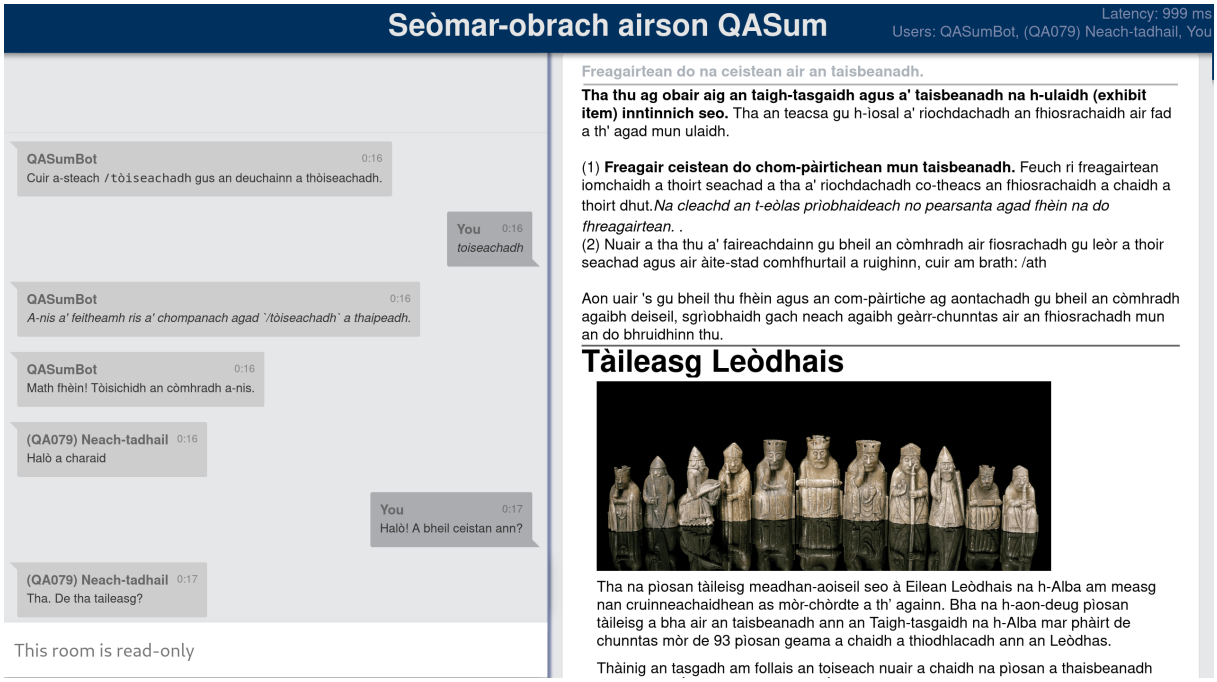

Figure 1: Our task interface. The left side of the screen contains a chat interface. Participants enter their messages to send in the text field at the bottom left. On the top right, a summary of the instructions for the user's role is displayed. Below this, the exhibit label and image are followed by the textual grounding material. This screenshot shows the RESPONDER view with enlarged text for legibility.

Q: Dè th' ann an tàileasg?	**Q**: What is chess?
R: 'S e geama a th' ann. Bidh dà chluicheadair a' gluasad phìosan timcheall air bòrd a' feuchainn ri 'rìgh' an neach eile a ghlacadh.	**R**: It's a game. Two players move pieces around a board trying to capture the other's 'king'.
Q: Cò na daoine a chruthaich na pìosan Thàileisg Leòdhais?	**Q**: Who were the creators of the Lewis Chess pieces?
R: Thathas den bheachd gun deach an cruathachadh ann an Nirribhidh. Tha iad air an dèanamh le ìbhri each-mara a thàinig à Graonlainn.	**R**: It is believed that they were hardened in Norway. They are made from walrus ivory that came from Greenland.
Q: Tha seo inntinneach. Carson a lorgar rudeigin air a dhèanamh ann an Nirribhidh ann an Leòdhas?	**Q**: This is interesting. Why find something made in Norway in Lewis?
R: Aig an àm bha Leòdhas, agus Innse Gall air fad, na phàirt den t-saoghal Lochlannach. Bhiodh daoine à Nirribhidh a' siubhal air ais is air adhart eadar Leòdhas is na dùthchannan Lochlannach agus a' toirt rudan mar seo leotha.	**R**: At the time Lewis, and the whole of Hebrides, was part of the Scandinavian world. People from Norway would travel back and forth between Lewis and the Scandinavian countries and bring things like this with them.
Q: Agus ciamar a chaidh na pìosan seo a lorg ann an Leòdhas? An robh iad am falach badeigin?	**Q**: And how were these pieces found in Lewis? Were they hiding somewhere?
R: Is dòcha gun deach an tìodhlachadh, gun deach an tasgadh. Chaidh an lorg faisg air Camas Ùige.	**R**: Maybe they were buried, they were deposited. They were found near Camas Uige.
Q: Gabhaibh mo leisgeul nach do dh'fhaighnich mi roimhe seo, ach dè gu dearbh a th' ann an ìbhri?	**Q**: Please excuse me for not asking before, but what exactly is ivory?
R: 'S e na fiaclan fada aig eich-mhara a th' ann. Canaidh sinn ivory ris sa Bheurla. Bidh daoine nas cleachdte ri ìbhri a gheibhear bho ailbheanan, 's dòcha!	**R**: It is the long teeth of walruses. We call it ivory in English. People will be more used to ivory from elephants, perhaps!
Q: Agus ceist mu dheireadh: cùin a chruthaich daoine na pìosan seo? Cò an linn?	**Q**: And a final question: when did people create these pieces? Whose era?
R: Chaidh an dèanamh uaireigin ron 12mh linn. Uaireigin sna meadhan aoisean.	**R**: They were made sometime before the 12th century. Sometime in the Middle Ages.
Q: Glè mhath. Taing mhòr.	**Q**: Very Good. Thanks a lot.

Table 1: Example dialogue between a QUESTIONER (Q) and RESPONDER (R) about the Lewis Chess Pieces in Scottish Gaelic, along with a translation to English.

Q (summary): 'S e geama-bùird a th' ann an Tàileasg, far am bhios dà chluicheadair a' gluasad phìosan timcheall air bòrd a' feuchainn ri 'rìgh' an neach eile a ghlacadh. Rinn cudeigin ann an Nirribhidh na pìosan Tàileisg Leòdhais, à ìbhri each-mara a thàinig a Graonlainn. Thàinig iad gu Leòdhas oir bha Leòdhas agus na h-Innse Gall air fad nam pàirt dhen 'shaoghal Lochlannach' aig an àm. Chaidh na pìosan an cruthachadh uaireigin ron 12mh linn, anns na meadhan aoisean. Chaidh na pìosan Thàileisg a lorg faisg air Camas Ùige. 'S e fiaclan fada aig eich-mhara a th' ann an ìbhri, ged a thigeadh ìbhri bhon ailbheann cuideachd.
R (summary): Bhruidhinn sinn mu fhir Thàileasg Leòdhais. Dè th' ann an Tàileasg, cò chruthaich na pìosan, carson a lorgadh rudan a chaidh a dhèanamh ann an Nirribhidh ann an Leòdhas agus ciamar a chaidh an lorg. Bhruidhinn sinn cuideachd air dè th' ann an ìbhir agus air cuin a chaidh na pìosan a chruthachadh.

Q (summary): Chess is a board game, where two players move pieces around a board trying to capture the other person's 'king'. Someone in Norway made the Lewis Chess pieces, from walrus ivory that came to Greenland. They came to Lewis because Lewis and the whole of the Hebrides were part of the 'Scandinavian world' at the time. The pieces were created sometime before the 12th century, in the Middle Ages. The pieces were found near Camas Uíge. Ivory is the long teeth of walruses, although ivory could also come from the elephant.
R (summary): We talked about the Lewis chess men. What Chess is, who created the pieces, why items made in Norway were found in Lewis and how they were found. We also discussed what ivory is and when the pieces were created.

Table 2: Summaries for the conversation in Table 1 written by the same users, along with translation to English.

with crowdsourcing-based studies on platforms like Prolific and Amazon's Mechanical Turk.

3 Corpus Analysis[4]

Demographics We recruited 19 participants, 13 of whom live in the Highlands and Islands and 4 of whom live in the Central Belt in Scotland. Nearly all report speaking (16) and hearing (17) Scottish Gaelic daily, with 15 also reading it daily and 11 writing daily. All participants used the language across all modalities at least monthly. Fourteen started learning the language before the age of 25. We had only one participant under 25 years old, with participants otherwise fairly evenly split among 25-34, 35-44, 45-54, and 55 or older.

Backing Data Statistics The corpus contains information about 12 different museum exhibits drawn from a pool of exhibits which were popular and/or relevant to Gaelic culture as recommended by a museum curator. We created summary texts in English based on these materials and translated them into Scottish Gaelic, using 1–3 images for each exhibit. QUESTIONERs received 7 keywords on average, with the text of the definitions about 170 words (std. dev. 69). The grounding texts for RESPONDERs typically contained 6 paragraphs and 385 words (std. dev. 53).

Conversation Statistics The dataset includes 43 conversations consisting of 870 turns. Table 1 shows one of the dialogues from our corpus. Conversations were 16.8 turns long on average (std. dev. 9.4) lasting about 11 min 53 seconds (std. dev. 5:31), with each turn taking an average of 51.6 seconds (std. dev. 18.4). The dataset has about 13.7k tokens (2.3k types), for an average of 16 words per turn (std. dev. 24). The QUESTIONER averaged 12 words/turn (std. dev. 24) versus 19 (std. dev. 24) for the RESPONDER, in line with roles each participant played.

Summary Statistics Across the 43 conversations, QUESTIONERs wrote 26 summaries and RESPONDERs wrote 35 summaries. The summaries contain 2.0k tokens (573 types), with an average of 33.7 words per message (std. dev. 16.8).

4 Discussion & Conclusion

Recruitment was a major obstacle to our data collection (cf. Sec. 2.3), with this kind of study being more akin to linguistic field work or a psycholinguistic lab experiment than typical crowdsourcing tasks. Our research greatly benefited from the expertise of our Gaelic specialist, who contributed invaluable cultural and linguistic knowledge.

This first dataset for Scottish Gaelic NLG is of the same order of magnitude as NLG datasets for English just a decade ago (e.g. Wen et al.'s (2015) restaurant corpus of 5k utterances), providing a solid starting point for developing prototype systems. Indeed, we are currently exploring response generation for grounded question answering, dialogue summarisation, and text summarisation in Scottish Gaelic using this dataset.

Beyond the dataset, our experimental interface enables other researchers to build multipurpose datasets combining summarisation and conversation and/or grounding in text and images, especially for low-resource languages. We look forward to ex-

[4] In addition to these summaries provided here, detailed statistics can be found in Appendix A.

panding this corpus in the future and enabling others to develop more datasets for more low-resource languages.

5 Ethical Implications & Limitations

Working on low-resource languages which have smaller communities of use raises a variety of important ethical considerations in common with other work in linguistics (cf. Rice, 2006; Eckert, 2014; D'Arcy and Bender, 2023). As part of the standard ethical considerations for research with human participants (e.g. consent, compensation, etc), we recognise both the increased importance of demographic information for identifying speakers' position within the language community and the increased risk of being de-anonymised based on exactly that information. To address these problems, we focused on broad geographic regions and age bands when collecting demographic information and did not collect additional information which was less important to our analyses (e.g. gender). This allows us to differentiate between speakers in regions where the language is spoken by larger proportions of the local populace and speakers elsewhere, as well as to see differences between younger learners and older, established speakers of the language while avoiding de-anonymising our participants. In addition to standard ethical considerations, we also gave participants the opportunity to be associated with the dataset by name, without having their name directly linked to their contributions, as a way of recognising their contributions to the development of language technologies for their language community. We also found that by working with a small number of participants who care about the language, we were able to collect high quality data.

Our dataset is representative of the way a relatively small number of speakers would discuss a dozen museum exhibits; while this is a useful starting point for developing NLG systems, we cannot claim that it is representative of conversational Scottish Gaelic more broadly.

Acknowledgements

DH and DG are supported under the EPSRC project 'NLG for low-resource domains' (EP/T024917/1). Data collection was funded by the Creative Informatics small grant 'Scottish Gaelic Generation for Exhibitions'.

References

Gábor Bella, Fiona McNeill, Rody Gorman, Caoimhin O Donnaile, Kirsty MacDonald, Yamini Chandrashekar, Abed Alhakim Freihat, and Fausto Giunchiglia. 2020. A major Wordnet for a minority language: Scottish Gaelic. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2812–2818, Marseille, France. European Language Resources Association.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.

Khyathi Raghavi Chandu, David M. Howcroft, Dimitra Gkatzia, Yi-Ling Chung, Yufang Hou, Chris Emezue, Pawan Rajpoot, and Tosin Adewumi. 2023. Lowrecorp: the low-resource nlg corpus building challenge. In *Proceedings of the 16th International Conference on Natural Language Generation*, Prague, Czech Republic and virtual meeting. Association for Computational Linguistics.

Alexandra D'Arcy and Emily Bender. 2023. Ethics in linguistics. *Annual Review of Linguistics*, 9(1):49–69.

Penelope Eckert. 2014. *Ethics in linguistic research*, page 11–26. Cambridge University Press.

Jana Götze, Maike Paetzel-Prüsmann, Wencke Liermann, Tim Diekmann, and David Schlangen. 2022. The slurk interaction server framework: Better data for better dialog models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4069–4078, Marseille, France. European Language Resources Association.

David M. Howcroft and Dimitra Gkatzia. 2022. Most NLG is low-resource: here's what we can do about it. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 336–350, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Julia Nee, Genevieve Macfarlane Smith, Alicia Sheares, and Ishita Rustagi. 2021. Advancing social justice through linguistic justice: Strategies for building equity fluent nlp technology. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO

'21, New York, NY, USA. Association for Computing Machinery.

Keren Rice. 2006. Ethical issues in linguistic fieldwork: An overview. *Journal of Academic Ethics*, 4(1-4):123–155.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.

Yingxiu Zhao, Zhiliang Tian, Huaxiu Yao, Yinhe Zheng, Dongkyu Lee, Yiping Song, Jian Sun, and Nevin Zhang. 2022. Improving meta-learning for low-resource text classification and generation via memory imitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 583–595, Dublin, Ireland. Association for Computational Linguistics.