Context-8: A Data Set for Evaluating Context Sensitivity in Machine Translation

Dongyue Wang, Kyo Kageura

The University of Tokyo
dongyue@g.ecc.u-tokyo.ac.jp, kyo@p.u-tokyo.ac.jp

Abstract

Context plays a crucial role in translation, enhancing both accuracy and fluency. With the advancement of machine translation (MT), the concept of context is now considered across an increasingly broader range of phenomena. Despite its importance, however, systematic definitions of context provided by communication studies and translation studies remain fragmented, and the concept of context remains elusive in MT research. To the best of our knowledge, no dataset currently exists that comprehensively evaluates MT's sensitivity to context. In this study, we propose a systematic taxonomy of context and introduce Context-8, an evaluation dataset designed to assess context sensitivity in MT for English-to-Japanese translation. The initial release includes 130 groups comprising 533 English-to-Japanese translation examples, each requiring different context categories to produce accurate and fluent translations. The data are taken from both hand-crafted and online materials. We release Context-8[1] to support the evaluation and benchmarking of MT systems with respect to context sensitivity.

Keywords: context, dataset, machine translation, English-to-Japanese translation

1. Introduction

Understanding context is considered a *sine qua non* for translation, and its importance is widely acknowledged in translation studies. In recent years, Machine Translation (MT) research has increasingly incorporated context at various levels, as the field moves from sentence-level towards text-level translation (Bahdanau et al., 2015; Tiedemann and Scherrer, 2017; Maruf et al., 2021). Despite its significance, the concept of context has not been systematically clarified. The term is used across numerous domains, including those that address beyond language-based communication (Brézillon, 2013), and is often "defined rather impressionistically" (van Dijk, 2006). This situation is unsurprising given the breadth and flexibility of the term's basic meaning. It is used to cover both linguistic and non-linguistic environments within which an utterance occurs.

If we compare the range of concepts of context addressed in translation studies and in MT research, we find that translation studies employ a broader conception that covers both linguistic and non-linguistic context, whereas MT research tends to adopt a narrower, linguistic view. This may reflect the data-driven framework adopted in MT research, which in turn can limit how context sensitivity is understood and thus evaluated. To evaluate fully how contexts are reflected in MT systems, we need test sets that cover cases the translation of which requires taking into account a wide range of context categories. While there are some datasets for MT evaluation that take context into account, they are defined in task-oriented manner and they focus on some specific context categories instead of offering a comprehensive coverage.

In this paper, we first clarify the concept of context in language communication by reviewing existing definitions and proposing a taxonomy, in which context is divided into linguistic context with 3 categories and non-linguistic context with 5 categories. We then introduce Context-8, a context-sensitive evaluation dataset for MT. Context-8 covers eight context categories and is composed of groups of English-to-Japanese translation examples where different context categories need to be considered for translation. Its first release contains 130 groups (533 examples, 1102 sentences in total). While examples within the same group share the same translation focus, they differ in the remaining parts, which often lead to different translations of the focus. The data are drawn from hand-crafted and online sources. We release the dataset to enable the assessment of context sensitivity in MT systems.

This paper's contributions are as follows:

- a taxonomy that gives an overall picture of the concepts of context based on the review of context addressed in translation studies and MT research;
- Context-8, a context-sensitive evaluation dataset for English-to-Japanese translation, which enables the evaluation and diagnosis of context sensitivity of MT models; and
- a lightweight data-construction workflow that integrates hand-crafted items with web-sourced examples for building similar datasets for other language pairs.

In the following sections, we first review related work on context addressed in translation studies and MT research. We then introduce a taxonomy of context and detail the construction of Context-8 including design principles, the constructing

[1] https://github.com/utlis/context-8

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4902–4920
11–16 May 2026.

procedures, data sources, dataset statistics and descriptions. We finally report illustrative experiments using Context-8 and conclude with future work and limitations.

2. Related Work

2.1 Context in Translation Studies and MT Research

Context in Translation Studies: The concepts of context in translation studies are broadly construed, encompassing both linguistic and non-linguistic context. Many studies refer to context as a critical factor for proper translation. We list below major categories of context introduced in translation studies:

- lexico-grammatical context, which refers to elements that occur within the intratextual span of texts, including not only the immediate context or surrounding units (Nida, 1969; Joos, 1972), but also the wider context (Nida, 1969; Newmark 1981).
- intertextual context, which refers to other texts except the current text, including cross-book parallels (Nida, 1969), documents with background information, existing translations (Nida, 1964), non-translational target-language texts on the same topic (Pym, 2014), and different versions of the source text in other languages (Pieta, 2023).
- pragmatic context, which refers to phenomena including implicature, speech acts, relevance and conversation, as well as nonverbal communication (Mey, 1993) and the total discourse (Nida, 1964; Lindley, 1964).
- situational context, which refers to the scene and the event in which the expression occurs (Bell, 1991; Gentzler, 1993); such scene and event is framed by field, tenor and mode features that specify the register of the text (Halliday, 1985).
- cognitive context, which includes the purposes of the author (Poyatos, 1997), the purposes of the translator (Danks and Shreve, 1997), decoding ability and potential interest of the audience (Gutt, 1991), and previous knowledge shared among communicators (Winter, 1964; Bell, 1991).
- cultural context, which refers to cultural rules, conditions and practices that govern how people talk (Lindstrom, 1992), and relevant aspects of the social setting of an utterance (Leech, 1983).

While these context categories together cover a wide range of context, they are not necessarily mutually exclusive or cover the same level of concepts of context. It is because context in translation studies is mostly discussed from a functional point of view, emphasizing the role of context in achieving translation equivalence rather than aiming to systematically categorize context. For human translators, these categories may be practically useful and sufficient, but further systematization would be still useful.

Context in MT Research: The concepts of context in MT research have recently been extended to a wider range, but they still mostly focus on linguistic elements that can be represented in text. At the level of individual definitions, explicit and straightforward definitions of context are often not given. Instead, context is understood as factors that contribute to disambiguation, cohesion, or style. MT treats context in an operational, problem-driven way.

Sentence-bound MT/SMT relied on intra-sentential cues, caches, and simple domain hints, where context means the local text itself (Carpuat and Wu, 2007; Huang et al., 2012), and character-surround context for transliteration (Goto et al., 2003) or phrase-surround context for bilingual embeddings (Wu et al., 2014). Separately, some works involve coarse side information such as domain or topic tags and conversational metadata such as speaker or time/place, used as priors rather than textual context (Zhang and Bach, 2009; Saluja et al., 2011).

Early context-aware NMT then added small cross-sentence windows on the source and target sides to handle pronouns, ellipsis, and lexical cohesion, typically using the previous sentence or next one or two sentences as context (Jean et al., 2017; Bawden et al., 2018; Voita et al., 2018).

Document-level MT consolidated this shift by conditioning on whole documents, treating context as all sentences with their discourse relations: via hierarchical attention (Maruf et al., 2019; Tan et al., 2019), selection of useful context sentences (Kang et al., 2020; Zhang et al., 2021), direct evaluation of deixis, ellipsis, and lexical cohesion (Voita et al., 2019), etc.

In parallel, work framed context as controllable attributes for usage norms, so context denotes extra-textual controls that condition form and style: domain labels (Zeng et al., 2018), honorific level from speaker relations (Hwang et al., 2021), formality signals (Kim et al., 2023), speaker role and film-genre metadata textualized with surrounding utterances (Vincent et al., 2023; Vincent et al., 2024), gender cues supplied as separate prompts (Sharma et al., 2022), etc.

Most recently, systems externalize context via retrieval and memory, so context comprises external textual resources and knowledge added to document-internal co-text: non-parametric kNN over translation states (Khandelwal et al., 2021; Wang et al., 2022), in-prompt exemplars and instructions (Agrawal et al., 2023; Hendy et al., 2023), cross-cultural entity knowledge from multilingual KGs (Conia et al., 2024), culture-specific evaluation frames (Yao et al., 2024), etc.

2.2 Datasets for MT Evaluation

MT evaluation hinges on the design and availability of suitable datasets. Over the past decades, researchers have created datasets to support system development, and facilitate both automatic and human evaluation, ranging from sentence-level benchmarks to text- and document-level resources.

Early evaluation datasets, such as Europarl (Koehn, 2005), JRC-Acquis (Steinberger et al., 2006), MultiUN (Eisele and Chen, 2010), and the WMT shared tasks sets (Callison-Burch et al., 2007), primarily consist of sentence-level parallel texts with adequacy and fluency scores. They contributed to the development and dissemination of widely used automated evaluation metrics, including BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020).

In parallel, several datasets have been proposed to support domain- or task-specific evaluation. These include the reading comprehension dataset for MT evaluation (Scarton and Specia, 2016), an extension from CREG dataset (Ott et al., 2012), M1.2 for MT engine evaluation (España-Bonet et al., 2018), a dataset for corporate websites MT evaluation (Rivera-Trigueros and Olvera-Lobo, 2021), and a human-made dataset for speech-to-text and speech translation evaluation (Miller et al., 2021). While valuable in their respective contexts, these datasets generally do not focus on contextual factors.

More recently, several datasets have been developed to explore context-dependent challenges. Bawden et al. (2018) introduces a dataset for evaluating discourse-level phenomena. Liu and Zhang (2020) presents a dataset for document-level NMT, emphasizing the necessity of broader context beyond sentence boundaries. MT-GenEval (Nadejde et al., 2022) evaluates gender translation accuracy using controlled counterfactual test cases. GuoFeng (Xu et al., 2022) and Zero Pronoun datasets (Shimazu et al., 2020) are valuable for testing zero pronoun recovery and translation. These datasets support fine-grained evaluation of linguistic phenomena, but most tend to concentrate on specific features, and may not cover a broader range of contextual considerations.

A broader orientation appears in DiaBLa (Bawden et al., 2019), for example, presents bilingual spontaneous written dialogue, enabling exploration of interactive context and speaker roles, and DOLFIN (Nakhlé et al., 2025) focuses on financial documents, where maintaining cross-sentence consistency is crucial. CoCoA-MT (Nadejde et al., 2022) explores non-linguistic contextual dimensions by providing parallel translations that vary in formality across multiple languages, making it suitable for evaluating formality control in MT output. MMTE (Wang et al., 2024) introduces a multilingual test set and associated metrics that account for metaphorical equivalence, emotional tone, and cultural nuance. These datasets mark promising directions; however, a dataset that systematically incorporate a wide range of context is still missing.

3. Taxonomy of Context

To construct Context-8, we reviewed the concepts of contexts in communication studies, translation studies and MT research, and defined a systematic taxonomy of contexts with wide coverage. The concept categories defined in this taxonomy are used in the construction of Context-8 as the contextual factors that affect translations.

3.1 Framework of the Concept of Context

To explore what context is, we reviewed 69 papers that invoke the concept of context and identified 53 distinct definitions spanning multiple levels and perspectives. We then defined a framework that organized different categories of context, referring to two communication models as a broad framework: Hymes' SPEAKING Model (Hymes, 1967) and SMCR model of communication (Berlo, 1960). Though these studies are old, they are well established as the basic model of communications and can be used as the basic starting point for our task of defining context categories. Table 1 presents the resulting framework.

A Linguistic Context	A1 Message
	A2 Language Events
	A3 Genre
B Non-linguistic Context	B1 Participants
	B2 Channel
	B3 Social-cultural System
	B4 Setting and Scene
	B5 Truth Value

Table 1: Framework of concepts of context

We then examined in detail the 53 definitions of context and grouped them by referring to the framework in Table 1. The resultant taxonomy is shown in Table 2 (See Appendix A for detailed category descriptions and examples).

As shown in Table 1, the framework of concepts of context consists of A Linguistic Context and B Non-linguistic Context.

3.2 Linguistic Context

Regarding A Linguistic Context, we define 3 categories that determine what is to be conveyed: A1 Message, A2 Language Events and A3 Genre. They correspond mainly to message and information to be delivered in the communication, including its contents and structure, and the communication process itself.

A Linguistic Context	**A1 Message**	A1.1 Code	any group of symbols that can be structured in a way that is meaningful to some person (Berlo, 1960)
		A1.2 Content	A1.2.a. discourse shared by speaker and hearer (Hurford, 2007)
			A1.2.b. information from the immediate setting surrounding an item in a text (Richards and Schmidt, 2010)
			A1.2.c. information that can be used to characterize the situation of an entity (a person/place/object) (Dey and Abowd, 2000)
		A1.3 Elements	A1.3.a. specific parts of an utterance (or text) near or adjacent to a unit which is the focus of attention (Crystal, 1991)
			A1.3.b. stretch of utterance or text in which a linguistic element occurs (Asher, 1994)
			A1.3.c. text that accompanies other text (Halliday and Hasan, 1989)
		A1.4 Structure	A1.4.a. the discourse around a word or expression (Asher, 1994)
			A1.4.b. intertextual context: relations with other texts (Halliday, 1985)
			A1.4.c. intratextual context: coherence within the text (Halliday, 1985)
			A1.4.d. a frame whose job is to frame content by delimiting that content from its surroundings (Bateson, 1972)
	A2 Language Events		A2.a. a frame that surrounds the event being examined (Goffman, 1974)
			A2.b. the configuration of field, tenor and mode features that specify the register of the text (Halliday, 1989)
	A3 Genre		categories such as poem, myth, tale, proverb, riddle, curse, prayer, oration, lecture, commercial, form letter, editorial, etc. (Hymes, 1974)
B Non-linguistic Context	**B1 Participants (Sender, Receiver, Others)**	B1.1 Communication Skills	B1.1.a. gestures of the speaker (Malinowski, 1923) (Duranti and Goodwin, 1992)
			B1.1.b. verbal and non-verbal action of the participants (Widdowson, 2004)
		B1.2 Attitudes	B1.2.a. the attentiveness of participants (Ochs, 1979)
			B1.2.b. an epistemological notion that reflects the properties of the thoughts that the speaker intended to convey (Jaszczolt, 2012)
		B1.3 Knowledge	B1.3.a. knowledge shared by the speaker and hearer of what has been said earlier (Lyons, 1968)
			B1.3.b. a lore of understanding, an approach, a perspective (Goffman, 1974)
	B2 Channel		a medium, a carrier of messages (Berlo, 1960)
	B3 Social-cultural System		B3.a. relevant conventions, beliefs, and presuppositions of the speech community (Lyons, 1968; Ochs, 1979; Halliday, 1985)
			B3.b. the social and spatial framework within which encounters are situated (Duranti and Goodwin, 1992; Lyons, 1977)
			B3.c. the total cultural setting of the utterance (Malinowski 1923; Leech, 1983; Levinson, 1983; Lindstrom, 1992)
	B4 Setting and Scene		B4.a. situations which prompt people to utter speech, include every object and happening in the universe (Bloomfield, 1953; Sperber and Wilson, 1986)
			B4.b. non-verbal goings-on — the total environment in which a text unfolds (Halliday, 1989)
			B4.c. information about its environment, such as location, time, temperature or user identity (Ryan et al., 1997)
	B5 Truth Value		the status of being true or false in a given interpretation

Table 2: Concepts of context

The Message is defined as a physical product of the source, with Code, Content, Elements, and Structure being its four features (Berlo, 1960). The Code refers to a sign system to construct messages and to express information or a system of rules to convert information from one form into another, namely language. The Content refers to what was selected by the source to express their communicative purpose, namely information. The Elements refer to individual signs and symbols used, namely specific parts. And the Structure refers to the way these elements are arranged to their order.

The Language Events refer to the process of delivering the message as a whole, and the relevant features of the situation of utterance (Lyons, 1995). And the process has an initial state, intermediary states, and a final state (van Dijk, 1977).

The Genre refers to the style or category of the message or a form or type of communication in any mode (written, spoken, digital, artistic, etc.). The notion of genre implies the possibility of identifying formal characteristics traditionally recognized (Hymes, 1974).

3.3 Non-linguistic Context

Regarding B Non-linguistic Context, we define 5 categories that dependent from linguistic range but influence communication: B1 Participants, B2 Channel, B3 Social-cultural System, B4 Setting and Scene, and B5 Truth Value. They correspond to actors involved in encoding and decoding messages, physical conditions and external environment of communication, and invisible circumstantial conditions that influence the process of message encoding and decoding.

The Participants are further divided into Sender, Receiver and Others, who can be individuals or groups. The Participants are characterized by 3 features: Communication Skills, Attitudes and Knowledge. The Communication Skills include encoding and decoding skills like gesturing to contribute to interpretation of messages being delivered. The Attitudes refer to a positive or negative stance the communicators take toward the topic of the communication. The Knowledge includes participants' understanding of the subject being discussed, the knowledge of prior, ongoing, and future actions (verbal, non-verbal), the knowledge of senders and receivers, their beliefs, and assumptions about temporal, spatial, and social settings (Ochs, 1979) and the knowledge of the world.

The Channel refers to the means the message being delivered. Choice of channel is the choice among oral, written, telegraphic, semaphore, or other medium of transmission of speech (Hymes, 1974).

The Social-cultural System refers to social and cultural rules governing the speech event and the participants' actions and reactions, including background beliefs and values common in this culture.

The Setting and Scene macroscopically refers to the physical circumstances as a frame including time and location, and microscopically refers to any other non-verbal goings-on including all objects and happenings involved in the process of communication.

The Truth Value refers to the status of being true or false in a given interpretation.

4. Data Set Construction

4.1 Design Principle

Drawing on the taxonomy of the concept of context, we define a methodology for constructing Context-8, a dataset for MT evaluation consisting of examples whose translation requires different context categories in order to produce accurate and fluent translations. The dataset is composed of groups of English-to-Japanese translation examples. We begin with the English-to-Japanese pair in the first release. While some items are language-pair-dependent, a sizable portion of Context-8 is driven by properties of the English source. Combined with our construction workflow, this can support the construction of analogous datasets for other language pairs, including those with non-English source languages. The source text (hereafter ST) consists of groups of English examples. Each group is a contrastive set containing at least two examples. Examples within the same group share the same specific part as their translation focus (hereafter FOCUS) and differ in the surrounding text (hereafter CONTEXT). The FOCUS may be a word, phrase, sentence, or longer unit. In our design, the FOCUS serves as the primary evaluation target. The CONTEXT refers to the surrounding text that influences the interpretation and translation of the FOCUS. The CONTEXT may be a phrase, sentence, paragraph, or longer unit. The first example in each group presents the FOCUS alone, and the remaining examples in that group introduce alternative CONTEXTs.

When adding CONTEXT, we consider all context categories introduced in Section 3. Holding FOCUS constant while varying CONTEXT allows controlled comparison and analysis of how different context categories affect translation. For each example, we provide a human translation as the target text (hereafter TT).

An example is defined as context-sensitive if the translation of its FOCUS, either in form or in meaning, cannot be determined accurately without reference to context. That is, the same FOCUS demands different translations in different CONTEXTs. The key guiding criterion is that

context has a direct and essential influence on the translation.

We present one example group below. The first example (4-0) presents the FOCUS, which is shown in italics. The other three examples (4-1, 4-2, 4-3) introduce different CONTEXTs but all contain the FOCUS. The translation of the FOCUS may differ across examples. For clarity, we also provide a literal English gloss for each Japanese translation in parentheses.

Example Group

ST:

4-0. *It's not working*.

4-1. Friday 3pm, hmm, *it's not working* for me.

4-2. I pressed the power button twice, but *it's not working*.

4-3. I apologized and wanted to make up with Alex, but *it's not working*.

TT:

4-0. *それは動いていません。* (*That is not moving.*)

4-1. 金曜日の午後 3 時ですか、うーん、私は*都合がつきません。*(Friday at 3 p.m., hmm, *I cannot make arrangements*.)

4-2. 電源ボタンを 2 回押したのですが、*動いていません。* (I pressed the power button twice, but *it is not moving*.)

4-3. アレックスに謝って仲直りしたかったのですが、*うまくいっていません。*(I wanted to apologize to Alex and make up, but *it is not going well*.)

4.2 Constructing Procedures

We collected examples partly from real-world sources and partly by constructing them manually. For real-world sources, we scanned multiple sources to locate segments whose FOCUS depends clearly on CONTEXT to be interpreted correctly. Candidate examples were selected based on their suitability for representing at least one context category.

During this process, outputs from existing MT systems were used internally to identify which context categories are useful at the initial stage. These MT outputs were not stored or included in the final dataset but served as diagnostic aids for selecting representative or challenging examples. The MT systems used were Google Translate[2], DeepL[3], and TexTra[4]. An LLM (ChatGPT[5]) was also used for translation. This may be regarded as begging the question, but we did not straightforwardly adopt the examples and we did not make choices depending on the performance of these systems.

When real-world CONTEXT was insufficient or when specific context categories were underrepresented, examples were manually created to simulate realistic but controlled contextual contrasts. Manual examples were constructed with careful attention to naturalness, relevance, and alignment with observed MT system weaknesses.

ST was reviewed by native English speakers to confirm grammatical accuracy and fluency. It was also reviewed by bilingual (English and Japanese) speakers to confirm that CONTEXT meaningfully affects the translation of the FOCUS. TT was provided by native Japanese speakers with advanced English proficiency.

4.3 Data Sources

As mentioned in 4.2, ST is either drawn from real-world sources or created by hand. Real-world sources include publicly available written and online materials, covering a wide range of text types and registers.

CONTEXT associated with each ST is similarly obtained from two sources. If a real-world ST contains sufficient surrounding text that serves as context, this CONTEXT is retained as-is. If ST does not include suitable CONTEXT, additional CONTEXT may be located from external real-world sources or manually written to fit the scenario. For manually created ST, CONTEXT is also obtained accordingly, either from real-world sources or manually constructed, to simulate natural yet controlled contrasts. This approach ensures flexibility in constructing context-sensitive examples while maintaining relevance to real-world translation challenges.

4.4 Dataset Statistics and Description

The first release of Context-8 consists of 130 groups, 533 examples, and 1102 sentences in total. Each example is annotated with one or more context-category labels from our taxonomy (A1–A3, B1–B5; see Section 3). In what follows, an example is assigned to a context category only when that category provides a relevant contextual cue for translating the focus accurately and fluently. Please note that context-category labels are non-exclusive: an example may be annotated with multiple context categories that jointly influence the translation of the FOCUS. We next report basic statistics of Context-8 with respect to the context categories listed in Table 2.

For A1 Message, all 130 groups of 533 examples are annotated as message-sensitive. These examples reflect a range of linguistic phenomena, including lexical translation choices such as nouns, verbs, or idiomatic expressions, cases

[2] https://translate.google.com/
[3] https://www.deepl.com/translator
[4] https://textra.nict.go.jp/
[5] https://openai.com/index/chatgpt/

where surrounding text determines the appropriate sense of a word, and cases where terms must be translated consistently across occurrences. We also cover structural phenomena, such as coherence within a text, intertextual references that shape translation, and intratextual cues that constrain how a given unit is rendered in the target language.

For A2 Language Events, 47 groups of 193 examples are annotated as language-event-sensitive, in which the type of communicative event provides the key contextual cue for translation. These examples reflect 6 types of language events with 14 categories, i.e. institutional and organizational communication (adjudication, education and academia, journalism, organizational and business communication, and public service encounters), task and operational communication (collaborative coordination, and procedural command communication), ritualized communication (ritual and ceremony), public and mediated communication (public address, and computer-mediated communication), interpersonal communication (phatic communion, and narrative discourse), and scripted and creative communication (literary adaptation, and audiovisual localization).

For A3 Genre, 24 groups of 98 examples are annotated as genre-sensitive, in which genre information provides the key contextual cue for translation. These examples reflect 30 genres (Buck, 2016), i.e. newspaper article, satire, personal letter, greeting card, academic essay, schedule, things-to-do list, political campaign posters, poetry, song lyrics, autobiographical essay, business letter or correspondence, speech or debate, testimonial, animation, textbook article, science article or report, encyclopaedia article, dialogue of a conversation among two or more people, ghost story, recipe, brochure, TV advertisement or infomercial, restaurant description or menu, how-to or directions booklet, wedding/graduation or special event invitation, website, blog posting, Facebook update, and news program story or announcement.

For B1 Participants, 30 groups of 136 examples are annotated as participant-sensitive, in which participant information provides the key contextual cue for translation. These examples reflect 12 types of interpersonal relationships, i.e. 6 personal-domain relationships (strangers, acquaintances, friends, close friends, romantic partners, and family hierarchical relations) and 6 institutional relationships (colleagues, superior and subordinate, teacher and student, professional and client, service staff and customer, and public authority and citizen); 13 common gestures/actions of participants in communication, i.e. nodding, head shaking, eye contact, smiling, frowning, hand raising, pointing, hand waving, shrugging shoulders, leaning forward, leaning back, crossing arms, and clapping; 13 types of attitudes as communicative stances, i.e. cooperative, defensive, supportive, critical, aggressive, passive, empathetic, enthusiastic, skeptical, polite, ironical, dominant, and submissive; and 7 audience types by level of understanding, i.e. expert audience, knowledgeable lay audience, general public, novice, child audience, multilingual or second-language audience, and special-needs audience.

For B2 Channel, 23 groups of 94 examples are annotated as channel-sensitive, in which channel information provides the key contextual cue for translation. These examples reflect 12 channel types, i.e. oral (face-to-face conversation, telephone, voice messaging, audio broadcast, video conference), written (handwritten letters and postcards, printed publications, email, text messaging, instant messaging, online forums), and visual-gestural (emoji).

For B3 Social-cultural System, 23 groups of 90 examples are annotated as social-cultural-sensitive, in which social-cultural information provides the key contextual cue for translation. These examples reflect culturally anchored phenomena that make context salient in translation, including culture-specific expressions, idioms and proverbs, conventional metaphors, allusions and intertextual references, genre conventions and fixed expressions, politeness norms and honorific systems, address forms and titles, taboo and euphemism, ritual or religious formulae, and measurement and calendrical systems.

For B4 Setting and Scene, 15 groups of 54 examples are annotated as setting-and-scene-sensitive, in which setting-and-scene information provides the key contextual cue for translation. These examples reflect the immediate environment of communication, including situations and objects that prompt speech, the non-verbal goings-on that form the background in which a text unfolds, and information about the physical environment.

For B5 Truth Value, 9 groups of 29 examples are annotated as truth-value-sensitive, in which truth-value information provides the key contextual cue for translation. These examples reflect 3 truth values, i.e. truth, false, and true in some world.

5. Experiments

5.1 Experimental Setup

To demonstrate the usefulness of Context-8, we conduct baseline experiments with three commercial MT systems and an LLM system. The purpose of this experiment is to provide the results of a single baseline as an initial comparison point for future efforts. Context-8 can be used either for developer-oriented diagnostics of MT/LLM systems under development or for user-oriented benchmarking of existing systems.

In this paper, we take the latter perspective and evaluate publicly available MT and LLM systems. Translations were produced by Google Translate[6], DeepL[7], TexTra[8], and an LLM (GPT-5.2)[9], yielding one Japanese output per example for each system. We normalized whitespace and preserved punctuation; no additional tokenization was applied.

5.2 Evaluations

We evaluate all examples in Context-8. We adopt the reference-free COMET-QE metric (Rei et al., 2020) with the public checkpoint *Unbabel/wmt20-comet-qe-da*. We also compute reference-based COMET scores using *Unbabel/wmt20-comet-da* to compare system outputs against the reference translation. Both metrics return a segment-level score for each example.

We also evaluate the translation of the FOCUS within each example. To obtain the corresponding target fragments, we first extract aligned spans from the system outputs using SimAlign (Jalili Sabet et al., 2020), a word alignment method based on multilingual BERT, and then manually correct the extracted fragments when necessary.

Following common practice, we report:

- system-level mean at the example level: arithmetic average across all examples
- system-level mean at the FOCUS level: arithmetic average across all FOCUS
- context-category-wise means at the example level: arithmetic averages over examples within each of the 8 context-category labels (A1–A3, B1–B5)[10]
- context-category-wise means at the FOCUS level: arithmetic averages over FOCUS within each of the 8 context-category labels (A1–A3, B1–B5)

Category	No.	COMET-QE (reference-free)				COMET-DA (reference-based)			
		Google	DeepL	TexTra	GPT-5.2	Google	DeepL	TexTra	GPT-5.2
Mean	533	0.4277	0.3883	0.3648	0.4205	0.6724	0.5937	0.6244	0.7652
A1	533	0.4277	0.3883	0.3648	0.4205	0.6724	0.5937	0.6244	0.7652
A2	193	0.4328	0.3695	0.3367	0.4101	0.7022	0.6037	0.6129	0.8061
A3	98	0.4069	0.3775	0.3547	0.3959	0.6806	0.5645	0.6067	0.7056
B1	136	0.4928	0.4525	0.4108	0.4708	0.8247	0.7785	0.7572	0.8662
B2	94	0.4608	0.4180	0.4195	0.4437	0.7849	0.6732	0.7259	0.8679
B3	90	0.2989	0.2609	0.2468	0.3046	0.2835	0.2297	0.3787	0.5300
B4	54	0.3624	0.3096	0.3169	0.3118	0.6624	0.5401	0.6520	0.7776
B5	29	0.5121	0.4972	0.5173	0.5180	0.8858	0.8106	0.8819	0.9375

Table 3: Mean COMET-QE and COMET-DA by context category (example level)

Category	No.	COMET-QE (reference-free)				COMET-DA (reference-based)			
		Google	DeepL	TexTra	GPT-5.2	Google	DeepL	TexTra	GPT-5.2
Mean	533	0.5293	0.4818	0.4661	0.4925	0.5940	0.5365	0.5645	0.6998
A1	533	0.5293	0.4818	0.4661	0.4925	0.5940	0.5365	0.5645	0.6998
A2	193	0.5162	0.4140	0.4148	0.4469	0.5522	0.4582	0.4692	0.7067
A3	98	0.5231	0.5130	0.5131	0.5331	0.6668	0.6015	0.5600	0.6503
B1	136	0.6307	0.6322	0.5450	0.6031	0.8517	0.8045	0.8230	0.9124
B2	94	0.5235	0.4850	0.4931	0.5134	0.8075	0.7084	0.7540	0.8996
B3	90	0.2966	0.2555	0.2313	0.2284	0.0077	-0.0372	0.1687	0.3003
B4	54	0.5054	0.3932	0.4571	0.4224	0.5257	0.4756	0.5583	0.7268
B5	29	0.6408	0.6516	0.6417	0.6574	0.9915	0.9904	1.0380	1.0586

Table 4: Mean COMET-QE and COMET-DA by context category (FOCUS level)

[6] https://translate.google.com/ (accessed 2026-02-28)
[7] https://www.deepl.com/translator (accessed 2026-03-01)
[8] https://textra.nict.go.jp/ (accessed 2026-03-02)

[9] We used OpenAI GPT-5.2 for English to Japanese translation (accessed 2026-03-04). Prompt: "You are a machine translation engine. Translate from English to Japanese. Do not add or omit information. Preserve meaning exactly. Output Japanese only." The model was applied independently to each example.

[10] As mentioned in Section 4.4, context-category labels are non-exclusive: an example may be annotated with multiple context categories that jointly influence the translation of the FOCUS.

5.3 Results

In this section, we summarize the experimental results. Table 3 reports mean COMET-QE and COMET-DA at the example level, and Table 4 reports the corresponding scores at the FOCUS level.

Across the MT systems, both COMET-QE (reference-free) and COMET-DA (reference-based) yield modest scores at both the example and FOCUS levels, suggesting that many items in Context-8 remain challenging for these systems. By contrast, the LLM (GPT-5.2) attains higher scores, especially under the reference-based COMET-DA metric.

Across the eight context categories, the broad ordering is similar between the example-level and FOCUS-level views, though the exact ranking varies by metric and system. In particular, B1 Participants and B2 Channel tend to be among the higher-scoring categories, whereas B3 Social-cultural System and B4 Setting and Scene tend to score lower across systems. These differences suggest that examples associated with different context categories may pose different difficulties; however, because examples can carry multiple context-category labels and data sizes are uneven, the per-category means combine system behaviour with corpus composition. Accordingly, the results should be interpreted as descriptive diagnostics, not as causal evidence that any context category is easier or harder for MT/LLM systems to handle. We therefore avoid drawing causal conclusions from the ranking alone.

Comparing the two metrics, COMET-DA scores are consistently higher than COMET-QE across systems and context categories. The absolute scales of the two metrics are not directly comparable, and we primarily use them as complementary views: COMET-QE provides a reference-free baseline, while COMET-DA provides a reference-based baseline.

Comparing the example-level and FOCUS-level results across Tables 3 and 4, we observe different tendencies under COMET-QE and COMET-DA. Under COMET-QE, FOCUS-level scores are often higher than example-level scores, suggesting that the FOCUS can be translated plausibly in isolation but becomes less appropriate when broader context is considered. In contrast, under COMET-DA, example-level scores are often higher, suggesting that matching the reference translation benefits from the broader textual context available at the example level. Since the FOCUS-level view is derived from extracted target fragments (with manual correction when necessary), we interpret such a gap as a descriptive diagnostic signal rather than a definitive measure of context benefit.

Overall, the results suggest that translating context-dependent text remains a challenging task, calling for further development. Reporting both per-example and per-FOCUS means provides complementary diagnostics, and can guide dataset growth and model development.

6. Conclusions

This study introduces a context-sensitive dataset for MT evaluation, motivated by the growing recognition of context as an important factor in both translation studies and MT research. We proposed a systematic taxonomy of context with broad coverage, clarifying that context encompasses both linguistic context and non-linguistic context, and compared how context is addressed in translation studies and in MT research; the broader treatment in translation studies offers guidance for incorporating richer contextual information into MT. We also presented a dataset that allows systematic exploration of how different context categories influence translation. The dataset is composed of groups of English-to-Japanese translation examples where FOCUS remains constant while CONTEXT varies, enabling controlled observation of context effects. The first version has demonstrated the feasibility of this approach, while also highlighting the complexity of capturing and evaluating context-sensitive translation phenomena. We hope this study will support further studies on context in MT evaluation and contribute to the development of more context-aware translation systems. Future work will expand and refine the dataset, focusing on improving the coverage and quality of examples that involve more complex context phenomena. We will also further validate the dataset with other automatic metrics and human evaluation and provide clear usage guidelines. We hope that the public release of the dataset will provide a useful resource for researchers interested in evaluating context-sensitive MT.

7. Limitations

One of the limitations in this work is that the current scale of the dataset is still small, which restricts fine-grained analyses by context category. Coverage is still uneven across context categories, and some non-linguistic categories (e.g., B4 Setting and Scene) are underrepresented. And most examples are short snippets, so results may not fully reflect long-range discourse effects.

8. Acknowledgments

We sincerely thank Professor Atsushi FUJITA, Professor Masaru YAMADA, and Professor Rei MIYATA for their encouragement, insightful comments, and hard questions. This work would not have been possible without their contributions. We also thank Mayuka YAMAMOTO for providing the Japanese reference translations for the corpus.

9. Bibliographical References

Agrawal, R., Turchi, M., and Negri, M. (2018). Contextual handling in neural machine translation: look behind, ahead and on both sides. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT 2018)*, pp. 31-40, Alicante, Spain, May. European Association for Machine Translation.

Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L. and Ghazvininejad, M. (2023). In-Context Examples Selection for Machine Translation. In *Proceedings of the Association for Computational Linguistics: ACL 2023*, pp. 8857-8873.

Al-Sulaimaan, M. M. D., and Khoshaba, L. M. (2018). Context as a basis for understanding pragmatic ambiguity with reference to Arabic. *International Journal of English Literature and Social Sciences, 3*(5), pp. 768-775. https://doi.org/10.22161/ijels.3.5.12

Akman, V., Bouquet, P., Thomason, R., and Young, R. A. (Eds.). (2001). *Modeling and using context: CONTEXT 2001*. Heidelberg, Germany: Springer.

Ahmed, M., Ouda, A., Abusharkh, M., Kohli, S., and Rai, K. (2023). An optimized approach to translate technical patents from English to Japanese using machine translation models. *Applied Sciences, 13*(12), 7126. https://doi.org/10.3390/app13127126

Asher, R. E., and Simpson, J. M. Y. (Eds.). (1994). *The encyclopedia of language and linguistics* (1st ed.). Tokyo, Japan: Pergamon Press.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, USA, May.

Baker, K. L., Franz, A. M., and Jordan, P. W. (2001). *Coping with ambiguity in knowledge-based natural language analysis* [Technical report]. Center for Machine Translation and Department of Philosophy, Carnegie Mellon University.

Bateson, G. (1972). *Steps to an ecology of mind*. New York, NY: Chandler.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018). Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1304-1313. New Orleans, LA, June. Association for Computational Linguistics.

Bawden, R., Rosset, S., Lavergne, T., and Bilinski, E. (2019). DiaBLa: A corpus of bilingual spontaneous written dialogues for machine translation. *arXiv*. https://arxiv.org/abs/1905.13354

Bazire, M., and Brézillon, P. (2005). Understanding context before using it. In *Modeling and using context: 5th International and Interdisciplinary Conference, CONTEXT 2005*, pp. 29-40. Paris, France.

Bell, R. T. (1991). *Translation and translating: Theory and practice*. London: Longman.

Berlo, D. K. (1960). *The process of communication: An introduction to theory and practice*. New York, NY: Holt, Rinehart and Winston.

Berry, D. M., Kamsties, E., and Krieger, M. M. (2003). *From contract drafting to software specification: Linguistic sources of ambiguity* (Version 1.0) [Technical report]. University of Waterloo.

Bloomfield, L. (1953). *Language*. London, UK: George Allen & Unwin Ltd.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Névéol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016). Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 131-198. Berlin, Germany, August. Association for Computational Linguistics.

Brézillon, P. (2013). Context-based development of experience bases. In *Modeling and Using Context (CONTEXT 2013)*, LNAI 8175, pp. 87-100. Berlin/Heidelberg: Springer.

Buck, D. (2016). Examples of genres. *Essentials for ENGL-121*. Howard Community College (Pressbooks). Accessed 21 February 2026.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (Meta-) Evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 136-158. Prague, Czech Republic, June. Association for Computational Linguistics.

Carpuat, M. and Wu, D. (2007). Context-dependent phrasal translation lexicons for statistical machine translation. In *Proceedings of Machine Translation Summit XI: Papers*, pp. 73-80, Copenhagen, Denmark, September.

Catford, J. C. (1965). *A linguistic theory of translation: An essay in applied linguistics*. Oxford, UK: Oxford University Press.

Conia, S., Lee, D., Li, M., Minhas, U. F., Potdar, S., and Li, Y. (2024). Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, pp. 16343-16360, Miami, Florida, USA, November. Association for Computational Linguistics.

Crystal, D. (1991). *A dictionary of linguistics and phonetics* (3rd ed.). Oxford, UK; Cambridge, USA: Blackwell.

Danks, J. H., Shreve, G. M., Fountain, S. B., and McBeath, M. K. (Eds.). (1997). *Cognitive*

processes in translation and interpreting. Thousand Oaks, CA: Sage Publications.
Dey, A. K., and Abowd, G. D. (2000). Towards a better understanding of context and context-awareness. In *Proceedings of the PrCHI 2000 Workshop on the What, Who, Where, When and How of Context-Awareness*, pp. 304-307, The Hague, Netherlands.
Duranti, A., and Goodwin, C. (Eds.). (1992). *Rethinking context: Language as an interactive phenomenon*. Cambridge, UK: Cambridge University Press.
Eisele, A., and Chen, Y. (2010). MultiUN: A multilingual corpus from United Nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, pp. 2868-2872, Valletta, Malta, May. European Language Resources Association (ELRA).
España-Bonet, C., Stiller, J., and Henning, S. (2018). M1.2 – Corpora for the machine translation engines. In *Proceedings of Computer Science, Linguistics*, pp. 1-11.
Fernandes, P., Yin, K., Neubig, G., and Martins, A. F. T. (2021). Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6467-6478. Online, August. Association for Computational Linguistics.
Fillmore, C. J. (1982). Frame semantics. In *Linguistics in the morning calm*, pp. 111-137. Seoul, South Korea: Hanshin.
Finkbeiner, R., Meibauer, J., and Schumacher, P. B. (Eds.). (2012). *What is a context? Linguistic approaches and challenges*. Amsterdam, Netherlands: John Benjamins.
Gentzler, E. (1993). *Contemporary translation theories*. London: Routledge.
Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. New York: Harper & Row.
Goto, I., Kato, N., Uratani, N., and Ehara, T. (2003). Transliteration considering context information based on the maximum entropy method. In *Proceedings of Machine Translation Summit IX: Papers*, pp. 125-132, New Orleans, LA, September.
Gutt, E.A. (1991). *Translation and relevance: Cognition and context*. Manchester: St Jerome.
Halliday, M. A. K. (1985). *An introduction to functional grammar*. London, England: Edward Arnold.
Halliday, M. A. K., and Hasan, R. (1989). *Language, context and text: Aspects of language in a social-semiotic perspective*. Oxford, UK: Oxford University Press.
Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (2018). Achieving human parity on automatic Chinese to English news translation. *arXiv*. https://arxiv.org/abs/1803.05567
Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., and Hassan Awadalla, H. (2023). How good are GPT models at machine translation? A comprehensive evaluation. *arXiv preprint* arXiv:2302.09210.
House, J. (1997). *Translation quality assessment: A model revisited*. Tübingen, Germany: Gunter Narr Verlag.
Hu, X., and Wan, X. (2023). Exploring discourse structure in document-level machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13889-13902. Singapore, December. Association for Computational Linguistics.
Huang, E., Socher, R., Manning, C., and Ng, A. (2012). Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012), Volume 1: Long Papers*, pp. 873-882, Jeju, Korea, July. Association for Computational Linguistics.
Hurford, J. R., and Heasley, B. (1983). *Semantics: A coursebook*. Cambridge, UK: Cambridge University Press.
Hurford, J. R., Heasley, B., and Smith, M. B. (2007). *Semantics: A coursebook*. Cambridge, UK: Cambridge University Press.
Hwang, Y., Kim, Y., and Jung, K. (2021). Context-aware neural machine translation for Korean honorific expressions. *Electronics*, 10(13):1589.
Hymes, D. (1967). Models of interaction of language and social setting. *Journal of Social Issues, 23*(2), pp. 8-28.
Hymes, D. H. (1974). *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia, USA: University of Pennsylvania Press.
Jalili Sabet, M., Dufter, P., Yvon, F. and Schütze, H. (2020). SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In *Proceedings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1627-1643.
Jaszczolt, K. M. (2012). Context: Gricean intentions vs. two-dimensional semantics. In R. Finkbeiner, J. Meibauer, and P. B. Schumacher (Eds.), *What is a context? Linguistic approaches and challenges*. Amsterdam, Netherlands: John Benjamins.
Jean, S., Lauly, S., Firat, O., and Cho, K. (2017). Does neural machine translation benefit from larger context? *arXiv preprint* arXiv:1704.05135.
Jiang, Y. E., Liu, T., Ma, S., Zhang, D., Sachan, M., and Cotterell, R. (2023). Discourse-centric evaluation of document-level machine

translation with a new densely annotated parallel corpus of novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7853-7872. Toronto, Canada, July. Association for Computational Linguistics.

Joos, M. (1972). Semantic axiom number one. *Language*, 48, pp. 257-265.

Kang, X., Zhao, Y., Zhang, J., and Zong, C. (2020). Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pp. 2242-2254, Online, November. Association for Computational Linguistics.

Khandelwal, U., Fan, A., Jurafsky, D., Zettlemoyer, L., and Lewis, M. (2021). Nearest neighbor machine translation. *arXiv preprint* arXiv:2010.00710.

Kim, D., Baek, Y., Yang, S., and Choo, J. (2023). Towards formality-aware neural machine translation by leveraging context information. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7384-7392, Singapore, December. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pp. 79-86. Phuket, Thailand, September. Asia-Pacific Association for Machine Translation.

Kudo, K., Deguchi, H., Morishita, M., Fujii, R., Ito, T., Ozaki, S., Natsumi, K., Sato, K., Yano, K., Takahashi, R., Kimura, S., Hara, T., Sakai, Y., and Suzuki, J. (2024). Document-level translation with LLM reranking: Team-J at WMT 2024 general translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pp. 210-226. Miami, Florida, USA, November. Association for Computational Linguistics.

Leech, G. (1983). *Principles of pragmatics*. London, England; New York, USA: Longman.

Levinson, S. C. (1983). *Pragmatics*. Cambridge, UK: Cambridge University Press.

Lindley, D. (1964). The editor's problem. In W. Arrowsmith and R. Shattuck (Eds.), *The Craft and Context of Translation: A Critical Symposium*, pp. 158-163, Garden City, NY. Anchor Books/Doubleday.

Lindstrom, L. (1992). Context contests: Debatable truth statements on Tanna (Vanuatu). In R. Finkbeiner, J. Meibauer, and P. B. Schumacher (Eds.), *What is a context? Linguistic approaches and challenges*. Amsterdam, Netherlands: John Benjamins.

Liu, S., and Zhang, X. (2020). Corpora for document-level neural machine translation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 3775-3781. Marseille, France, May. European Language Resources Association (ELRA).

Lyons, J. (1968). *Introduction to theoretical linguistics*. London, UK: Cambridge University Press.

Lyons, J. (1977). *Semantics* (Vols. 1 & 2). Cambridge, England: Cambridge University Press.

Lyons, J. (1995). *Linguistic Semantics: An Introduction*. London, UK: Cambridge University Press.

Malinowski, B. (1923). The problem of meaning in primitive languages. In C. K. Ogden and I. A. Richards (Eds.), *The meaning of meaning*. London, UK: K. Paul, Trench, Trubner.

Maruf, S., Martins, A. F. T., and Haffari, G. (2019). Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Volume 1 (Long and Short Papers)*, pp. 3092-3102, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Maruf, S., Saleh, F., and Haffari, G. (2021). A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys*, 54(2), pp. 1-36.

Mey, J. L. (1993). *Pragmatics: An introduction*. Malden, Mass: Blackwell.

Miller, C., Tzoukermann, E., Doyon, J., and Mallard, E. (2021). Corpus creation and evaluation for speech-to-text and speech translation. In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pp. 44-53. Virtual, August. European Association for Machine Translation (EAMT).

Nadejde, M., Currey, A., Hsu, B., Niu, X., Federico, M., and Dinu, G. (2022). CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 616-632. Seattle, United States, July. Association for Computational Linguistics.

Nakhlé, M., Dinarelli, M., Qader, R., Esperança-Rodier, E., and Blanchon, H. (2025). DOLFIN – Document-level financial test set for machine translation. *arXiv*. https://arxiv.org/abs/2502.03053

Newmark, P. (1981). *Approaches to translation*. Oxford: Pergamon.

Nida, E. A. (1964). *Toward a science of translating: With special reference to principles and procedures involved in Bible translating*. Leiden, Netherlands: E. J. Brill.

Nida, E. A. (1969). *The theory and practice of translation*. Leiden: E.J. Brill.

Nida, E. A. (2001). *Contexts in translating*. Amsterdam, Netherlands: John Benjamins.

Oaks, D. D. (2010). *Structural ambiguity in English: An applied grammatical inventory*. London, UK: Continuum.

Ochs, E. (1979). Introduction: What child language can contribute to pragmatics. In E. Ochs and B. B. Schieffelin (Eds.),

Developmental Pragmatics. New York, USA: Academic Press.

Ott, N., Ziai, R., and Meurers, D. (2012). Creation and analysis of a reading comprehension exercise corpus: Towards evaluating meaning in context. In T. Schmidt et al. (Eds.), *Multilingual corpora and multilingual corpus analysis* (Hamburg Studies on Multilingualism, Book 14), pp. 47-69. Amsterdam, Netherlands: John Benjamins.

Papineni, K., Roukos, S., Ward, T., and 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. Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Partee, B. H. (1970). Opacity, coreference, and pronouns. *Synthese, 21*(3-4), pp. 359-385. https://doi.org/10.1007/BF00484805

Pięta, H., Bueno Maia, R. and Torres-Simón, E. (2023). *Indirect translation explained*. London: Routledge.

Poyatos, F. (1997). *Nonverbal communication and translation: New perspectives and challenges in literature, interpretation, and the media*. Amsterdam: John Benjamins.

Pym, A. (2014). *Exploring translation theories*. London: Routledge.

Rauf, S. A., and Yvon, F. (2020). Document-level contexts for neural machine translation. [Research Report] 2020-003, LIMSI-CNRS. https://hal.science/hal-03687190v2

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685-2702. Online, November. Association for Computational Linguistics.

Richards, J. C., and Schmidt, R. (2010). *Longman dictionary of language teaching and applied linguistics* (4th ed.). Harlow, England; Tokyo, Japan: Longman.

Rivera-Trigueros, I., and Olvera-Lobo, M.-D. (2021). Building a corpus for corporate websites machine-translation evaluation: A step-by-step methodological approach. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pp. 93-101. Online: INCOMA Ltd.

Ryan, N. S., Pascoe, J., and Morse, D. R. (1999). Enhanced reality fieldwork: The context-aware archaeological assistant. In L. Dingwall, S. Exon, V. Gaffney, S. Laflin, and M. van Leusen (Eds.), Archaeology in the Age of the Internet: CAA97. *Computer Applications and Quantitative Methods in Archaeology. Proceedings of the 25th Anniversary Conference, University of Birmingham, April 1997 (BAR International Series 750)*, pp. 269-274, Oxford, UK. Archaeopress.

Saluja, A., Lane, I., and Zhang, Y. (2011). Context-aware language modeling for conversational speech translation. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China, September.

Scarton, C., and Specia, L. (2016). A reading-comprehension corpus for machine-translation evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 16)*, pp. 3652-3658. Portorož, Slovenia, May. European Language Resources Association (ELRA).

Sharma, S., Dey, M., and Sinha, K. (2022). How sensitive are translation systems to extra contexts? Mitigating gender bias in neural machine translation models through relevant contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1968-1984, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Shimazu, S., Takase, S., Nakazawa, T., and Okazaki, N. (2020). Evaluation dataset for zero pronoun in Japanese-to-English translation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 3630-3634. Marseille, France.

Siripragada, S., Philip, J., Namboodiri, V. P., & Jawahar, C. V. (2020). A multilingual parallel-corpora collection effort for Indian languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 3743-3751. Marseille, France, May. European Language Resources Association (ELRA).

Sperber, D., and Wilson, D. (1986). *Relevance: Communication and cognition*. Oxford, UK: Blackwell.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy, May. European Language Resources Association (ELRA).

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior, 18*(6), pp. 645-659. https://doi.org/10.1016/S0022-5371(79)90355-4

Tan, X., Zhang, L., Xiong, D., and Zhou, G. (2019). Hierarchical modeling of global context for document-level neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1576-1585. Hong Kong, China, November. Association for Computational Linguistics.

Tiedemann, J. and Scherrer, Y. (2017). Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pp. 82-92,

Copenhagen, Denmark, September. Association for Computational Linguistics.

van Dijk, T. A. (1977). *Text and context: Explorations in the semantics and pragmatics of discourse*. London, UK; New York, USA: Longman.

van Dijk, T. A. (2006). Discourse, context and cognition. *Discourse Studies*, 8 (1):159-177.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, *Vol. 30*, pp. 5998-6008, Long Beach, CA, USA, December. Curran Associates, Inc.

Vincent, S., Flynn, R., and Scarton, C. (2023). MTCue: learning zero-shot control of extra-textual attributes by leveraging unstructured context in neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8210-8226, Toronto, Canada, July. Association for Computational Linguistics.

Vincent, S., Prescott, C., Bayliss, C., Oakley, C., and Scarton, C. (2024). A case study on contextual machine translation in a professional scenario of subtitling. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pp. 561-572, Sheffield, UK, June. European Association for Machine Translation (EAMT).

Voita, E., Serdyukov, P., Sennrich, R., and Titov, I. (2018). Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), Volume 1: Long Papers*, pp. 1264-1274, Melbourne, Australia, July. Association for Computational Linguistics.

Voita, E., Sennrich, R., and Titov, I. (2019). When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pp. 1198-1212, Florence, Italy, July. Association for Computational Linguistics.

Wang, D., Fan, K., Chen, B., and Xiong, D. (2022). Efficient cluster-based k-nearest-neighbor machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022), Volume 1: Long Papers*, pp. 2175-2187, Dublin, Ireland, May. Association for Computational Linguistics.

Wang, S., Zhang, G., Wu, H., Loakman, T., Huang, W., and Lin, C. (2024). MMTE: Corpus and metrics for evaluating machine-translation quality of metaphorical language. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11343-11358. Miami, Florida, USA, November. Association for Computational Linguistics.

Wasow, T., Perfors, A., and Beaver, D. (2005). The puzzle of ambiguity. In C. O. Orgun & P. Sells (Eds.), *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe*, pp. 265-282. Stanford, CA: CSLI Publications.

Widdowson, H. G. (2004). *Text, context, pretext: Critical issues in discourse analysis*. Oxford, UK: Blackwell.

Winter, W. (1964). Impossibilities of translation. In W. Arrowsmith and R. Shattuck (Eds.), *The Craft and Context of Translation, A Critical Symposium*, pp. 93-112, Garden City, NY. Anchor Books/Doubleday.

Wu, H., Dong, D., Hu, X., Yu, D., He, W., Wu, H., Wang, H., and Liu, T. (2014). Improve statistical machine translation with context-sensitive bilingual semantic embedding model. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 142-146, Doha, Qatar, October. Association for Computational Linguistics.

Xu, M., Wang, L., Wong, D. F., Liu, H., Song, L., Chao, L. S., Shi, S., and Tu, Z. (2022). GuoFeng: A benchmark for zero-pronoun recovery and translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11266-11278. Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Yao, B., Jiang, M., Bobinac, T., Yang, D., and Hu, J. (2024). Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13078-13096, Miami, Florida, USA, November. Association for Computational Linguistics.

Zeng, J., Su, J., Wen, H., Liu, Y., Xie, J., Yin, Y., and Zhao, J. (2018). Multi-domain neural machine translation with word-level domain context discrimination. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pp. 447-457, Brussels, Belgium, October–November. Association for Computational Linguistics.

Zhang, Y. and Bach, N. (2009). Virtual babel: towards context-aware machine translation in virtual worlds. In *Proceedings of Machine Translation Summit XII: Posters*, Ottawa, Canada, August 26-30.

Zhang, L., Zhang, Z., Chen, B., Luo, W., and Si, L. (2022). Context-adaptive document-level neural machine translation. In *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6232-6236. Singapore, May. IEEE.

Zhang, L., Zhang, T., Zhang, H., Yang, B., Ye, W., and Zhang, S. (2021). Multi-hop transformer for document-level machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies (NAACL-HLT 2021), pp. 3953-3963, Online, June. Association for Computational Linguistics.

Zwicky, A. M., and Sadock, J. M. (1975). Ambiguity tests and how to fail them. In J. P. Kimball (Ed.), *Syntax and semantics*, *Vol. 4*, pp. 1-36. New York, NY: Academic Press.

Appendix

A: Context Category Descriptions and Examples

This appendix provides detailed descriptions of the eight context categories defined in Section 3. For each category, we provide a representative example group with the source text (ST) and a reference translation (TT), and, where helpful, we add a brief literal gloss in parentheses.

A1 Message: Message as context, as defined in Section 3.2, concerns information that is internal to the message itself and that is necessary to interpret and translate the FOCUS within the same communicative message. It covers both what is encoded in the message and how it is packaged, including the code and symbolic resources used to express meaning, the discourse content made available within the ongoing message, relevant co-textual elements that accompany the FOCUS and constrain its interpretation, and structural relations that organize the message and connect the FOCUS to other parts of the text, for example, coherence relations and requirements for consistent rendering across mentions. As illustrated below, the word “bank” is underdetermined in isolation, but the surrounding text clarifies whether it refers to a financial institution or a riverbank, leading to different translations.

Example Group (Excerpt)

ST:

1-0. *He went to the bank*.

1-1. *He went to the bank* and changed his money.

1-2. *He went to the bank* and took a walk along the river.

TT:

1-0. *彼は銀行に行きました*。(*He went to the bank*.)

1-1. *彼は銀行に行って*両替をしました。(*He went to the bank* and exchanged money.)

1-2. *彼は川岸に行って*散歩をしました。(*He went to the riverbank* and took a walk.)

A2 Language Events: Language Events as context, as defined in Section 3.2, concern the situational context in which an utterance is produced and interpreted, namely the communicative event and its relevant features that constrain what the speaker is doing with the utterance and how it should be understood. This includes the activity in which the speech is embedded and the event-level frame that provides resources for interpretation, such as what type of interaction it is, what the participants are doing, and what counts as an appropriate contribution in that situation. As illustrated below, the same surface expression “Cheers!” receives different translations depending on the language event, for example, a toast in a drinking situation versus a conventional response in everyday exchange, and recognizing the event is necessary to choose an appropriate translation.

Example Group (Excerpt)

ST:

2-0. *Cheers*!

2-1. A: "It has been a long time since I had a beer."

B: "*Cheers*!"

A: "*Cheers*!"

2-2. A: "I will send you the photos later."

B: "*Cheers*!"

2-3. A: "Oh thank you for helping me!"

B: "*Cheers*!"

TT:

2-0. *乾杯*！(*Cheers*!)

2-1. A：「久しぶりにビールを飲みます。」

B：「*乾杯*！」

A：「*乾杯*！」(A: “It’s been a long time since I had a beer.” / B: “*Cheers*!” / A: “*Cheers*!”)

2-2. A:「後で写真を送りますね。」

B:「*ありがとうございます*！」(A: “I’ll send you the photos later.” / B: “*Thank you*!”)

2-3. A:「ああ、手伝ってくれてありがとうございます！」

B:「*とんでもないです*！」(A: “Ah, thank you for helping me!” / B: “*You’re welcome*.”)

A3 Genre: Genre as context, as defined in Section 3.2, concerns the style or category of the message, that is, the type of communicative activity in which an utterance is produced, in which genre expectations are necessary to interpret the FOCUS and choose an appropriate translation. Genre cues can come from lexical and structural patterns typical of particular text types, such as news reporting, official statements, or narrative writing, and they guide how information should be framed, what register is appropriate, and what wording sounds natural in the target language. As illustrated below, the same core proposition is expressed in different genres, and

the translations differ accordingly in register, phrasing, and discourse style.

Example Group (Excerpt)

ST:

82-0. *The person was later found deceased*.

82-1. Following an extensive investigation, it was concluded that *the person was later found deceased* due to hypothermia, with no evidence of foul play.

82-2. According to the police, *the person* went missing from her home on the morning of Dec. 20, 2024, and *was later found deceased*.

82-3. They clung to hope as the hours dragged on, but by dawn, their worst fears were confirmed: *the person was later found deceased*, lying silently beneath the snow.

TT:

82-0. *その人はのちに死亡しているのが確認されました*。(*It was later confirmed that the person had died*.)

82-1. 大規模な捜査の結果によりますと、*その人は*低体温症によって*死亡していたとのちに判明し*、事件性はないとされました。(According to the results of a large-scale investigation, *it was later found that the person had died* of hypothermia, and the case was deemed not to involve any foul play.)

82-2. 警察によりますと、*その人は*2024 年 12 月 20 日の朝、自宅から行方不明となり、*その後死亡しているのが確認されました*。(According to the police, *the person* went missing from home on the morning of December 20, 2024, *and was later confirmed dead*.)

82-3. 時間が経過する中、彼らは希望に縋りついていた。しかし夜明けには、彼らにとって最悪な絶望が現実のものとなった。*その人は遺体で発見されたのである*。雪の下で、静かに横たわっていた。(As time passed, they clung to hope. But by dawn, their worst despair had become reality. *The person was found dead*, lying quietly beneath the snow.)

B1 Participants: Participants as context, as defined in Section 3.3, concern information about the sender, the receiver, and other relevant parties involved in a communicative event, in cases where such information is necessary to interpret the FOCUS and choose an appropriate translation. Other relevant parties may include additional addressees in group communication, overhearers or bystanders present in the situation, and, in translation, the translator as an active participant. This category includes participant roles and relationships, as well as participant-related factors that shape meaning and form, such as communicative actions (verbal and non-verbal), interactional skills, stances and attitudes, and knowledge assumed to be shared between participants. It also includes the receiver's assumed knowledge and level of expertise, for instance whether the message is addressed to a child or an adult, or to a lay audience or a specialist audience, since these assumptions affect how information should be framed and expressed. As illustrated below, the same English expression "Thank you" can function as a polite response, a sincere expression of appreciation, or an interactionally marked utterance, and the appropriate rendering varies accordingly.

Example Group (Excerpt[11])

ST:

47-0. *Thank you*.

47-2. "Then why do you never come near me?" she said. "Or kiss me, or talk to me. Why do you think I can bear it, what reason have you for treating me like that? Have you any reason?"

"Yes," I said, "I have a reason," and added very softly, "My God."

…

"*Thank you* for clarifying," she said, raising her eyebrows.

47-3. "I'm glad to hear that! You did an excellent job, and your hard work made a big difference."

"I appreciate that. It means a lot to have your guidance. *Thank you* again!"

47-5. "Dad, Grandma's on the phone!" Gary ambled across the yard.

"Dad!"

"*Thank you*, Aaron, I heard you the first time."

TT:

47-0. *ありがとうございます*。(*Thank you*.)

47-2. 「じゃあ、どうして私の近くに来ないの？」と彼女は言った。「キスもしないし、話しかけてもくれない。どうして私が耐えられると思うの？どうして私をあんな風に扱うの？何か理由があるの？」

「ええ」と私は答えた。「理由があるの」そして、とても優しく「マイゴッド」と付け加えた。

. . .

「はっきり言ってくれて*ありがとう*」と彼女は眉を上げて言った。("Then why don't you come near me?" she said. "You don't even kiss

[11] The example numbers correspond to the original corpus IDs and may therefore be non-consecutive.

me, and you won't even talk to me. Why do you think I can stand it? Why do you treat me like that? Is there some reason?" "Yes," I answered. "There is a reason," and I added very gently, "my God." … "*Thank you* for saying it plainly," she said, raising her eyebrows.)

47-3. 「それは嬉しいです！本当に素晴らしいお仕事でした。その努力が大きな違いを生みました。」

「ありがとうございます。ご指導いただけて本当に助かります。*ありがとうございます*！」("That's wonderful! You did a truly great job. Your effort made a big difference." "Thank you. Your guidance really helped me. *Thank you*!")

47-5. 「お父さん、おばあちゃんから電話だよ！」ゲイリーは庭をのんびりと横切った。

「お父さん！」

「*わかってるよ*、アーロン。一回目で聞こえたよ。」("Dad, Grandma's on the phone!" Gary strolled across the yard. / "Dad!" / "*I know*, Aaron. I heard you the first time.")

B2 Channel: Channel as context, as defined in Section 3.3, concerns the medium through which a message is delivered and the communicative conventions associated with that medium. It also includes mode-related features such as spoken versus written communication and how the text is presented in the communicative event, in cases where such channel features are necessary to interpret the FOCUS and choose an appropriate translation (Berlo, 1960; Halliday, 1985; Halliday and Hasan, 1976). As illustrated below, the channel cues whether the translation should read like a brief instruction, a legal clause, a system guide entry, or a formal customer notice, and this affects lexical choices and politeness/register in the translation.

Example Group

ST:

83-0. *To cancel, please contact us one month in advance*.

83-1. Termination of this agreement requires prior notice. *To cancel, please contact us one month in advance*.

83-2. Top > Appointment Booking System Usage Guide

2023/12/18

To cancel, please contact us one month in advance.

83-3. Dear Customer,

Thank you for choosing our service. *To cancel, please contact us one month in advance*. Let us know if we can assist any further.

Best regards,

TT:

83-0. *キャンセルされる場合は、1ヶ月前までにご連絡ください。*(*If you wish to cancel, please contact us at least one month in advance*.)

83-1. 本契約の解約には事前のご連絡が必要となります。*ご希望の場合は、1 ヶ月前までにお知らせください。*(Termination of this contract requires prior notice. *If you wish to terminate, please inform us at least one month in advance*.)

83-2. トップ > 予約システム利用ガイド

2023/12/18

予約をキャンセルされる場合は、1 ヶ月前までにご連絡ください。

(Top > Reservation System User Guide / 2023/12/18 / *If you wish to cancel your reservation, please contact us at least one month in advance*.)

83-3. お客様各位

この度は当社のサービスをご利用いただきありがとうございます。*本サービスの解約をご希望の場合は、1 ヶ月前までにご連絡ください。*他にご不明点がございましたら、お気軽にお問い合わせください。(Dear Customers, / Thank you for using our service. *If you wish to cancel this service, please contact us at least one month in advance*. If you have any other questions, please feel free to contact us.)

B3 Social-cultural System: Social-cultural System as context, as defined in Section 3.3, concerns the shared social and cultural background that makes an utterance interpretable in a particular community. This includes culturally shared beliefs and values, expectations about acceptable behaviour, and community-level presuppositions that are taken for granted in communication (Malinowski, 1923; Berlo, 1960; Lyons, 1968; Halliday, 1985; Duranti and Goodwin, 1992; Lindstrom, 1992). Importantly, this involves not only the source-language culture that motivates the meaning of an expression, but also the target-language culture that provides conventionalized ways of expressing that meaning (Givón, 1984; Crystal, 2003; Bach, 2005). As illustrated below, a culture-bound idiom, for example, can be translated using an established target-language idiom or a paraphrase.

Example Group

ST:

102-0. *Don't count your chickens before they hatch*.

102-1. Bonuses aren't final yet. *Don't count your chickens before they hatch*.

102-2. The manager warned: “Let’s *not count our chickens before they hatch*. There’s still work left to get the deal over the line.”

102-3. You may get the job, but *don’t count your chickens before they hatch*; wait for the offer letter before you throw a party.

TT:

102-0. *捕らぬ狸の皮算用はしてはいけません。* (*Don’t count on profits from a raccoon dog you haven’t caught yet*, don’t count your chickens before they hatch.)

102-1. ボーナスはまだ確定していません。*喜ぶのはまだ早いです。* (The bonus isn’t confirmed yet. *It’s too early to celebrate*.)

102-2. マネージャーは警告しました。「*捕らぬ狸の皮算用はやめましょう*。取引を成立させるには、まだやるべきことが残っています。(The manager warned, “Let’s *stop counting on profits from a raccoon dog before they hatch*. There’s still work to do to close the deal.”)

102-3. 仕事は決まるかもしれませんが、*まだ確定したわけではありません*。パーティーを開く前に、まずは内定通知が来るのを待ちましょう。(You might get the job, but *it’s not confirmed yet*. Before throwing a party, wait for the offer letter first.)

B4 Setting and Scene: Setting and Scene as context, as defined in Section 3.3, concerns situational and environmental information that accompanies an utterance and is relevant to its interpretation, such as where and when an utterance occurs, what objects or events are present, and what non-verbal circumstances are ongoing, in cases where such information is necessary to interpret the FOCUS and choose an appropriate translation. As illustrated below, the surrounding setting (platform signage, police tape, a protest placard) determines what “the line” refers to and therefore affects the most natural Japanese rendering.

Example Group

ST:

119-0. *Do not cross the line*.

119-1. The yellow edge stencil on the platform reads: “*Do not cross the line*.”

119-2. Blue-and-white police tape reads: “*Do not cross the line*.”

119-3. A union steward held a placard: “*Do not cross the line*.”

TT:

119-0. *線を越えないで*。(*Don’t cross the line*.)

119-1. ホームの黄色い縁には「*線の内側までお下がりください*」と書かれています。(On the yellow edge of the platform it says, “*Please stand back behind the line*.”)

119-2. 青と白の警察用テープには「*立入禁止*」と書かれています。(On the blue-and-white police tape it says, “*No entry*.”)

119-3. 労働組合の担当者は「*このラインを越えないで*」と書かれたプラカードを掲げた。(A union representative held up a placard that said, “*Don’t cross this line*.”)

B5 Truth Value: Truth Value as context, as defined in Section 3.3, concerns whether a proposition is to be interpreted as true or false in the relevant discourse world, and what information is needed to establish that interpretation. Following Schumacher (2012), we treat truth value not as an abstract property of an utterance, but as the situational or factual conditions that make the utterance true or false, which can in turn affect the appropriate translation. B5 is related to B4 Setting and Scene, but B4 concerns the physical setting or ongoing scene, whereas B5 concerns facticity and world knowledge that determine the intended truth-conditions of the message. As illustrated below, the word “oldest” can be misconstrued as “oldest in history”; however, even when historical information is mentioned, world knowledge about what would make the statement true suggests that the intended meaning is better captured as “oldest in population”.

Example Group

ST:

122-0. *Japan is the world's oldest country*.

122-1. *Japan is the world's oldest country*, with more than 30 percent being over 60 years old.

122-2. *Japan is the world's oldest country*, with its origin going back to 660 BC.

TT:

122-0. *日本は世界で最も高齢化が進んでいる国です*。(*Japan is the country in which aging has progressed the most in the world*.)

122-1. *日本は世界で最も高齢化が進んでいる国であり*、全人口のうち30パーセントを超える人が 60 歳以上です。(*Japan is the country in which aging has progressed the most in the world*, and among the total population more than 30 percent of people are 60 years old or above.)

122-2. *日本は世界で最も高齢化が進んでいる国であり*、その起源は紀元前 660 年に遡ります。(*Japan is the country in which aging has progressed the most in the world*, and its origin goes back to 660 BC.)

B: Data Format (JSON)

This appendix describes the released JSON format of Context-8. The dataset is provided as a single JSON file with three top-level fields: schema_version, category_map, and items.

category_map specifies the mapping from context-category codes to category names. items contains a list of examples. Each example includes the following fields: group_id (group identifier), sid (example identifier within group), st (English source text), ref (Japanese reference translation), context_impact (whether the context mainly affects accuracy or fluency), and context_categories (a non-exclusive list of assigned context-category codes). Labels are non-exclusive, so one example may appear under multiple categories.

JSON example (one item):


```
{
    "group_id": 1,
    "sid": "1-0",
    "st": "He went to the bank.",
    "ref": "彼は銀行に行きました。",
    "context_impact": "accuracy",
    "context_categories": [
      "A1",
      "A2"
    ]
}
```


C: FOCUS-level Scoring Procedure

In addition to example-level evaluation, we also report scores at the FOCUS level. For each example, we extract the target-language fragment in the system output that corresponds to the FOCUS in the source. We first obtain an automatic word alignment between the source sentence and the system output using SimAlign (Sabet et al., 2020), and then derive a minimal contiguous target span aligned to the FOCUS. Because automatic alignment can be imperfect, we manually correct the extracted target fragments when necessary, primarily to adjust span boundaries (e.g., inclusion or exclusion of function words and punctuation) while preserving the system output. These extracted fragments are used to compute FOCUS-level COMET scores.

D: LLM Prompt and Inference Settings

This appendix documents the prompt used for generating LLM translations in our experiments, to support reproducibility.

Model: GPT-5.2 (accessed on 2026-03-04).

System prompt:

None (the system prompt was left empty).

User prompt:

“You are a machine translation engine. Translate from English to Japanese. Do not add or omit information. Preserve meaning exactly. Output Japanese only.”

Output constraints: The model was instructed to output only the Japanese translation, without additional explanations or formatting.

Inference settings: We used the default API settings for this model. When an explicit temperature parameter was not supported, we used the model’s default settings.