Can True Zero-shot Methods with Large Language Models be Adopted for Sign Language Machine Translation?

Euan McGill
Universitat Pompeu Fabra
Barcelona, Spain
euan.mcgill@upf.edu

Horacio Saggion
Universitat Pompeu Fabra
Barcelona, Spain
horacio.saggion@upf.edu

1 Introduction

'Long-tail' or low resource languages are spoken by communities which are often left out of technological advancements, and therefore further endanger a given language's survival (Kornai, 2013; Joshi et al., 2020). They can be identified in typological resources such as Ethnologue (Eberhard et al., 2024) with metrics such as Language Vitality and Digital Language Support (Simons et al., 2022). The possibility of generating and translating text into these languages may enable the empowerment of these communities and enduring linguistic diversity.

The rise of data-intensive and large language model (LLM)-based language technologies for tasks like machine translation (MT), automatic speech recognition, and named entity recognition has enabled the inclusion of low-resource spoken languages in these technologies. Within MT, practical multilingual *few-shot* and *zero-shot* models have been created for nearly all of the $1,500$ languages[1] where there is text data that can be mined from the web (Bapna et al., 2022; Goyal et al., 2022; Federmann et al., 2022; Maillard et al., 2023; FitzGerald et al., 2023; Ruder et al., 2023) and also multimodal data (Bugliarello et al., 2022).

For the other $c.6,000$ languages, however, there exists either little or no digital presence. Resources may be confined to restricted dictionaries or wordlists, for example gathered in linguistic fieldwork studies.

As shown in Figure 1, Ethnologue's 159 documented Sign Languages (SLs) are all digitally low-resource. They cover the full spectrum of Language Vitality - but no SL has a Digital Language Support status higher than 'Emerging[2]'. SLs are characterised by multimodality (Bragg et al., 2019) and there is a lack of agreement on standardising textual SL data (Cormier et al., 2016; De Sisto et al., 2022), if there is textual data at all[3].

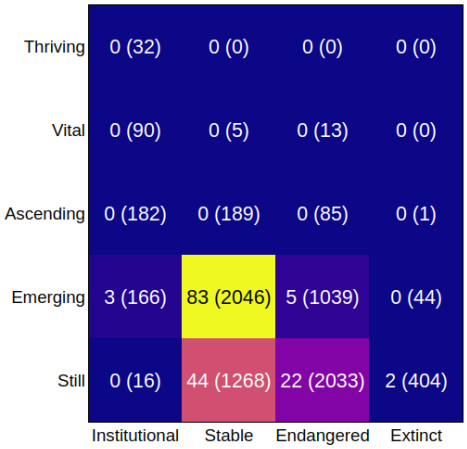

Figure 1: Labelled heatmap of the 159 SLs categorised by Language Vitality (x-axis) and Digital Language Support (y-axis). In brackets, figures for all Ethnologue languages

The unique challenge of SL data means that the methods mentioned so far may be unsuitable. Most rely on text mined from the web, while the digital resources available for SLs are usually in image or video format. In addition, other methods such as data augmentation have been attempted but have reached a performance ceiling because of the lack of parallel data available and the prospect of real, large-scale data collection efforts (De Coster et al., 2023).

[1] https://newsletter.ruder.io/p/true-zero-shot-mt provides an overview of current efforts towards true-zero shot machine translation (MT) for extremely low resource languages, and serves as the inspiration for this investigation

[2] "...some content in digital form and/or encoding tools"

[3] Moryossef (2021) characterises SLs as *extremely* low resource languages

Proceedings of the First International Workshop on Knowledge-Enhanced Machine Translation, pages 40–43
June 27, 2024 ©2024 European Association for Machine Translation

1.1 True zero shot methods

A recent work, **"Machine Translation from One Book (MTOB)"** (Tanzer et al., 2024), creates a benchmark which shows that LLMs show promise in learning sequences of a language which does not exist on the web, and is therefore completely opaque to any LLM's training data.

The authors use a *true zero-shot* approach (see also Zhang et al. (2024a) and Zhang et al. (2024b)) enabled by advances in LLMs whose prompting context window can be sufficiently long to contain book-length resources - such as a descriptive linguistic fieldwork grammar - and even multimodal data in text, audio and video (*e.g.* Gemini 1.5 Pro (Reid et al., 2024)).

It is hoped that leveraging the techniques of MTOB can be transferable to MT involving SLs (SLMT). The rest of this extended abstract describes the additional challenges foreseen by attempting this, and some methodological choices that will need to be made.

2 Resources, Challenges and Evaluation

Resources: According to repositories like Glottolog[4] (Hammarström et al., 2024), there appears to be a broad range of language grammars, dictionaries and textbooks describing numerous SLs - at least as many as for spoken languages (Zhang et al., 2024b). Resources not yet made publicly available on the web would be the most important to analyse, in order to appraise the MTOB approach on a SL unseen to any LLM training. It would also be important to adopt techniques for LM efficiency in low-resource scenarios (Warstadt et al., 2023).

Representations: Decisions around the appropriate representation in text, or even the medium itself (visual *versus* textual) are perhaps the most important that need to be made for the proposed approach.

SL grammars are likely to use glosses[5] to represent signs in examples and glossaries as well as in parallel corpora with continuous SL data[6]. Otherwise, a notation system such as SignWriting[7] could be used. It is compatible with the MTOB approach, as its characters are encoded in Unicode or translatable to ASCII (Jiang et al., 2023).

As for the medium - the multimodality of SLs alongside the ability of models like Gemini 1.5 Pro (Reid et al., 2024) to interpret visual, audio, or text data make a *true zero-shot* study a complex, but exciting prospect.

Evaluation: Model output in MTOB and other *few* and *zero-shot* methods has been evaluated with automatic metrics solely on text. Character based metrics such as CHrF (Popović, 2015; Bapna et al., 2022; Ruder et al., 2023), have been used for languages which are low resource, do not have clear token boundaries, or using non-romanised characters (Tanzer et al., 2024). These metrics may be suitable for SLs which are low resource, and may be notated in a system like SignWriting.

It may be possible to use BLEU (Papineni et al., 2002), standard in MT, but is known to be problematic in languages where there is only one reference translation. In addition, if SL data is presented as linear glosses, BLEU (which relies on tokenised text) may be an appropriate metric.

Further considerations: The principal users and guardians of SLs, and their related technologies, is the Deaf and Hard-of-Hearing (DHH) community. As such, it is essential to work under the principle of "nothing about us without us" (Vandeghinste et al., 2023). DHH stakeholders must consent to this technology being investigated, the use of SL data and resources, as well as being involved in the research itself.

3 Call to arms

In summary, recent research has shown that it is possible to show multimodal LLMs, within prompts, entire language descriptions with examples from book-length texts. Then, they have been shown to be able to provide translations between English and a language which has never been seen by the LLM.

This extended abstract shows the potential of extending this methodology to SLs, and intends to begin a discussion towards experimenting in LLMs with long prompt windows and SL data.

However, there remains the following open questions in order to develop this technology: **(1)** Which language pairs to target?, **(2)** How to incorporate non-text modalities?, **(3)** How to integrate image content in linguistic texts into multimodal models?, **(4)** What are the computing resources required to conduct this research?, **(5)** How to integrate the DHH community at each stage?

[4] e.g. https://glottolog.org/resource/languoid/id/cata1241 as an example for Catalan Sign Language

[5] A lexical representation based on a spoken language

[6] https://how2sign.github.io/related_datasets.html

[7] https://www.sutton-signwriting.io/

Acknowledgements

This work is part of Maria de Maeztu Units of Excellence Programme CEX2021-001195-M, funded by MCIN/AEI /10.13039/501100011033

References

Bapna, Ankur, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages.

Bragg, Danielle, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 16–31, New York, NY, USA. Association for Computing Machinery.

Bugliarello, Emanuele, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. Iglue: A benchmark for transfer learning across modalities, tasks, and languages. In *International Conference on Machine Learning*, pages 2370–2392. PMLR.

Cormier, Kearsy, Onno Crasborn, and Richard Bank. 2016. Digging into signs: Emerging annotation standards for sign language corpora. In Efthimiou, Eleni, Stavroula-Evita Fotinea, Thomas Hanke, Julie A. Hochgesang, Jette Kristoffersen, and Johanna Mesch, editors, *Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining*, pages 35–40, Portorož, Slovenia, May. European Language Resources Association (ELRA).

De Coster, Mathieu, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2023. Machine translation from signed to spoken languages: State of the art and challenges. *Universal Access in the Information Society*, pages 1–27.

De Sisto, Mirella, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov, and Horacio Saggion. 2022. Challenges with sign language datasets for sign language recognition and translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2478–2487, Marseille, France, June. European Language Resources Association.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig. 2024. Ethnologue: Languages of the World. Twenty-seventh edition.

Federmann, Christian, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In Ahuja, Kabir, Antonios Anastasopoulos, Barun Patra, Graham Neubig, Monojit Choudhury, Sandipan Dandapat, Sunayana Sitaram, and Vishrav Chaudhary, editors, *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online, November. Association for Computational Linguistics.

FitzGerald, Jack, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada, July. Association for Computational Linguistics.

Goyal, Naman, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Glottolog 5.0.

Jiang, Zifan, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2023. Machine translation between spoken languages and signed languages represented in SignWriting. In Vlachos, Andreas and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1706–1724, Dubrovnik, Croatia, May. Association for Computational Linguistics.

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

Kornai, András. 2013. Digital language death. *PloS one*, 8(10):e77056.

Maillard, Jean, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small data, big impact: Leveraging minimal

data for effective machine translation. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada, July. Association for Computational Linguistics.

Moryossef, Amit, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. In Shterionov, Dimitar, editor, *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 1–11, Virtual, August. Association for Machine Translation in the Americas.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Isabelle, Pierre, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Bojar, Ondřej, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Reid, Machel, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Ruder, Sebastian, Jonathan Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. 2023. XTREME-UP: A user-centric scarce-data benchmark for under-represented languages. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1856–1884, Singapore, December. Association for Computational Linguistics.

Simons, Gary F., Abbey L. L. Thomas, and Chad K. K. White. 2022. Assessing digital language support on a global scale. In Calzolari, Nicoletta, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Tanzer, Garrett, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book.

Vandeghinste, Vincent, Dimitar Shterionov, Mirella De Sisto, Aoife Brady, Mathieu De Coster, Lorraine Leeson, Josep Blat, Frankie Picron, Marcello Paolo Scipioni, Aditya Parikh, Louis ten Bosch, John O'Flaherty, Joni Dambre, Jorn Rijckaert, Bram Vanroy, Victor Ubieto Nogales, Santiago Egea Gomez, Ineke Schuurman, Gorka Labaka, Adrián Núnez-Marcos, Irene Murtagh, Euan McGill, and Horacio Saggion. 2023. SignON: Sign language translation. progress and challenges. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 501–502, Tampere, Finland, June. European Association for Machine Translation.

Warstadt, Alex, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell, editors. 2023. *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, Singapore, December. Association for Computational Linguistics.

Zhang, Chen, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024a. Teaching large language models an unseen language on the fly. *arXiv preprint arXiv:2402.19167*.

Zhang, Kexun, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024b. Hire a linguist!: Learning endangered languages with in-context linguistic descriptions. *arXiv preprint arXiv:2402.18025*.