Developing New Theories for New Realities: Reduce Linguistic Inequality with Multimodal Machine Translation

Xiaojun Zhang Xi'an Jiaotong-Liverpool University

Zipei Zhou Nanjing Foreign Language School

This year (2023) marks the invention of artificial intelligence generated content (AIGC) product such as OpenAI's ChatGPT and Google's BART which are anecdotally believed to change translation environment a lot. In this paper, the major changes in the translation environment over the past 30 years are reviewed and some new realities that we believe merit increased attention are proposed: the growth of linguistic inequality caused by cultural conceptualization and re-conceptualization in cross-language communication, code-switching in multilingual conversations, and sign language applied in human activities and massive media, and new technologies employed in translation practices. These realities are not completely new and scholars have already begun to explore their implications, but both the scope and the depth of these implications are growing and evolving. The aim of this research is calling for new empirical insights, new theoretical lenses, and new perspectives to shed light on these issues that are increasingly having profound impacts on society, on language strategies, and on cross-language communication.

Keywords: linguistic inequality, multimodal machine translation, code-switching, sign language translation

Linguistic inequality is a producer and reproducer of wider social, economic, and cultural inequalities due to the indexical character of language (Bonnie, 2013). This inextricable codependence between linguistic and wider social inequalities requires an interdisciplinary approach to a multidimensional phenomenon. The linguistic side involves a conception of "multilingualism" that is different from the traditional notion of "language," and a different method of analysis for language contact, displacing our vision from a "distributional" conception towards a pragmatics of intercultural communication (Moyer, 2011). It also makes us review the conception of "translation" in the new realities: Could we reduce the linguistic inequality with the succession of "turns" (Snell-Hornby, 2006) or a certain translation turn? For example, one advantage of the technical turn of translation studies is reducing the linguistic inequality of the minor or low sourced language speakers. Technologies will enable the majority to understand the minority, and vice visa, not only from the perspective of linguistics but of culture and society. Can we learn more about the phenomenon of translation if we take into account its non-lingual types? Let's start with Shashi's story in the film *English Vinglish* (2012).

Mrs Shashi Godbole in the movie *English Vinglish* is a middle-aged Maharashtrian housewife and she might be weak in English speaking. She was treated unequally at a Café shop in New York City when she ordered a sandwich because of her insufficient communication performance. Here are the subtitles of their conversation between Shashi and the crew in the Café (Figure 1). Actually, the inequality occurs in her daily life as well when she talked with her well-educated husband and daughter. Due to the linguistic inequality that lies at the core of language diversity, Shashi depressed and felt being "otherised" in the cultural shock during her visiting in New York City.

Beyond the descriptions/subtitles in Figure 1, the Indian habit of shaking her head when talking in the scenario of food ordering also makes the crew puzzled. For Indians, shaking head has its own meaning, because the movement is

---

Xiaojun Zhang's publications may be found at https://www.xjtlu.edu.cn/en/study/departments/school-of-humanities-and-social-sciences/translation-and-interpreting/department-staff/academic-staff/staff/xiaojun-zhang01. We have no conflicts of interest to disclose.

This work was supported by XJTLU Research Development Funding, Grant RDF-22-01-053.

Correspondence concerning this article should be addressed to Xiaojun Zhang, Department of Translation and Interpreting, Xi'an Jiaotong-Liverpool University, 111 Ren'ai Road, Dushuhu High Education District, Suzhou, 215123, China. E-mail: xiaojun.zhang01@xjtlu.edu.cn

00:41:12,666 --> 00:41:13,999
CC: How you doing today ma'am?
00:41:14,416 --> 00:41:16,040
SG: I want...
00:41:16,333 --> 00:41:18,415
CC: I asked how you were doing today
00:41:19,458 --> 00:41:22,165
SG: Doing...I'm doing...
00:41:22,750 --> 00:41:24,249
SG: I'm doing...
00:41:27,083 --> 00:41:28,415
CC: You can't take all that time
00:41:30,791 --> 00:41:32,790
CC: Sorry... what to eat?
00:41:33,083 --> 00:41:34,999
CC: Are you kidding me right now...
00:41:39,375 --> 00:41:40,832
SG: Vegetarian...
00:41:42,958 --> 00:41:44,540
CC: Vegetarian is fine...
00:41:44,916 --> 00:41:46,832
CC: What do you want to eat?
00:41:47,791 --> 00:41:49,082
SG: Only vegetarian...

00:41:49,208 --> 00:41:51,707
CC: A bagel...a wrap...a sandwich?
00:41:51,916 --> 00:41:52,832
SG: Sandwich
00:41:53,958 --> 00:41:55,790
CC: And what kind of filling do you want inside?
00:41:55,875 --> 00:41:56,874
CC: Do you want cheese...
00:41:56,958 --> 00:41:58,415
CC: Tomatoes... lettuce..?
00:42:00,500 --> 00:42:03,249
CC: Lady...you're holding up my line...
00:42:03,333 --> 00:42:04,790
CC: This is not rocket science
00:42:05,083 --> 00:42:06,165
CC: Cheese?
00:42:06,333 --> 00:42:07,540
SG: Yes... cheese...
00:42:07,875 --> 00:42:09,207
CC: Yes to cheese!
00:42:10,208 --> 00:42:11,540
CC: Anything to drink?
00:42:12,166 --> 00:42:13,207
SG: Water...

00:42:13,291 --> 00:42:14,790
CC: Still or sparkling?
00:42:15,750 --> 00:42:17,249
SG: Only water
00:42:19,291 --> 00:42:21,707
CC: Still or sparkling?
00:42:24,208 --> 00:42:25,082
SG: Coffee..?
00:42:26,125 --> 00:42:27,832
CC: Americano? Cappuccino? Latte?
00:42:27,958 --> 00:42:29,374
CC: Lady...I ain't got all day...
00:42:29,458 --> 00:42:34,124
CC: Americano? Cappuccino? Latte?
00:42:34,708 --> 00:42:35,999
SG: 'Nescoffee'
00:42:36,833 --> 00:42:37,457
CC: What?
00:42:37,541 --> 00:42:38,540
SG: 'Nescoffee'
00:42:38,666 --> 00:42:40,082
CC: Yes we have nice coffee...
00:42:40,166 --> 00:42:41,624
CC: We have the best coffee in Manhattan

00:42:41,750 --> 00:42:43,540
CC: I'll just give you an Americano
00:42:43,625 --> 00:42:44,874
CC: Small or medium?
00:42:46,083 --> 00:42:47,124
SG: Small
00:42:47,291 --> 00:42:49,290
CC: Small. Is that it?
00:42:50,125 --> 00:42:51,582
CC: $10.20
00:42:52,000 --> 00:42:53,165
SG: 10 dollars...
00:43:11,458 --> 00:43:14,165
CC: Hello...the least you could do is say thank you..!
00:43:14,250 --> 00:43:15,249
SG: Sorry...thank you...
00:43:17,208 --> 00:43:18,665
CC: Stupid idiot!
00:43:19,250 --> 00:43:19,832
SG: Sorry...
00:43:21,125 --> 00:43:22,874
CC: I am not cleaning that up!
00:43:26,083 --> 00:43:27,749
CC: What a stupid woman

Figure 1. Conversation Between Shashi and the Crew in the Café (CC: Café crew; SG: Shashi Godbole)

non-verbal communication. Usually someone shakes his/her head as a sign of refusing or saying "no." However, shaking head for Indian society means "good" or "yes, I understand." Actually, shaking head when listening to other people speak in India is considered politer than silence or actually not seen listening to the words of the person who is talking.

Linguistic Inequality, Code-Switching and Machine Translation

One of the most solid achievements of linguistics in the twentieth century has been to eliminate the idea that some languages or dialects are inherently "better" than others (Hudson, 1996, p. 203). Linguists recognize that some varieties of language are considered by lay people to be better than others, but they point out that each variety displays characteristics common to all human language, such as being complex and rule-governed, and that even the least prestigious language varieties reveal an impressively rich set of structural patterns. Linguists would claim that if they were simply shown the grammars of two different varieties of a completely unfamiliar language, one with high and the other with low prestige, they could not tell which was which.

Moreover, most linguists would probably say the same about linguistic differences between individual speakers (Hudson, 1996, p. 204): if there are differences between the grammars of two people, there is no way of knowing which has the higher prestige in society simply by studying the grammars. Admittedly there are individuals who clearly have inherently incomplete grammars, such as small children, foreigners and people with mental disabilities, but these deviations are easy to explain and predict, and leave intact the claim that all normal people are equal with regard to their grammars. Of course, there is no shortage of differences between grammars, whether of individuals or whole communities, but there are no purely linguistic grounds for ranking any of the grammars higher than others.

The multi-lingual speakers and language learners alternate between one or many languages always. This results in a new form of a hybrid language form called code-switching language. Code-switching can be defined as "the use of several languages or dialects in the same conversation or sentence by bilingual people" (Gardner-Chloros, 2009). It is used in particular circumstances by bilingual people who alternate between languages in an unchanged setting (Bullock & Toribio, 2009). The switches could be happened between the "major" language and the "minor" language, or between native speaking and foreign language, or between "standard" language and dialect, or between verbal and sign language, or

between language and other sounds (for example, music), or between human language and non-human language etc.

Translation technology has a significant place in almost every field of traditional human translation and change the translators' working way. Recently, machine translation (MT) has demonstrated significant progress in terms of translation quality. However, most of the research has focused on translating with pure monolingual texts in the source and the target side of the parallel corpora, when in fact code-switching is very common in communication nowadays. Despite the importance of handling code-switching in the translation task, existing MT systems fail to accommodate the code-switching content. In this paper, we examine the phenomenon of code-switching in machine translation for multimodalities.

Humans are able to make use of complex combinations of visual, auditory, tactile and other stimuli, and are capable of not only handling each sensory modality in isolation, but also simultaneously integrating them to improve the quality of perception and understanding (Stein et al., 2009). From a computational perspective, natural language processing (NLP) requires such abilities, too, in order to approach human-level grounding and understanding in various artificial intelligence (AI) tasks (Sulubacak et al., 2020). The multimodality brings new horizon for both machine translation and translation studies and a new branch of multimodal NLP in machine translation (MT), named multimodal machine translation (MMT), is growing fast, which involves both multiple modalities and different input and output languages.

Multimodal Machine Translation

Given that the additional modalities will include relevant alternative views of the input data, multimodal machine translation entails extracting information from more than one modality. An example from Specia et al. (2021) shows that the German translation of the word "hat" in the sentence "Woman covering her face with her hat" is ambiguous. To choose the appropriate word—"Hut (summer hat)" rather than "Mütze (winter hat)," the image is required. An image will be processed using multimodal machine translation and translated into another language.

The cutting-edge performance and architectural flexibility of neural sequence-to-sequence models are currently the primary factors driving the growing interest in MMT (Bahdanau et al., 2015; Sutskever et al., 2014; Vaswani et al., 2017). Since these approaches are end-to-end, this flexibility provides the potential to reunite the communities of speech, language, and vision. However historically speaking, even before the development of effective statistical machine translation (SMT) models, there was a large amount of curiosity in performing machine translation (MT) with non-text modalities. One of the earliest initiatives was the Automatic Interpreting Telephony Research project (Morimoto, 1990), a proposal from 1986 that sought to construct a pipeline of automatic speech recognition, rule-based MT, and speech synthesis, forming a comprehensive speech-to-speech translation system. Several other speech-to-speech translation systems have been developed as a result of more study (Lavie et al., 1997; Takezawa et al., 1998; Wahlster, 2000). In MMT, information is gathered from multiple modalities since it is believed that the additional modalities will provide helpful alternative perspectives on the input data. Three multimodal translation tasks—image-guided translation (IGT), video-guided translation (VGT), and spoken language translation (SLT)—were identified by Sulubacak et al. (2020) in comparison to the other two unimodal translation tasks, text-based machine translation and speech-to-speech translation.

(1) Image-guided translation (IGT) Image-guided translation is indeed a contextual grounding task in which the objective is to improve the translation of the texts by utilizing the images' semantic correspondence to the documents. One of the key motivations for this task is the effort to overcome ambiguities by visual cues.

Image caption translation, where the correspondence is linked to sentences being the descriptions of the images, is a widely used application of IGT. The first translations of image captions were mainly pipeline methods: Elliott et al. (2015) offered a pipeline of visually conditioned neural language models, and Hitschler et al. (2016) developed a new multimodal retrieval and re-ranking approach to the issue. IGT gained a lot more focus from the scientific community once the WMT multimodal translation shared task was introduced (Specia et al., 2016). The popular methods used nowadays depend on visually conditioning end-to-end neural MT systems with visual information extracted from cutting-edge pretrained convolutional neural networks (CNN).

Using images when translating captions is theoretically extremely helpful to manage grammatical features (e.g., noun genders) in translating between different languages, as well as addressing translational ambiguities, despite current debate on the efficacy of the visual modality under specific dataset and task conditions (Caglayan et al., 2019; Elliott 2018). Additionally, Caglayan et al. (2019) demonstrated

how cutting-edge models may employ the visual signal when source captions are purposefully distorted in a simulated low-resource setting.

(2) Video-guided translation (VGT) In contrast to image-guided translation, which focuses on static images connected with the textual input, although video-guided translation (VGT) is also a multimodal machine translation task, it addresses video clips (and possibly audio clips as well) instead of static images. There may be variations in video-guided translation based on the textual content. The source text could be a textual explanation of a scene or an action shown in the video, frequently written for persons who are blind or visually impaired. The source text can also include speech transcripts from the video, which are typically split as standard subtitles. Consequently, both SLT (time-variant audiovisual input) and IGT (indirect correspondence between source modalities) may cause considerable challenges for video-guided translation.

The relative insufficiency of datasets is a significant obstacle impeding development in video-guided translation. Despite the fact that a sizable collection, like the OpenSubtitles corpus (Lison & Tiedemann, 2016), can give access to a vast number of parallel subtitles, there is no associated audiovisual content because the relevant movies are not freely accessible. This problem has begun to be addressed by recent initiatives to compile freely available data for video-guided translation, such as the How2 (Sanabria et al., 2018) and VaTeX (Wang et al., 2019) datasets. We hope that these projects will stimulate more studies on video-guided translation, even though there hasn't been enough time to fully assess their effects.

(3) Spoken language translation (SLT) Commonly referred to as voice-to-text translation or automatic speech translation, it is the method of translating spoken words from one source language to another. As such, the source-side modality distinguishes it from traditional MT. Systems must establish a complicated input-output mapping because they have to conduct both modality conversion and translation at the same time, which is a challenging task. The SLT task has been influenced by several early, significant works (such as Ney, 1999 and Vidal, 1997), and it has been supported since 2004 by the speech translation tasks of the IWSLT evaluation campaign.

A pipeline strategy was traditionally used to handle SLT, effectively dividing multimodal MT into modality conversion and unimodal MT. End-to-end systems, usually based on NMT architectures, have been proposed more recently. In these systems, the source language audio sequence is directly translated into the target language text sequence (Bérard et al., 2018; Weiss et al., 2017). Although end-to-end methodologies have just recently been developed, they are quickly catching up to the pipeline systems paradigm, which has been the standard for decades.

In this paper, we intend to do some pilot studies on the MT of "minor" languages in movies which might be treated unequally. Comparing with subtitle MT of the major language in a movie, it's hard for MT to deal with the minor language phenomena including code-switching (e.g., Vinglish and Hindi in *English Vinglish* (2012)), and sign languages transferring (e.g., sign language in *CODA* (2021)) because they lack of sufficient training data for MT system. Our study will make these tasks fall under the umbrella of MMT. We will take the unequal treatment on Shashi due to her insufficient English in the movie *English Vinglish* (2012) as a case to explore IGT in using, make an experimental study on Irish sign language's translation to verify the VGT's quality in MMT.

MMT in Food Ordering Scenario

It's impossible for Shashi to hire an interpreter in New York City. The W instant translator or the other scan-for-translation tool is helpful for written texts' translation but invalid in food-ordering scenarios. Might a voice avatar translator be helpful? It depends on how accuracy of the automatic sound recognition in a noisy Café shop.

A possible and practical way is to utilize the information from image modality. Once a well-known IGT of image caption translation is realized, where the correspondence is related to sentences being the descriptions of the images, Shashi needs to take several food pictures that she wants to order and reads the simple output of English translations from the MMT system that "Good afternoon lady, I wanna order a sandwich with cheese, and a glass of water, no still no sparkling, just tap water, please, and thank you." What a good picture! Furthermore, IGT can be employed into more complicated scenarios of food-ordering, for example, to recognize local menu, or, to know of cooking receipts.

There are few things so fundamental to the human experience as food. Its consumption is intricately linked to our health, our feelings and our culture. Even migrants starting a new life in a foreign country often hold on to their ethnic food longer than to their native language. Vital as it is to our lives, food also offers new perspectives on topical challenges in computer vision like finding representations that are robust to occlusion and deformation (as occur during ingredient

processing).

We introduce Recipe1M+, a new large-scale, structured corpus of over 1m cooking recipes and 800k food images. Comparing with Food-101 dataset[1](Bossard et al., 2014) containing 101 food categories and 1,000 images for each one of these 101 food categories, totaling up to 101,000 images, Recipe1M+ is the largest publicly available collection of recipe data. It affords the ability to train high-capacity models on aligned, multi-modal data. Using these data, Marín et al. (2021) train a neural network to find a joint embedding of recipes and images that yields impressive results on an image-recipe retrieval task. Additionally, we demonstrate that regularization via the addition of a high-level classification objective both improves retrieval performance to rival that of humans and enables semantic vector arithmetic. We postulate that these embeddings will provide a basis for further exploration of the Recipe1M+ dataset and food and cooking in general. Code, data and models are publicly available[2]. Here is an example of extracting entry from Receipt 1 M+ (Figure 2):

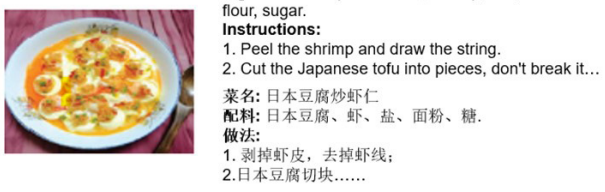

Figure 2. Extract Entry from Receipt 1 M+

Technically, natural language descriptions sometimes accompany visualizations to better communicate and contextualize their insights, and to improve their accessibility for readers with disabilities. However, it is difficult to evaluate the usefulness of these descriptions, and how effectively they improve access to meaningful information, because we have little understanding of the semantic content they convey, and how different readers receive this content. In response, we introduce a conceptual model for the semantic content conveyed by natural language descriptions of visualizations. Developed through a grounded theory analysis of 2,147 sentences, Zhu et al. (2020) developed a model which spans four levels of semantic content: enumerating visualization construction properties (e.g., marks and encodings); reporting statistical concepts and relations (e.g., extrema and correlations); identifying perceptual and cognitive phenomena (e.g., complex trends and patterns); and elucidating domain-specific insights (e.g., social and political context).

To demonstrate how their model (Zhu et al., 2020) can be applied to evaluate the effectiveness of visualization descriptions, we reduplicated the model of recipe embeddings learning (Marín et al., 2021) and generated 2,147 sentences of food image descriptions (same number with Zhu's model), and then we conducted a mixed-methods evaluation with 30 blind and 90 sighted readers, and find that these reader groups differ significantly on which semantic content they rank as most useful. Together, our findings suggest that access to meaningful information is strongly reader-specific, and that research in automatic visualization captioning should orient toward descriptions that more richly communicate overall trends and statistics, sensitive to reader preferences. Our work further opens a space of research on natural language as a data interface coequal with visualization.

A picture is worth a thousand words! The self-service is more helpful. It's even better that multimodal machine translation. KFC, in partnership with Chinese search engine giant Baidu, has opened the world's first human-free fast food restaurant in Shanghai in 2016[3]. Actually, a California-based company, Momentum Machines, has applied AI technology to create burgers since 2010, but these gaudy machines have been relegated to the kitchen, not to interact with customers. However, as to a foreigner with low-sufficient English, a picture-guided self-service machine is good enough. The tourists can move their fingers to select their favorite foods (in pictures) and click "payment" to complete this ordering (Figure 3 shows Self-service device at Burger King). And more conveniently, s/he can sit in the Café and scan a sort code to order food.

Figure 3. Burger King Self-Service

[1] https://www.kaggle.com/datasets/dansbecker/food-101.

[2] http://im2recipe.csail.mit.edu/dataset/.

[3] https://www.digitaltrends.com/cool-tech/kfc-ai-robot-restaurant.

However, beyond the scenarios of food ordering, for many people with various types of verbal disabilities such as the deaf, it's a great barrier for communication between them and the hearable speakers. Linguistic inequality caused by language disabilities is also paid more attention in the civilizing world. In this case, technology-based sign language translation finds a way to automatically translate the sign languages to natural languages.

Sign Language Translation

In the Apple TV's movie CODA (2021), American sign language (ASL) is switched with the hearing language (English) frequently. The issue raised on the created languages in the both films is their translation, subtitling and dubbing. The translators obviously lack of bilingual performance in their translating. The brand-new source language will be a challenge for every professional and skillful translator.

To solve this problem, the first way we are thinking is spoken translation. That is, we need to collect illocutors' voices and gestures and recognize them as written forms, then translate these recognized texts into target language automatically. The second way is end-to-end intersemiotic translation. That is, machine translation system will interpret the sign images or sign languages into texts directly. The assumption is that context provided by non-verbal modalities can help ground the meaning of the text and as a consequence, generate more adequate translations (Specia et al., 2021). Both of them need a sign language database which can tell a specified sign in a specified language can be corresponded to a certain word in the certain target language (for example, English). Let's take ISL STEM Glossary as an example.

ISL STEM Glossary[4]

Approximately 5,000 people in Ireland use Irish Sign Language (ISL) as their first language. ISL is a full and complex language with its own grammar, syntax and structure. But there are sometimes gaps in the vocabulary of ISL when it comes to technical subjects like Science, Technology, Engineering and Maths (STEM). To tackle this problem, Deaf Community representatives, academics from Dublin City University and other professionals came together in 2018 to build an Irish Sign Language STEM Glossary (Figure 4), led by Dr Elizabeth Mathews. When we have agreed on a sign, we record it on a phone or a tablet so we have a reference point for when we are recording during formal filming. When we are ready for filming, we record in front of a green screen in DCU with high quality video footage to work with. Because each country has its own sign Language, a number of STEM glossaries have developed in other countries. Part of the reason we decided to have an ISL STEM Glossary was because we could see the need for the glossary here and the success of glossaries in other countries. We wanted to share some of those glossaries here, but remember that the signs used here will not be suitable for learners working in Irish Sign Language.

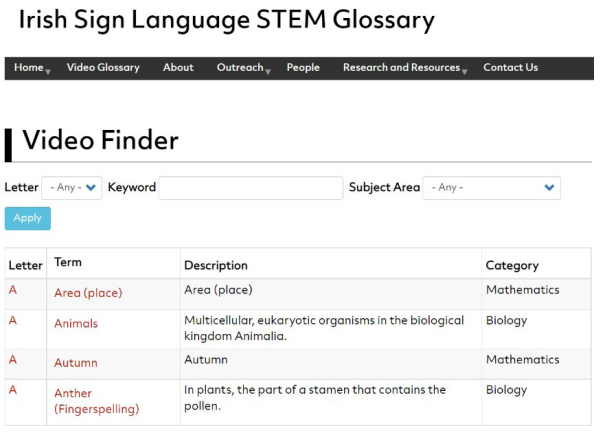

Figure 4. Irish Sign Language STEM Glossary (https://www.dcu.ie/islstem)

At present, three STEM sign language glossaries of DeafTEC, ASL STEM Forum, and ASL STEM Clear have been developed in the United States of America; two of Scottish Science Sensory and Signing Biotechnology in the United Kingdom; New Zealand Sign Language Dictionary in New Zealand; and Swedish Sign Language Dictionary in Sweden.

ISL STEM Glossary Evaluations

To verify the validity of ISL STEM Glossary in using among the deaf people in term of reducing linguistic inequality, an empirical user evaluation and an experimental technical evaluation need to be conducted in its application. Dr. Elizabeth Mathews conducted the user evaluation on maths glossary and we completed the experimental evaluation based on a machine translation system respectively.

User Evaluation

Dr. Elizabeth Mathews and her colleagues evaluated the

[4] https://www.dcu.ie/islstem.

Maths glossary[5] using three methods of data collection: (a) They carried out questionnaires with people before and after using the glossary, which collected data on attitudes especially towards talking to deaf children about maths; (b)The uptake of the evaluation was measured using user-statistics from the glossary website as well as analytics from the Youtube page hosting the videos; and (c) They interviewed teachers about their experience of using the glossary.

The results of the survey revealed the fact that 56% of participants either agreed or strongly agreed that deaf children find maths more difficult than hearing children. Perhaps one of the reasons of the difficulty in discussing maths with deaf children with 62% of participants either agreeing or strongly agreeing that discussing maths with deaf children is more difficult than with hearing children. That said, 54% of the sample felt confident discussing maths with deaf children.

They counted that their maths glossary videos had over 5,000 views over the course of the first year. We collected user data for the first 10 months: 808 entries in total to view the glossary are made up of 237 times a teacher entered, 108 times an interpreter or interpreting student entered, 105 times a parent entered and 79 times a pupil entered. The remaining 280 "other" entries indicates that there is a wide reach to groups outside the target groups.

When they interviewed teachers about their experience of using the glossary, teachers were positive overall and a number of common themes emerged. For example, it was common for teachers not to have a sign for science terms: "but I don't have a sign for Quadratic, I finger spell it and then I just do X squared." And, teachers regularly improvised when they didn't have a sign for a concept. They used finger spelling a lot "sometimes on the spot in a class I'd have to try and come up with a sign but generally to err on the side of caution I would just finger spell for the minute." The teachers also had some recommendations for how our glossary might improve.

> "I really have always felt that we needed like standardised signing and we needed teachers to be using the same signs, there was definitely a need. And I always wanted for this for a very long time. ...I think it will be a fantastic resource for all of the teachers, for educators, for interpreters, for parents, for everyone involved in educating deaf children."

The interviewer also asked parents and teachers before the evaluation began whether or not they had a sign for a selection of mathematical terms. Two interesting observations can be made from their responses. First and unsurprisingly, participants are much more likely to know signs from maths that have high application to everyday life. In particular, terms used to describe time (last year, minute, weekly, early) and location (right, on top, beside) feature in the top ten most known signs. The second observation is on professional and technical signs in maths. The signs relating to the more technical aspects of maths (improper fraction, predicting, ordinal) are much less frequently known. The low scores can to some extent be explained by the lack of terminology for some of these signs (e.g., improper fraction, standard deviation) prior to this project.

Technical evaluation

Sign Language Recognition (SLR) is a Computer Vision (CV) and Machine Learning (ML) task, with potential applications that would be beneficial to the Deaf community, which includes not only deaf persons but also hearing people who use Sign Languages. SLR is particularly challenging due to the lack of training datasets for CV and ML models, which impacts their overall accuracy and robustness. Most published results for Sign Language fingerspelling recognition, including ISL, have been obtained in controlled environments where the training and test data are derived from the same subjects and in similar data capture conditions. The performance degradation of such models, when applied to unseen and "real world" domains, known as "domain adaptation," is well documented (Charles et al., 2013).

In this experiment, we explore the use of synthetic images to augment a dataset of fingerspelling Irish signs and we evaluate whether this could be used to reliably increase the performance of an SLR system. Our model is based on a fine-tuned pretrained CNN, using skeletal wireframe image training datasets, and tested using ISL STEM Maths Glossary, a corpus dataset of real recordings of native signers. Techniques using synthetic data have been applied to problems in object detection and segmentation, face and text recognition, image classification and pose estimation (Bayraktar et al., 2019; Goyal et al., 2018; Hinterstoisser et al., 2019; Nikolenko, 2019; Peng et al., 2014). Our approach differs from the previous research by using larger synthetic datasets than those available in Sign Language corpora. Through the use of an automated framework, we can control the variations within the training dataset and can generate ground-truth frame-level annotation automatically. Furthermore, we adopt a pose

[5] The outcomes of this research are not published yet but have been reported on the page of its website: https://www.dcu.ie/islstem/maths-glossary-evaluation-isl-stem.

estimation model, MediaPipe[6], in the recognition pipeline to reduce the domain shift between training and test datasets. We use customized wireframe skeletal images to exploit the performance of current CNN models through transfer learning techniques.

An accuracy of 73.5% recognition was achieved in our experiment using synthetic images and using the pose estimation model in the test pipeline, which is better than the state-of-the-art CV models with 62.3% accuracy based on "in-the-wild" fingerspelling test datasets (Shi et al., 2019).

To avoid misunderstanding and linguistic inequality, it is possible but sounds funny for Shashi to communicate with the Café crew in sign languages. In this case, our model would achieve 73.5% accuracy in sign recognition and might read Shashi's poses as "this…this…and that, please. Thank you!" easily.

Conclusion and Further Studies

Jakobson (1959) divided translation into three types as intralingual translation, interlingual translation, and intersemiotic translation. His tripartite division of translation has been proposed and developed more than 60 years. Unfortunately, its current situation in the academic world (Jia, 2017) does not cover the phenomenon of linguistic inequality. The verbal English codes of Hinglish (Hindu English), Chicano English (Mexican-American English), Paklish (Pakistani English) and Fringlish (French English) in *English Vinglish* (2012) are very different and they may cause linguistic inequality in the community of New York English. Different registers, genres and styles of English might cause inequality. For example, Posh fascinates many non-native speakers and many English language students try to copy the accent. Some native people also try to copy the accent to make themselves seem more upper class and wealthy[7].

The multimodality brings new horizon for machine translation and translation studies. The fact that non-professional translators are doing some translation jobs with the assistance of machine translation has been accepted, no matter how you want or you do not want it happens, which even caused a threatening if robot will replace human translators in the future (Barnatt, 2017). The multimodal machine translation involves spoken language translation, image-guided translation, and video-guided translation. We introduced, duplicated and demonstrated some applications of multimodal machine translation in terms of reducing linguistic inequality, say, automatic food ordering to overcome the cultural shock, and sign language translation for deaf people. Since these researches are all pilot studies, the significance of this paper might not to verify a specified research hypothesis or to answer a detailed research question, but contribute to do the "future" translation, and more ambitious, to renew the concept of translation.

For example, human's hands are the most natural way of interacting with the world. They have 54 degrees of freedom in which they move, and no one knows how to use their hands better than the Deaf. Language should bring us together but not separate us. That's what motivated us to make a world with linguistic equality. Right now, we are starting with fingerspelling, and fingerspelling for any type of word that you want to learn. The sign language lens will help Deaf and hearing people receive feedback on their signing. The importance of this moment is that we will see a Deaf-led augmented reality (AR) evolution. It's that the Deaf will be at the centre of next generation of AR products including sign language translation with AR toolkits and technologies.

[6] https://google.github.io/mediapipe/.

[7] https://www.bloomsbury-international.com/images/ezone/ebook/how-to-be-posh.pdf.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *The 3rd International Conference on Learning Representations (ICLR 2015)*.

Barnatt, C. (2017). Replaced by the robot. *The Linguist*, *57*(4), 8–9.

Bayraktar, E., Yigit, C., & Boyraz, P. (2019). A hybrid image dataset toward bridging the gap between real and simulation environments for robotics. *Machine Vision and Applications*, *30*(1), 23–40.

Bérard, A., Besacier, L., Kocabiyikoglu, A. C., & Pietquin, O. (2018). End-to-end automatic speech translation of audiobooks. *International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*.

Bonnie, J. (2013). New Dimensions of Linguistic Inequality: An Overview. *Language and Linguistics Compass, 7*(9), 500–509.

Bossard, L., Guillaumin, M., & Van Gool, L. (2014). Food-101-mining discriminative components with random forests. *European Conference on Computer Vision*.

Bullock, B., & Toribio, A. J. (2009). Themes in the Study of Code-Switching. In B. E. Bullock & A. J. Toribio (Eds.), *The Cambridge Handbook of Linguistic Code–switching* (pp. 1–18). Cambridge: Cambridge University Press.

Caglayan, O., Madhyastha, P., Specia, L., & Barrault, L. (2019). Probing the need for visual context in multimodal machine

translation. *The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Charles, J., Pfister, T., Magee, D., Hogg, D., & Zisserman, A. (2013). Domain Adaptation for Upper Body Pose Tracking in Signed TV Broadcasts. *British Machine Vision Conference 2013*.

Elliott, D. (2018). Adversarial evaluation of multimodal machine translation. *The 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Elliott, D., Frank, S., & Hasler, E. (2015). Multi-language image description with neural sequence models. https://arxiv.org/abs/1510.04709.

Gardner-Chloros, P. (2009). *Code Switching*. Cambridge: Cambridge University Press.

Goyal, M., Rajpura, P., Bojinov, H., & Hegde, R. (2018). Dataset augmentation with synthetic images improves semantic segmentation. *National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics*.

Hinterstoisser, S., Lepetit, V., Wohlhart, P., & Konolige, K. (2019). On pretrained image features and synthetic images for deep learning. *Computer Vision – ECCV 2018 Workshop*.

Hitschler, J., Schamoni, S., & Riezler, S. (2016). Multimodal pivots for image caption translation. *The 54th annual meeting of the association for computational linguistics (ACL)*.

Hudson, R. (1996). Linguistic and social inequality. In R. A. Hunson (Eds.), *Sociolinguistics* (pp. 203–227). Cambridge: Cambridge University Press.

Jakobson, R. (1959). On Linguistic Aspects of Translation. In L. Venuti (Eds.), *The Translation Studies Reader* (pp. 113–118), London: Routledge.

Jia, H. (2017). Roman Jakobson's Triadic Division of Translation Revisited. *Chinese Semiotic Studies*, *13*(1), 31–46.

Lavie, A., Waibel, A., Levin, L., Finke, M., Gates, D., Gavalda, M., Zeppenfeld, T., & Zhan, P. (1997). JANUS-III: Speech-to-speech translation in multiple languages. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1997)*.

Lison, P., & Tiedemann, J. (2016). OpenSubtitles2016: extracting large parallel corpora from movie and TV subtitles. *The 10th International Conference on Language Resources and Evaluation (LREC)*.

Marín, J., Biswas, A., Ofli, F., Hynes, N., Salvadorm, A., Aytar, Y., Weber, I., & Torralba, A. (2021). Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(1), 187–203.

Morimoto, T. (1990). Automatic interpreting telephony research at ATR. W*orkshop on Machine Translation*.

Moyer, M. (2011). What multilingualism? Agency and unintended consequences of multilingual practices in a Barcelona health clinic. *Journal of Pragmatics, 43*(5), 1209–1221.

Ney, H. (1999). Speech translation: coupling of recognition and translation. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1999)*.

Nikolenko, S. I. (2019). *Synthetic Data for Deep Learning*. Berlin: Springer.

Peng, X., Sun, B., Ali, K., & Saenko, K. (2014). Exploring invariances in deep convolutional neural networks using synthetic images. https://arxiv.org/abs/1412.7122v2.

Sanabria, R., Caglayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., & Metze, F. (2018). How2: a large-scale dataset for multimodal language understanding. *NeurIPS, Workshop on Visually Grounded Interaction and Language (ViGIL)*.

Shi, B., Martinez Del Rio, A., Keane, J., Brentari, D., Shakhnarovich, G., & Livescu, K. (2019). Fingerspelling recognition in the wild with iterative visual attention. https://arxiv.org/abs/1908.10546.

Snell-Hornby, M. (2006). *The Turns of Translation Studies: New Paradigm or Shifting Viewpoints*. Amsterdam: John Benjamins.

Specia, L., Frank, S., Sima'an, K., & Elliott, D. (2016). A shared task on multimodal machine translation and crosslingual image description. *The 1st Conference on Machine Translation*.

Specia, L., Wang, J., Lee, S., Ostapenko, A., & Madhyastha, P. (2021). Red, spot and translate. *Machine Translation*, *35*(2), 145–165.

Stein, B., Stanford, T., & Rowland, B., (2009). The neural basis of multisensory integration in the midbrain: its organization and maturation. *Hearing Research*, *258*(1), 4–15.

Sulubacak, U., Caglayan, O., Gronroos, S., Rouhe, A., Elliott, D., Specia, L., & Tiedemann, J., (2020). Multimodal machine translation through visuals and speech. *Machine Translation*, *34*(3), 97–147.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. T*he 27th International Conference on Neural Information Processing Systems (NeurIPS)*, 3104–3112.

Takezawa, T., Morimoto, T., Sagisaka, Y., Campbell, N., Iida, H., Sugaya, F., Yokoo, A., & Yamamoto, S. (1998). A Japanese-to-English speech translation system: ATR-MATRIX. T*he 5th International Conference on Spoken Language Processing (ICSLP)*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NIPS2017)*.

Vidal, E. (1997). Finite-state speech-to-speech translation. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1997)*.

Wahlster, W. (2000). Mobile speech-to-speech translation of spontaneous dialogs: an overview of the final Verbmobil system. In W. Wahlster (Eds.), *Verbmobil: foundations of speech-to-speech translation* (pp. 3–21). Berlin: Springer.

Wang, X., Wu, J., Chen, J., Li, L., Wang, Y., & Wang, W. (2019). VATEX: a large-scale, high-quality multilingual dataset for video-and-language research. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Weiss, R. J., Chorowski, J., Jaitly, N., Wu, Y., & Chen, Z. (2017). Sequence-to-sequence models can directly translate foreign speech. *Interspeech*.

Zhu, B., Ngo, C., & Chen, J. (2020). Cross-domain Cross-modal Food Transfer. T*he 28th ACM International Conference on Multimedia (MM'20)*.