SURVEYS · TECHNOLOGY AND THE FUTURE

The Use of Machine Translation in Multilingual Survey Knowledge Collection

Machine translation can produce survey questionnaires and translate open-ended answers in seconds. This article explains how modern machine translation works, where it can help in multilingual surveys, the risks it brings for comparability and how to combine it with human expertise.

ModQu research article · 2,182 words · 21 references (APA 7th edition)

Surveys that cover several languages face a constant challenge: questions must mean the same thing in every language, and answers in many languages must be understood and analysed together. Good translation has traditionally required skilled human translators working through a careful team process, which is slow and expensive. **Machine translation**, the automatic translation of text by computer, now produces fluent translations in seconds and at almost no cost. It is tempting to use it for questionnaires, invitations and open-ended answers.

This article explains how machine translation works, what it does well and badly, where it can safely support multilingual survey work and how to keep human judgement in charge of the result.

How machine translation works

Machine translation is one of the oldest goals of computing. Research began in the 1950s, when early systems used dictionaries and rules written by linguists to convert sentences word by word and then rearrange them. These systems handled simple, controlled text but struggled with ambiguity and idiom. From the 1990s, statistical systems learned translation patterns from large collections of documents that had already been translated by people, such as the proceedings of multilingual parliaments. They produced more natural results for common phrases but often lost the structure of longer sentences. Since the mid-2010s, **neural machine translation** has become standard. Neural systems learn to translate whole sentences, taking account of context, by training on millions of pairs of sentences in two languages (Bahdanau et al., 2015). The **transformer** architecture introduced by Vaswani et al. (2017) greatly improved quality and is now the basis of most translation systems and of large language models, which can also translate (Koehn, 2020; Stahlberg, 2020).

Modern systems produce translations that are usually fluent and grammatical. For widely used language pairs and general text, their quality can approach that of professional translators on some measures. Claims of "human parity", however, have been criticised for relying on weak evaluation methods; when evaluated by professional translators, looking at whole documents rather than isolated sentences, human translations are still generally preferred (Läubli et al., 2020).

Large language models as translators

General-purpose large language models can translate as well as answer questions and write text. They offer some advantages over dedicated translation systems. A researcher can give them instructions in plain language, such as "translate this survey question into Portuguese for respondents in Brazil, keeping a neutral tone, using informal address and translating the term 'household' as defined in the notes". They can be asked to explain their choices, to offer alternative wordings or to flag terms that have no exact equivalent. These features make them useful partners in the review stage of a team translation.

They also bring additional risks. Their output can vary from one request to the next, so the same question may be translated differently on different occasions. They may follow some instructions and ignore others. They can produce confident explanations for choices that are wrong. Because commercial models are updated frequently, a translation produced one month may not be reproducible the next. Researchers who use language models for translation should save the instructions and outputs, fix the model version where possible and treat explanations as suggestions to check rather than as expert advice.

Where machine translation can help

Figure 1 shows the main points in the survey process where machine translation is used, and the human role at each.

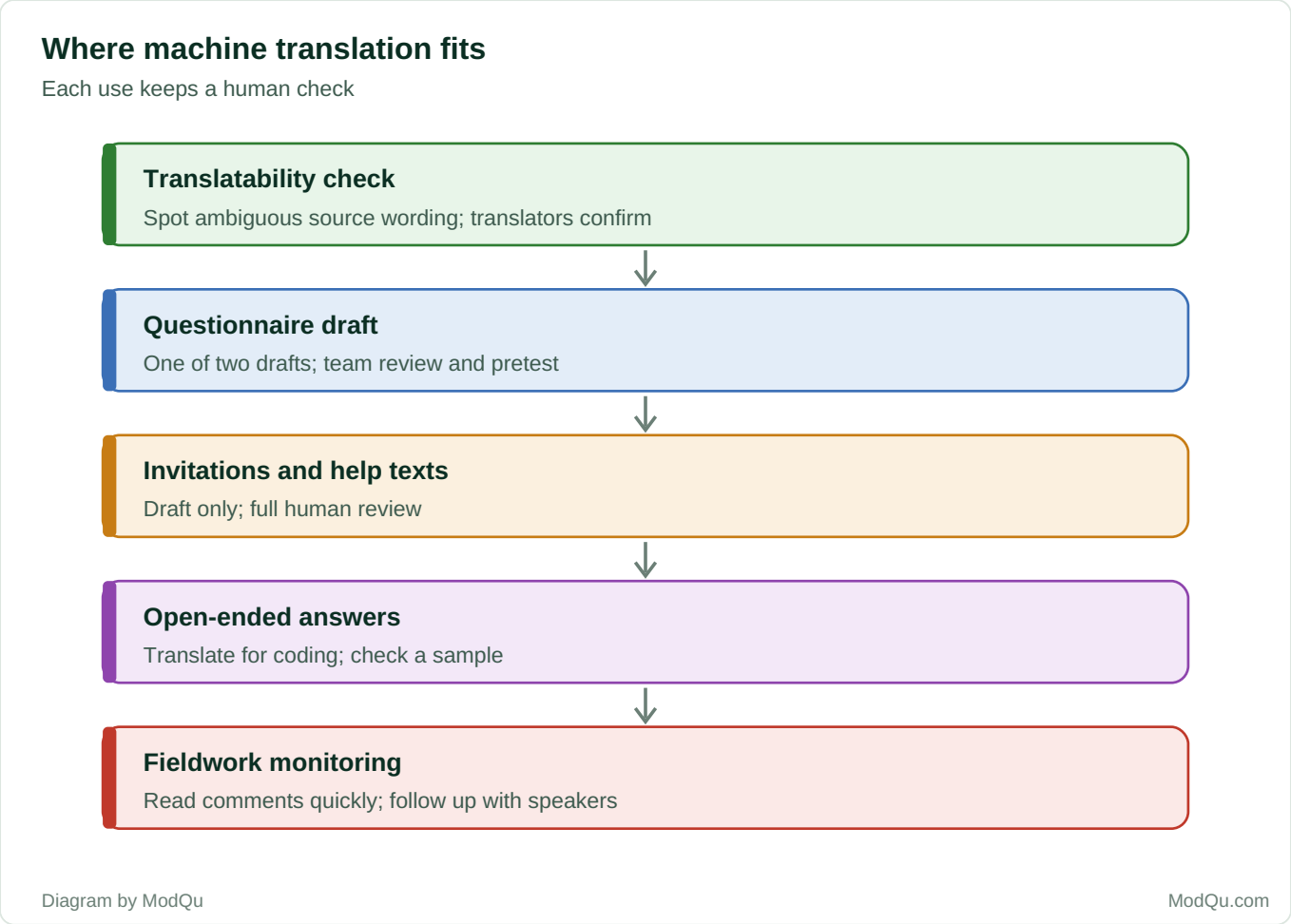


Figure 1. Uses of machine translation in multilingual surveys and the human check at each step. Diagram by ModQu.

Drafting questionnaire translations

Best practice in survey translation is a team approach: two independent translations, a review meeting, adjudication by an expert, pretesting and documentation (Harkness, 2003; Survey Research Center, 2016). Machine translation can provide one of the initial drafts, which human translators then review. In this role, it can speed up the process and highlight alternative wordings. Starting from a machine draft can also influence reviewers, who may accept fluent wording without asking whether a better choice exists, so the second, independent human translation remains important as a point of comparison. Machine translation should not replace the review, adjudication and pretesting steps that ensure functional equivalence, meaning that respondents in each language understand the question in the same way.

The editing of machine output by human translators is called **post-editing**. Studies in commercial translation have found that post-editing can be faster than translating from scratch for suitable text (Plitt & Masselot, 2010). For survey questions, where small differences in wording can change answers, post-editing must be thorough rather than light.

Checking translatability

Before a source questionnaire is finalised, researchers can translate it automatically into several languages to spot terms and constructions likely to cause difficulty. If a machine produces very different translations of a question, or translates a term inconsistently, the source wording may be ambiguous. Such checks supplement, but do not replace, review by human translators.

Translating supporting materials

Invitation letters, reminders, help texts and consent information also need translation. Machine translation can produce drafts quickly, but these texts affect whether people take part and whether their consent is informed, so they need the same human review as questions.

Translating open-ended answers

In a multilingual survey, open-ended answers arrive in many languages. Machine translation allows analysts to read and code all answers in a common language. This is often its most practical use. Errors in translating an individual answer usually have less effect than errors in a question, because they affect one response rather than all. Analysts should still check a sample of translations, especially for languages where translation quality is lower, and code answers in the original language where coders with the right skills are available.

Monitoring fieldwork

During fieldwork, machine translation can help central teams understand comments from interviewers and respondents in other languages and spot emerging problems quickly.

Risks and limitations

Fluent but wrong

Machine translations are usually fluent, which makes errors harder to spot. A translation may read naturally yet change the meaning of a question, for example by choosing a word with a stronger or weaker sense, losing a distinction between "often" and "sometimes" or dropping a qualification such as "in the last 12 months". Large language models used for translation can also add or omit content, a form of the problem known as **hallucination** (Ji et al., 2023). Because survey answers depend on precise wording, such errors can create differences between language groups that are not real.

Survey-specific requirements

General translation systems are not designed for survey questions. They do not know that response scales should be balanced, that labels on a scale should be equally spaced in meaning, or that a question must remain neutral. They may translate the same term differently in different questions, breaking consistency. They may choose a formal or informal form of address inconsistently. These are matters of survey design, which human translators with survey training are taught to handle (Harkness et al., 2010).

Uneven quality across languages

Translation quality depends on the amount of training data available. For major languages, quality is generally high; for many languages spoken by smaller or less-resourced communities, it is much lower. Joshi et al. (2020) showed that most of the world's languages receive little attention in language technology. A survey that relies on machine translation may therefore serve speakers of major languages well and speakers of minority languages poorly, creating unequal measurement quality between groups.

Figure 2 illustrates how the share of translated questions needing substantial correction might vary between languages.

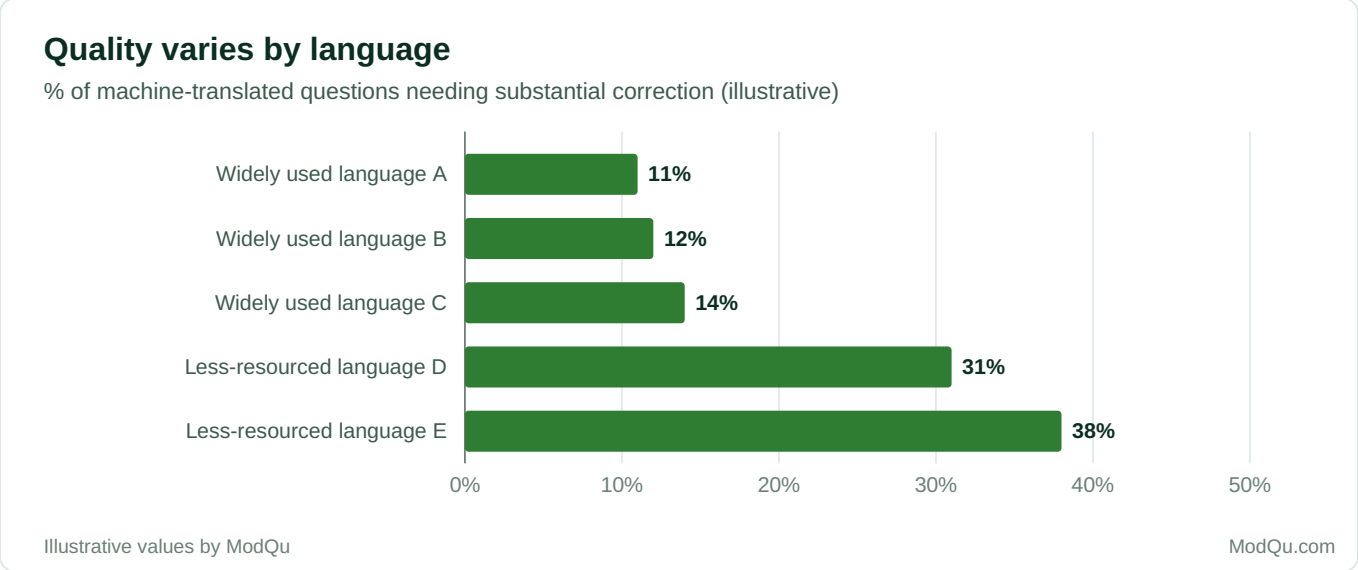


Figure 2. Share of machine-translated questions needing substantial correction by language in a hypothetical review (illustrative values). Chart by ModQu.

Bias

Machine translation systems can introduce social bias. When translating from languages without grammatical gender into languages with it, systems often assign gender according to stereotypes, such as translating a neutral word for "doctor" as male and "nurse" as female (Stanovsky et al., 2019; Prates et al., 2020). In a survey, this could change the meaning of questions about work or family roles. Language models more broadly can reproduce biases present in their training data (Bender et al., 2021).

Confidentiality

Open-ended answers often contain personal information. Sending them to an external translation service may conflict with promises made to respondents and with data protection law (European Parliament & Council of the European Union, 2016). Researchers should check where data are processed, whether they are stored or used for training and whether a secure, locally hosted system is available.

Evaluating machine translation for surveys

Automatic quality scores, such as BLEU, compare machine output with reference translations by counting shared words and phrases (Papineni et al., 2002). They are useful for comparing systems on large amounts of text but say little about whether an individual survey question keeps its meaning. For survey purposes, evaluation needs human judgement:

- **Expert review** by translators trained in survey methods, comparing machine output with the source question for meaning, neutrality and consistency.
- **Cognitive interviews** in each language, asking respondents to explain how they understood the translated question (Willis & Miller, 2011).
- **Statistical checks** after data collection, testing whether questions function equivalently across language groups (Davidov et al., 2014).

Back translation, in which a translation is translated back into the source language for comparison, is not a sufficient check for human translation (Behr, 2017), and it is even less useful for machine translation, because a machine may reproduce its own errors in both directions.

The changing role of translators

Machine translation does not remove the need for skilled translators; it changes their work. Instead of producing every word, translators increasingly review, compare and correct machine output, and spend more of their time on the decisions that matter most for surveys: preserving the meaning and tone of questions, keeping response scales balanced and ensuring that terms are used consistently. These tasks require an understanding of survey methodology as well as of language and culture.

This shift has consequences for how translation work is organised and paid. If translators are asked only to make light corrections to machine drafts, important problems may be missed. Survey organisations should allow enough time and budget for thorough review, and should involve translators in pretesting and in discussions with questionnaire designers. A translator who understands why a question was written as it was is far better placed to judge whether a machine draft has kept its meaning.

When not to rely on machine translation

Some situations call for particular caution. Questions on sensitive topics, such as health, sexuality, religion or crime, depend on carefully chosen wording that is respectful and non-judgemental in each culture; a slightly wrong tone can increase refusals or socially desirable answers. Legal texts, such as consent statements, must be accurate in every detail. Surveys of communities whose languages are poorly served by translation technology need human translators from those communities. And long-running surveys that compare results over time should not switch to machine translation without testing whether the change affects answers.

A responsible workflow

The following workflow combines the speed of machine translation with the safeguards of established practice:

1. **Prepare the source.** Write clear source questions and translation notes explaining the intended meaning of key terms.
2. **Generate a machine draft** as one of two initial translations, with a human translator producing the other.
3. **Review** both drafts in a team meeting with a survey methodologist, choosing and editing the best wording for each question.
4. **Adjudicate** remaining issues and check consistency of terms and scales across the questionnaire.
5. **Pretest** with speakers of each language.
6. **Document** which parts were machine-drafted, what was changed and why.
7. **For open answers,** translate automatically for analysis, check a sample by hand for each language and keep the original text.

This follows the structure recommended by survey translation guidelines (Survey Research Center, 2016; Pan & de la Puente, 2005) and similar guidance for translating health questionnaires (Wild et al., 2005), with the machine taking the place of one translator rather than the whole team.

An example

A research agency conducts a customer survey in six languages. To save time, it uses machine translation to produce first drafts. A bilingual reviewer for each language compares the drafts with a human translation. For the four most widely spoken languages, reviewers make minor edits to about one question in eight. For two less widely spoken languages, about one question in three needs substantial changes, including a satisfaction scale whose middle label had been translated as "indifferent", which carries a negative tone. Cognitive interviews with five respondents per language confirm the revised versions. Open-ended answers are machine-translated for coding, and a native speaker checks 10 per cent of translations in each language. The agency's report describes this process (illustrative example).

Conclusion

Machine translation has become fast, fluent and widely available, and it can make multilingual survey work quicker and cheaper. Its fluency hides errors that matter greatly for surveys: shifts in meaning, inconsistent terms, unbalanced scales, gender bias and uneven quality across languages. Used as one input to a human-led translation process, as a tool for checking translatability and as a way to read open-ended answers, it can strengthen multilingual survey knowledge. Used as a replacement for that process, it risks producing data that look comparable but are not.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*. https://arxiv.org/abs/1409.0473

Behr, D. (2017). Assessing the use of back translation: The shortcomings of back translation as a quality testing method. *International Journal of Social Research Methodology, 20*(6), 573–584. https://doi.org/10.1080/13645579.2016.1252188

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*, 55–75. https://doi.org/10.1146/annurev-soc-071913-043137

European Parliament & Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). *Official Journal of the European Union, L 119*, 1–88.

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. J. R. van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). Wiley.

Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L., Mohler, P. Ph., Pennell, B.-E., & Smith, T. W. (Eds.). (2010). *Survey methods in multinational, multiregional, and multicultural contexts*. Wiley. https://doi.org/10.1002/9780470609927

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (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* (pp. 6282–6293). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

Koehn, P. (2020). *Neural machine translation*. Cambridge University Press. https://doi.org/10.1017/9781108608480

Läubli, S., Castilho, S., Neubig, G., Sennrich, R., Shen, Q., & Toral, A. (2020). A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research, 67*, 653–672. https://doi.org/10.1613/jair.1.11371

Pan, Y., & de la Puente, M. (2005). *Census Bureau guideline for the translation of data collection instruments and supporting materials: Documentation on how the guideline was developed* (Research Report Series, Survey Methodology No. 2005-06). U.S. Census Bureau.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Plitt, M., & Masselot, F. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics*, *93*, 7–16.

Prates, M. O. R., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications*, *32*(10), 6363–6381. https://doi.org/10.1007/s00521-019-04144-6

Stahlberg, F. (2020). Neural machine translation: A review. *Journal of Artificial Intelligence Research*, *69*, 343–418. https://doi.org/10.1613/jair.1.12007

Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1679–1684). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1164

Survey Research Center. (2016). *Guidelines for best practice in cross-cultural surveys*. Survey Research Center, Institute for Social Research, University of Michigan. https://ccsg.isr.umich.edu/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30* (pp. 5998–6008). Curran Associates.

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erikson, P. (2005). Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value in Health*, *8*(2), 94–104. https://doi.org/10.1111/j.1524-4733.2005.04054.x

Willis, G. B., & Miller, K. (2011). Cross-cultural cognitive interviewing: Seeking comparability and enhancing understanding. *Field Methods*, *23*(4), 331–341. https://doi.org/10.1177/1525822X11416092