Anusandhanvallari
Vol 2026, No.1
October 2026
ISSN 2229-3388

Decoding Cultural Semantics: AI as a Tool for Cross-Language Meaning and Contextual Transfer

Kantubhuktha Surya Naveen[1], Abdul Rasheed P[2], Dr. Ramesh Kumar[3], Ms. Nanda Makhija[4], Akansh Garg[5]

[1]Assistant Librarian, Gujarat National Law University - Silvassa Campus, Sachivalaya, Secretariat Building, Amli, Silvassa, Dadra & Nagar Haveli - 396230 DNHDD, SilvassaDadra & Nagar Haveli, India,

[2]Assistant Professor, Department of English, EMEA College of Arts and Science, Kondotty Malappuram, Kerala,

[3]Ph.D, Independent Researcher, Dhanbad, Jharkhand,

[4]Assistant Professor, Department of English, JSPM University, Pune, Maharashtra nandamakhija@gmail.com

[5] Director Array Research Pvt Ltd,

Abstract: Translation has always required more than swapping words, because meaning depends on context, register, and cultural background. AI systems now translate and interpret text across hundreds of languages, but it is unclear how much cultural meaning they carry across. This paper reviews the empirical evidence on AI for cross-language meaning and contextual transfer. It combines linguistic and translation theory with studies of cross-lingual representations, massively multilingual machine translation, large language models, cultural-bias benchmarks, and evaluation methods. Five findings emerge. First, coverage has grown quickly but remains small: the largest published translation system covers 200 languages, about 3 percent of the more than 7,000 living languages. Second, transfer is uneven: a multilingual encoder improved average accuracy on a 15-language inference benchmark by 14.6 percent over its predecessor, yet for the lowest-resource 30 percent of languages the predecessor was often no better than non-BERT baselines. Third, models default to Western norms: GPT models answered a values survey like English-speaking and Protestant European countries, while cultural prompting reduced the mismatch for 71 to 81 percent of countries in newer models, so context supplied at inference is a real lever. Fourth, pivoting through English discards distinctions such as politeness level, and an illustrative calculation shows a loss of about 1.5 bits per second-person pronoun. Fifth, the same text can need up to 15 times more tokens in some languages than in others. Evaluation is weak: common metrics miss cultural fidelity, and a 55 percent preference rate on 500 comparisons carries a margin of about ±4.4 points. Two comparative tables map approaches to evidence and failure types to mitigations. The paper then proposes a context-aware cultural transfer pipeline with a native-rater scoring index. This framework is a conceptual synthesis of the reviewed literature and has not been tested experimentally.

Keywords: Cross-Lingual Transfer, Machine Translation, Cultural Semantics, Pragmatics, Large Language Models, Cultural Bias, Low-Resource Languages.

I. Introduction

Meaning does not move between languages as a fixed cargo. Whorf proposed that habitual language shapes habitual thought [1]. The strong form is rejected, but weaker effects have been measured: Russian speakers, whose language has separate basic terms for lighter and darker blue, were faster at discriminating blues that crossed that lexical boundary [2]. Berlin and Kay compared basic colour vocabulary in 20 languages and found at most 11 basic categories, added in a constrained order [3]. Wierzbicka's natural semantic metalanguage proposes roughly sixty candidate semantic primes from which culture-specific meanings are built [4]. At the pragmatic level, Hall distinguished high-context communication, where meaning rests on shared background, from low-context

communication, where words carry it explicitly [5], and Hofstede's survey work compared national value dimensions [6]. Translation theory drew the consequence. Nida argued for dynamic equivalence, matching the response of receivers instead of the form of the source [7], while Venuti criticized domesticating translation for erasing foreignness [8].

AI approaches rest on a different idea, the distributional hypothesis. Harris argued that linguistic units can be characterized by the contexts in which they occur [9], and Firth summarized it as knowing a word by the company it keeps [10]. Word embeddings operationalized it at scale [11]. Mikolov and coauthors showed that embedding spaces for different languages can be related by a linear map, which allowed word translation [12], and Conneau and coauthors later did this without parallel data [13]. Bender and Koller argued that systems trained only on linguistic form cannot, in principle, learn meaning without grounding in something outside the text [14]. Whether AI systems capture culture-specific meaning is therefore an empirical question.

The scale of the problem is large. Ethnologue lists more than 7,000 living languages [15]. In India alone, the 2011 Census recorded 121 languages with at least 10,000 speakers, and the Constitution schedules 22 [16]. A person writing in Bengali, Tamil, or Maithili is served unevenly by technology built mostly around English.

This paper uses contextual transfer to mean carrying across languages not only denotation but also implicature, register, and culturally specific reference. It reviews how AI systems approach each, how they are evaluated, and where the evidence is weak.

Research Objectives

- To review linguistic and translation theory on culture-specific meaning and its implications for AI.
- To compare cross-lingual representation, machine translation, and large-language-model approaches to meaning transfer.
- To assess evidence on cultural bias, cultural knowledge, and figurative and pragmatic competence.
- To evaluate the adequacy of current metrics and human evaluation for cultural fidelity.
- To propose a pipeline and an evaluation index, and to identify research priorities.

II. Literature Review

2.1 Culture-Specific Meaning and Translation Strategies

Aixelá defined culture-specific items as those whose function and connotations in the source text cause a translation problem because the item has no equivalent, or a different status, in the target culture [17]. Davies analyzed the translation of the Harry Potter books into Hebrew and French and identified a repertoire of strategies: preservation, addition, omission, globalization, localization, transformation, and creation [18]. This repertoire shows what an AI system must choose among: the question is not whether a term is translated, but which strategy fits the reader and purpose.

Pragmatics adds a second layer. Brown and Levinson described politeness as a set of strategies that speakers use to manage face, realized very differently across languages [19]. Hindi marks three degrees of formality in the second-person pronoun (tū, tum, āp), whereas English has one you. Sennrich, Haddow, and Birch showed that a neural translation model can be steered for formal or informal address in English-to-German translation by attaching a side constraint to the input [20]. Hovy and Yang argued more broadly that language technology has under-modeled social factors such as the relationship between speaker and listener [21].

An illustration shows the cost of pivoting. Suppose the three Hindi forms occur with assumed frequencies of 50, 30, and 20 percent. The entropy of that choice is $-(0.5 \log_2 0.5 + 0.3 \log_2 0.3 + 0.2 \log_2 0.2) \approx 1.49$ bits. A system

that translates Hindi into English and back has seen only you, so its best guess for the original form is the most common one, right 50 percent of the time. The 1.49 bits must be restored from context, such as the relationship between the speakers, or they are lost. The frequencies are assumed, but the structure holds for any distinction that the pivot language does not mark.

2.2 Cross-Lingual Representations

Cross-lingual embedding maps depend on similar geometry across languages. Søgaard, Ruder, and Vulić showed that unsupervised alignment fails for distant language pairs because their embedding spaces are not isomorphic [22]. Multilingual BERT was pretrained on Wikipedia text in 104 languages [23]. Pires, Schlinger, and Garrette found that it transfers zero-shot between languages, more successfully between typologically similar ones [24]. K and coauthors found that word-order similarity and model depth mattered more than shared vocabulary [25]. Wu and Dredze found that for the lowest-resource 30 percent of the languages they tested, ranked by Wikipedia size, multilingual BERT was often no better than, and sometimes worse than, non-BERT baselines [26].

XLM-R scaled this to 100 languages and 2.5 TB of filtered Common Crawl text. Its authors reported a 14.6 percent average accuracy gain over multilingual BERT on XNLI, and described a “curse of multilinguality” in which adding languages dilutes per-language capacity at fixed model size [27]. XNLI covers 15 languages [28]. XTREME extended evaluation to 40 languages, nine tasks, and 12 language families and reported large gaps between English and other languages, especially on structured prediction and question answering [29]. These benchmarks test denotational understanding such as entailment, entities, and answers. They do not test whether a model grasps a custom, an honorific, or a joke.

2.3 Massively Multilingual Translation and Large Language Models

Johnson and coauthors showed that one model can translate among many languages and even translate between pairs it never saw in training [30]. Arivazhagan and coauthors trained a model on more than 25 billion sentence pairs across 103 languages and documented trade-offs between high- and low-resource languages [31]. The NLLB team extended coverage to 200 languages, reported an average improvement of 44 percent in BLEU over prior state of the art, and added human-evaluated safety checks, including detection of added toxicity [32]. These are major advances, but 200 languages is about 3 percent of living languages (200 ÷ 7,000, derived).

Benchmarks shape what is measured. FLoRes-101 provides 3,001 sentences per language across 101 languages, sampled from English Wikipedia and translated into the others [33]. Because the source content is English-origin, the benchmark mostly tests translation of content that was written about topics salient in English. It does not sample cultural content native to other languages, an inference from the design and not a finding of the paper.

For India, Ramesh and coauthors compiled 49.7 million English–Indic sentence pairs across 11 languages [34], and IndicTrans2 extended translation to all 22 scheduled languages [35]. Large language models are a separate route. Hendy and coauthors found that GPT models were competitive with strong translation systems for high-resource languages and weaker for low-resource ones [36]. Cost compounds this. Petrov and coauthors found that the same text can differ in tokenized length by up to 15 times across languages, so users of some languages pay more for the same content and fit less in the context window [37]. As an illustration with assumed numbers, if a 750-word English document is 1,000 tokens and a given language expands it 4×, then processing 100,000 such documents at an assumed $10 per million tokens costs $1,000 in English and $4,000 in the other language. An 8,000-token context window holds eight English documents but two in the other language.

2.4 Evaluating Meaning Transfer

BLEU scores n-gram overlap with reference translations [38]. Callison-Burch, Osborne, and Koehn showed that it can rank systems differently from human judges, especially when systems differ in kind [39]. COMET trains a neural model on human judgments and correlates better with them [40]. Kocmi and coauthors compared metrics

for deployment decisions and found overlap-based metrics less reliable than neural ones [41]. Freitag and coauthors compared crowd ratings, expert error annotation under the Multidimensional Quality Metrics (MQM) framework, and automatic metrics on WMT 2020 outputs, and found that crowd ratings and metrics correlated weakly with expert annotation, which they treated as more reliable [42]. MQM defines a typology of error categories that evaluators mark and weight [43].

Human evaluation needs statistical care. Krippendorff recommended that conclusions rest on agreement of at least $\alpha = 0.80$, with tentative conclusions only for α between 0.667 and 0.80 [44]. Sample size matters too. With 500 pairwise comparisons, a 55 percent preference for one system has a 95 percent interval of $55 \pm 1.96 \times \sqrt{(0.55 \times 0.45 \div 500)} \approx \pm 4.4$ points, so the lower bound barely clears 50 percent. Standard adequacy and fluency scores do not capture whether a politeness level or a culture-specific reference was handled well unless the annotation scheme includes those categories.

2.5 Cultural Bias and Cultural Competence in Models

Much of the evidence underlying AI development comes from a narrow slice of humanity. Henrich, Heine, and Norenzayan reported that 96 percent of subjects in leading psychology journals came from Western industrialized countries holding 12 percent of the world's population [45]. Hershcovich and coauthors organized the challenges of cross-cultural NLP into differences in linguistic form, common ground, aboutness, and values [46].

Several benchmarks test cultural knowledge directly. MaRVL asks models to reason about images and statements from Indonesian, Mandarin, Swahili, Tamil, and Turkish speakers, with concepts chosen by native speakers, and found a large gap between models and humans [47]. GeoMLAMA probes knowledge of country-specific customs in multilingual models and found that larger models did not necessarily know more, and that prompting in a country's own language did not reliably improve accuracy on that country's customs [48]. Naous and coauthors built CAMeL, with hundreds of naturally occurring Arabic prompts and about 20,000 entities, and found that models still favored Western entities in Arabic contexts, a bias they linked to training data [49]. Tao and coauthors compared five GPT versions with survey data from many countries and found that responses resembled those of English-speaking and Protestant European countries. Cultural prompting reduced the mismatch for 71 to 81 percent of countries and territories in newer models [50]. Durmus and coauthors found that default model responses to a global opinion survey were closest to the opinions of the United States and some European and South American countries [51].

Bias also appears in word associations and translation. Bolukbasi and coauthors showed that word embeddings encode gender stereotypes about occupations [52], and Caliskan, Bryson, and Narayanan showed that embeddings reproduce known human implicit associations [53]. Stanovsky, Smith, and Zettlemoyer built WinoMT, with 3,888 sentences, to test gender in translation into eight languages and found that systems relied on stereotypes over context [54]. Savoldi and coauthors surveyed this literature and noted that gender-neutral source pronouns and gendered target languages make the problem acute [55].

Figurative language is another weak point. Dankers, Lucas, and Titov found that neural translation systems often translate idioms literally [56]. Kabra and coauthors tested figurative-language understanding in seven languages, including Hindi and Kannada, and reported large gaps relative to English, especially in lower-resource languages [57]. More broadly, Blasi, Anastasopoulos, and Neubig showed systematic inequalities in technology performance across the world's languages [58], and Bird argued that language technology work must be decolonized by working with communities on their own terms [59].

2.6 Research Gap

Representation studies measure denotation, translation studies measure overlap with references, and cultural studies measure bias in model outputs. Few connect these into a procedure that selects a translation strategy,

retrieves cultural context, and evaluates cultural fidelity with native raters. This review addresses the gap with two comparative tables and a pipeline.

III. Methodology

This study is a qualitative, comparative review with a conceptual synthesis step. Studies were included if they (i) examined cross-lingual meaning representation, machine translation, or large language models in relation to cultural, pragmatic, or figurative meaning, or evaluation of such systems; (ii) reported benchmark, experimental, or survey-based evidence, or were established theoretical works; and (iii) appeared in a peer-reviewed journal or proceedings, a standard scholarly book, or a recognized preprint archive. Studies were compared along four axes: the kind of meaning tested (denotation, pragmatics, figurative, or cultural knowledge); languages and data; evaluation method; and reported finding. Reported figures are those of the original authors, are not re-run, and are not comparable across studies because tasks, languages, and metrics differ. Illustrative calculations (the pronoun entropy, token cost, preference interval, and index example) use assumed inputs and are labeled as such. No new experiments are reported. Several sources are preprints and may have been revised, and because most cultural benchmarks cover a few languages, claims for the remaining languages, including most Indian languages, are extrapolations.

IV. Results and Discussion

4.1 Approaches, Evidence, and Limits

Table 1 maps the principal approaches onto what they transfer and the evidence on their limits.

Table 1. AI Approaches to Cross-Language Meaning: What Is Transferred, Evidence, and Limits

Approach	What It Transfers	Evidence Reviewed	Limit	Ref.
Aligned word embeddings	Word-level denotation	Linear and unsupervised alignment	Fails for distant language pairs	[12], [13], [22]
Multilingual encoders	Sentence-level meaning for inference, entities, answers	104 and 100 languages; +14.6% XNLI vs. mBERT	Often no better than baselines for lowest-resource 30%	[23], [26], [27]
Multilingual benchmarks	Task performance across languages	XNLI 15 languages; XTREME 40 languages	Test denotation, not culture	[28], [29]
Massively multilingual NMT	Text translation	103 to 200 languages; +44% BLEU reported	Benchmarks English-origin; about 3% of languages	[31], [32], [33]
Indic translation resources	Translation for Indian languages	49.7M pairs, 11 languages; 22 scheduled languages	Cultural fidelity not reported	[34], [35]
LLM translation	Translation with context in the prompt	Competitive for high-resource languages	Weaker for low-resource; token cost up to 15×	[36], [37]
Politeness control	Register	Side constraints steer formal or informal output	Requires register labels	[20]
Cultural probing benchmarks	Cultural knowledge and values	MaRVL, GeoMLAMA, CAMeL, global opinion tests	Western default; gaps vs. humans	[47], [48], [49], [51]

4.2 Discussion: What the Evidence Supports

Table 1 shows that progress is clearest on denotation. Aligned embeddings, multilingual encoders, and multilingual translation systems each improved benchmark results, and the 14.6 percent XNLI gain and the 44 percent reported BLEU gain are substantial [27], [32]. But the benchmarks measure whether a statement entails another, an entity is found, or a sentence overlaps a reference. They do not measure whether the right honorific, idiom, or culture-specific reference is used.

Cultural competence is weaker, and its pattern is consistent across independent studies. Models favor Western entities in Arabic prompts, answer value questions like English-speaking countries, and trail humans on culture-specific reasoning in Indonesian, Swahili, Tamil, and other languages [47], [49], [50]. The GeoMLAMA finding that size and prompt language did not reliably help suggests that scale alone will not close the gap [48]. The training-data explanation offered in the CAMeL study is plausible but not proven for every model [49].

The cultural-prompting result gives a practical lever. If telling a model to answer as a person from a given country shifts responses for 71 to 81 percent of countries in newer models, then context supplied at inference can substitute partly for context missing from training [50]. The remaining 19 to 29 percent of countries are a reminder that the lever is partial and uneven. The pivot problem is structural. Many systems translate through English, and the pronoun illustration shows that the information a pivot discards cannot be recovered without context [20], [21]. This affects not only politeness but also gender, evidentiality, and kinship terms, which many languages mark and English does not [55].

Cost and access form a second layer of inequity. If the same content needs up to 15 times more tokens in some languages, speakers of those languages face higher prices and shorter effective context, even where translation quality is equal [37]. The finding that low-resource languages are served worse by both encoders and LLMs compounds this [26], [36], [58].

4.3 Failure Types, Evidence, and Mitigations

Table 2 sets the principal failure types against evidence and mitigations.

Table 2. Failure Types in Cross-Cultural Meaning Transfer: Evidence and Mitigation

Failure Type	Evidence Reviewed	Mitigation	Ref.
Loss of register and honorifics	Pivot discards a three-way distinction (illustration: about 1.49 bits)	Explicit register control; speaker-relationship context	[19], [20], [21]
Literal idiom translation	NMT often translates idioms word for word; gaps in figurative understanding across seven languages	Idiom detection; figurative-language data; native review	[56], [57]
Gender and stereotype errors	Stereotype reliance across eight target languages; biased embeddings	Gender-neutral handling; bias tests in evaluation	[52], [53], [54], [55]
Western default	Western entities preferred in Arabic prompts; values near English-speaking countries	Cultural prompting; local data; cultural benchmarks	[49], [50], [51]
Mishandled culture-specific items	Strategy repertoire not selected deliberately	Strategy selection by purpose; native review	[17], [18]

Tokenization inequity	Up to 15× difference in tokens across languages	Tokenizer redesign; transparent pricing	[37]
Metric blindness	BLEU diverges from human ranking; crowd ratings weak vs. experts	MQM with native experts; agreement thresholds	[39], [41], [42], [44]
Data and technology inequality	Systematic performance gaps across languages	Community-led data collection	[58], [59]

4.4 Discussion: Evaluation and Governance

Table 2 shows that the failures are partly technical and partly a matter of who defines quality. Metrics built on overlap with one reference reward fidelity to form, which the Nida–Venuti debate shows is only one of several goals [7], [8], [38]. The Freitag results suggest that conclusions drawn from crowd ratings or metrics alone can mislead [42].

Human evaluation is necessary but should be reported with its uncertainty. The 55 percent preference example shows that moderate sample sizes cannot separate close systems [44]. Annotators need to be native speakers from the target community, and agreement should be reported by error category, since cultural fidelity is likely to show lower agreement than fluency.

Governance follows from the language-equity evidence. The inequalities documented by Blasi and coauthors and the argument by Bird imply that communities whose languages are being modeled should have a role in data collection, evaluation, and deployment decisions [58], [59]. For high-stakes content such as legal, medical, or religious text, a system that cannot establish its cultural fidelity should route to a human.

4.5 A Context-Aware Cultural Transfer Pipeline and Index

The reviewed evidence suggests a design and evaluation approach. It is a conceptual synthesis from the evidence above and has not been tested.

1. **Classify the unit.** Label each segment as literal, idiom or figurative, culture-specific item, register-marked, or humor, using the categories in the strategy and error literature [17], [18], [43].
2. **Retrieve cultural context.** For flagged segments, retrieve entries from a curated, community-maintained glossary of culture-specific items, idioms, and register rules, and supply them to the model, following retrieval-augmented generation [60].
3. **Set the context explicitly.** Supply the target audience, the relationship between speaker and listener, and the purpose, and use cultural prompting for the target region [20], [50].
4. **Select a strategy.** Choose among preservation, addition, localization, and the other strategies according to purpose, and record the choice [18].
5. **Generate and check.** Produce candidates and check for gendered or stereotyped output with bias test sets [54].
6. **Score with native raters.** Compute the Cultural Transfer Index (CTI) below on a sample, and route low-scoring categories and high-stakes content to human review [42], [44].
7. **Audit and feed back.** Monitor by language and category, publish results, and update the glossary with community input [58], [59].

The CTI is defined as CTI = 100 × (w_A·A + w_P·P + w_C·C). A is semantic adequacy, P is pragmatic appropriateness (register, politeness, implicature), and C is cultural fidelity (culture-specific items, idioms, values). Each is the mean of at least three native raters' scores, scaled from 0 to 1. The weights sum to 1 and depend on the use: legal text weights A, marketing weights C. A component is reported only if Krippendorff's α is at least 0.80, and as tentative if α is between 0.667 and 0.80 [44]. As an illustration with assumed values, A = 0.9, P = 0.7, C = 0.6, and weights of 0.4, 0.3, and 0.3 give CTI = 36 + 21 + 18 = 75. Adequacy alone would suggest 90, so the index shows a deficit in pragmatic and cultural handling that adequacy hides. Whether CTI predicts reader outcomes better than existing metrics is an empirical question that the validation step is designed to test.

V. Conclusion

This paper reviewed AI approaches to cross-language meaning and contextual transfer. Three conclusions follow. First, progress is clear on denotation: multilingual encoders and translation systems reached 100 to 200 languages and showed gains such as 14.6 percent on XNLI and 44 percent in reported BLEU, though that still covers about 3 percent of living languages [15], [27], [32]. Second, cultural and pragmatic competence lags. Models default to Western norms, translate idioms literally, rely on gender stereotypes, and lose distinctions such as honorifics in pivoting [49], [50], [54], [56]. Third, evaluation and equity are limiting factors: overlap metrics miss cultural fidelity, small human studies cannot separate close systems, and some languages pay up to 15 times more in tokens [37], [42], [44]. The practical recommendation is to supply cultural context explicitly, evaluate with native raters and reported agreement, and route high-stakes or low-confidence content to humans.

VI. Future Work

Future research should prioritize several areas. First, cultural benchmarks need to be built by native speakers of many more languages, including Indian languages, since existing ones cover few [47], [49]. Second, translation benchmarks should sample content native to the source culture, not only English-origin text [33]. Third, controlled studies should test whether retrieval of community-maintained cultural glossaries improves cultural fidelity beyond prompting alone [50], [60]. Fourth, register and honorific handling needs evaluation sets with relationship labels, extending work on politeness control [20], [21]. Fifth, tokenizer and pricing studies should measure the cost burden across languages and test remedies [37]. Finally, the pipeline and index in Section 4.5 should be implemented and validated against reader outcomes, with native-rater agreement reported by category.

References

[1] B. L. Whorf, Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf, J. B. Carroll, Ed. Cambridge, MA, USA: MIT Press, 1956.

[2] J. Winawer, N. Witthoft, M. C. Frank, L. Wu, A. R. Wade, and L. Boroditsky, "Russian blues reveal effects of language on color discrimination," Proc. Nat. Acad. Sci. USA, vol. 104, no. 19, pp. 7780–7785, May 2007.

[3] B. Berlin and P. Kay, Basic Color Terms: Their Universality and Evolution. Berkeley, CA, USA: Univ. California Press, 1969.

[4] A. Wierzbicka, Semantics: Primes and Universals. Oxford, U.K.: Oxford Univ. Press, 1996.

[5] E. T. Hall, Beyond Culture. Garden City, NY, USA: Anchor Press/Doubleday, 1976.

[6] G. Hofstede, Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations, 2nd ed. Thousand Oaks, CA, USA: Sage, 2001.

[7] E. A. Nida, Toward a Science of Translating. Leiden, The Netherlands: E. J. Brill, 1964.

[8] L. Venuti, The Translator's Invisibility: A History of Translation. London, U.K.: Routledge, 1995.

[9] Z. S. Harris, “Distributional structure,” Word, vol. 10, no. 2–3, pp. 146–162, 1954.

[10] J. R. Firth, “A synopsis of linguistic theory, 1930–1955,” in Studies in Linguistic Analysis. Oxford, U.K.: Blackwell, 1957, pp. 1–32.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv:1301.3781, 2013.

[12] T. Mikolov, Q. V. Le, and I. Sutskever, “Exploiting similarities among languages for machine translation,” arXiv:1309.4168, 2013.

[13] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, “Word translation without parallel data,” in Proc. Int. Conf. Learn. Representations (ICLR), 2018.

[14] E. M. Bender and A. Koller, “Climbing towards NLU: On meaning, form, and understanding in the age of data,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2020, pp. 5185–5198.

[15] D. M. Eberhard, G. F. Simons, and C. D. Fennig, Eds., Ethnologue: Languages of the World, 26th ed. Dallas, TX, USA: SIL International, 2023.

[16] Office of the Registrar General and Census Commissioner, India, “Census of India 2011: Language – India, States and Union Territories (Table C-16),” New Delhi, India, 2018.

[17] J. Aixelá, “Culture-specific items in translation,” in Translation, Power, Subversion, R. Álvarez and M. C.-Á. Vidal, Eds. Clevedon, U.K.: Multilingual Matters, 1996, pp. 52–78.

[18] E. E. Davies, “A goblin or a dirty nose? The treatment of culture-specific references in translations of the Harry Potter books,” The Translator, vol. 9, no. 1, pp. 65–100, 2003.

[19] P. Brown and S. C. Levinson, Politeness: Some Universals in Language Usage. Cambridge, U.K.: Cambridge Univ. Press, 1987.

[20] R. Sennrich, B. Haddow, and A. Birch, “Controlling politeness in neural machine translation via side constraints,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL-HLT), 2016, pp. 35–40.

[21] D. Hovy and D. Yang, “The importance of modeling social factors of language: Theory and practice,” in Proc. NAACL-HLT, 2021, pp. 588–602.

[22] A. Søgaard, S. Ruder, and I. Vulić, “On the limitations of unsupervised bilingual dictionary induction,” in Proc. 56th Annu. Meeting ACL, 2018, pp. 778–788.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[24] T. Pires, E. Schlinger, and D. Garrette, “How multilingual is multilingual BERT?” in Proc. 57th Annu. Meeting ACL, 2019, pp. 4996–5001.

[25] K. K, Z. Wang, S. Mayhew, and D. Roth, “Cross-lingual ability of multilingual BERT: An empirical study,” in Proc. ICLR, 2020.

[26] S. Wu and M. Dredze, “Are all languages created equal in multilingual BERT?” in Proc. 5th Workshop Representation Learn. NLP, 2020, pp. 120–130.

[27] A. Conneau et al., “Unsupervised cross-lingual representation learning at scale,” in Proc. 58th Annu. Meeting ACL, 2020, pp. 8440–8451.

[28] A. Conneau et al., "XNLI: Evaluating cross-lingual sentence representations," in Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP), 2018, pp. 2475–2485.

[29] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, "XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation," in Proc. 37th Int. Conf. Mach. Learn. (ICML), 2020, pp. 4411–4421.

[30] M. Johnson et al., "Google's multilingual neural machine translation system: Enabling zero-shot translation," Trans. Assoc. Comput. Linguistics, vol. 5, pp. 339–351, 2017.

[31] N. Arivazhagan et al., "Massively multilingual neural machine translation in the wild: Findings and challenges," arXiv:1907.05019, 2019.

[32] NLLB Team, "Scaling neural machine translation to 200 languages," Nature, vol. 630, pp. 841–846, Jun. 2024.

[33] N. Goyal et al., "The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation," Trans. Assoc. Comput. Linguistics, vol. 10, pp. 522–538, 2022.

[34] G. Ramesh et al., "Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages," Trans. Assoc. Comput. Linguistics, vol. 10, pp. 145–162, 2022.

[35] J. Gala et al., "IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages," Trans. Mach. Learn. Res., Dec. 2023.

[36] A. Hendy et al., "How good are GPT models at machine translation? A comprehensive evaluation," arXiv:2302.09210, 2023.

[37] A. Petrov, E. La Malfa, P. H. S. Torr, and A. Bibi, "Language model tokenizers introduce unfairness between languages," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 36, 2023.

[38] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in Proc. 40th Annu. Meeting ACL, 2002, pp. 311–318.

[39] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluating the role of BLEU in machine translation research," in Proc. 11th Conf. Eur. Chapter ACL (EACL), 2006, pp. 249–256.

[40] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for MT evaluation," in Proc. EMNLP, 2020, pp. 2685–2702.

[41] T. Kocmi, C. Federmann, R. Grundkiewicz, M. Junczys-Dowmunt, H. Matsushita, and A. Menezes, "To ship or not to ship: An extensive evaluation of automatic metrics for machine translation," in Proc. 6th Conf. Mach. Transl. (WMT), 2021, pp. 478–494.

[42] M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, and W. Macherey, "Experts, errors, and context: A large-scale study of human evaluation for machine translation," Trans. Assoc. Comput. Linguistics, vol. 9, pp. 1460–1474, 2021.

[43] A. Lommel, H. Uszkoreit, and A. Burchardt, "Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics," Tradumàtica, no. 12, pp. 455–463, 2014.

[44] K. Krippendorff, Content Analysis: An Introduction to Its Methodology, 2nd ed. Thousand Oaks, CA, USA: Sage, 2004.

[45] J. Henrich, S. J. Heine, and A. Norenzayan, "The weirdest people in the world?" Behav. Brain Sci., vol. 33, no. 2–3, pp. 61–83, Jun. 2010.

[46] D. Hershcovich et al., "Challenges and strategies in cross-cultural NLP," in Proc. 60th Annu. Meeting ACL, 2022, pp. 6997–7013.

[47] F. Liu, E. Bugliarello, E. M. Ponti, S. Reddy, N. Collier, and D. Elliott, "Visually grounded reasoning across languages and cultures," in Proc. EMNLP, 2021, pp. 10467–10485.

[48] D. Yin, H. Bansal, M. Monajatipoor, L. H. Li, and K.-W. Chang, "GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models," in Proc. EMNLP, 2022, pp. 2039–2055.

[49] T. Naous, M. J. Ryan, A. Ritter, and W. Xu, "Having beer after prayer? Measuring cultural bias in large language models," in Proc. 62nd Annu. Meeting ACL, 2024, pp. 16366–16393.

[50] Y. Tao, O. Viberg, R. S. Baker, and R. F. Kizilcec, "Cultural bias and cultural alignment of large language models," PNAS Nexus, vol. 3, no. 9, Art. no. pgae346, Sep. 2024.

[51] E. Durmus et al., "Towards measuring the representation of subjective global opinions in language models," arXiv:2306.16388, 2023.

[52] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 29, 2016, pp. 4349–4357.

[53] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," Science, vol. 356, no. 6334, pp. 183–186, Apr. 2017.

[54] G. Stanovsky, N. A. Smith, and L. Zettlemoyer, "Evaluating gender bias in machine translation," in Proc. 57th Annu. Meeting ACL, 2019, pp. 1679–1684.

[55] B. Savoldi, M. Gaido, L. Bentivogli, M. Negri, and M. Turchi, "Gender bias in machine translation," Trans. Assoc. Comput. Linguistics, vol. 9, pp. 845–874, 2021.

[56] V. Dankers, C. G. Lucas, and I. Titov, "Can Transformer be too compositional? Analysing idiom processing in neural machine translation," in Proc. 60th Annu. Meeting ACL, 2022, pp. 3608–3626.

[57] A. Kabra, E. Liu, S. Khanuja, A. F. Aji, G. Winata, S. Cahyawijaya, A. Aremu, P. Ogayo, and G. Neubig, "Multi-lingual and multi-cultural figurative language understanding," in Findings Assoc. Comput. Linguistics: ACL 2023, 2023, pp. 8269–8284.

[58] D. Blasi, A. Anastasopoulos, and G. Neubig, "Systematic inequalities in language technology performance across the world's languages," in Proc. 60th Annu. Meeting ACL, 2022, pp. 5486–5505.

[59] S. Bird, "Decolonising speech and language technology," in Proc. 28th Int. Conf. Comput. Linguistics (COLING), 2020, pp. 3504–3519.

[60] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 9459–9474.