# BOUQuET 💐: dataset, Benchmark and Open initiative for Universal Quality Evaluation in Translation

**The Omnilingual MT Team, Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Mark Dupenthaler, Nate Ekberg, Cynthia Gao, Daniel Licht, Jean Maillard, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenko, Albert Ventayol-Boada, Shireen Yates**

FAIR, Meta; University of London (UCL); University of the Basque Country (UPV/EHU)

costajussa@meta.com

## Abstract

BOUQuET is a multi-way, multicentric and multi-register/domain dataset and benchmark, and a broader collaborative initiative. This dataset is handcrafted in 8 non-English languages (i.e. Egyptian Arabic and Modern Standard Arabic, French, German, Hindi, Indonesian, Mandarin Chinese, Russian, and Spanish). Each of these source languages are representative of the most widely spoken ones and therefore they have the potential to serve as pivot languages that will enable more accurate translations. The dataset is multicentric to enforce representation of multilingual language features. In addition, the dataset goes beyond the sentence level, as it is organized in paragraphs of various lengths. Compared with related machine translation datasets, we show that BOUQuET has a broader representation of domains while simplifying the translation task for non-experts. Therefore, BOUQuET is specially suitable for crowd-source extension for which we are launching a call aiming at collecting a multi-way parallel corpus covering any written language. The dataset is freely available at https://huggingface.co/datasets/facebook/bouquet.

## 1 Introduction

Although multilingual large language model (LLM) evaluation benchmarks are only starting (Dac Lai et al., 2023), there is a rich research history in multilingual evaluation datasets for natural language processing; e.g., (Sun and Duh, 2020; Malmasi et al., 2022; Yu et al., 2022), with Machine Translation (MT) being the task with the highest investment in multilinguality (Kocmi et al., 2024). This is evident from the nearly 20-year history of the Conference on Machine Translation (formerly a workshop, WMT), which has established an international evaluation campaign (Kocmi et al., 2024). The campaign has compiled a comprehensive collection of parallel corpus evaluations covering a broad range of language pairs, domains, tasks and recently, investing in a multi-way parallel dataset expanding in languages (Deutsch et al., 2025). However, the largest multi-way parallel evaluation dataset to date was introduced with FLORES-101 (Goyal et al., 2022), later expanded to FLORES-200 (NLLBTeam, 2024), FLORES+[1] and to 2M-FLORES (Costa-jussà et al., 2024).

These existing datasets and benchmarks fall short due to having an English-centric focus, a narrow selection of registers, compromised quality from automated construction and mining, limited language coverage, or a static nature, in addition to being prone to contamination (Sainz et al., 2023). Similarly, in parallel with the previous progress, there have been several initiatives that called for data annotation in a collaborative and open way, such as the translation data collection initiative (Singh et al., 2024).

Recently, Wu et al. (2025) evaluate multilingual benchmarking and make a call for action for the need for accurate, contamination-free, challenging, practically relevant, linguistically diverse, and culturally authentic evaluations. This call and the urgent need of progressing in multilingual benchmarking set the stage for the introduction of a new multilingual multi-way parallel MT evaluation dataset and benchmark. BOUQuET, which additionally combines community efforts, relies on text written from scratch (contamination-free[2]) by native speakers in 8 different major languages (linguistically diverse). Text includes a variety of 8 practical domains (practically relevant) that represent localised knowledge (culturally diverse). BOUQuET is aligned at the sentence and paragraph-level and it relies on a mixture of com-

[1] https://oldi.org/

[2] Note that BOUQuET is free from contamination in each initial state because it is originally created and not mined. However, from the moment we open-source certain splits, BOUQuET will risk to leak into training. Therefore, we keep one split hidden to avoid this.

*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27515–27535
November 4-9, 2025 ©2025 Association for Computational Linguistics

missioned and openly collected human annotations to extend to any language.

The organisation of the paper is as follows. First, the paper details how we develop the Source-BOUQuET dataset (Section 3), which is the necessary stepping stone towards an open initiative. Second, we benchmark BOUQuET for the 8 pivot languages plus English (Section 4). Finally, Section 5 presents how we design the open initiative itself, which aims to build the Full-BOUQuET dataset; i.e., Source-BOUQuET translated into any written language. At the time of publication of this paper, BOUQuET includes 55 multi-way parallel completed languages (Table 6).

## 2 Definitions and background

**Definitions** Before describing the Source-BOUQuET dataset's characteristics and building methodology, we define our use of some frequently encountered terms that may cover a variety of meanings.

**Domain** By the term *domain*, we mean different spaces in which language is produced in speech, sign, or writing (e.g., books, social media, news, Wikipedia, organization websites, official documents, direct messaging, texting). In this paper, we focus solely on the written modality.

**Register** We understand the term *register* as a functional variety of language that includes socio-semiotic properties, as expressed in Halliday and Matthiessen (2004), or more simply as a "contextual style," as presented in Labov (1991, pp.79–99). In that regard, a register is a specific variety of language used to best fit a specific communicative purpose in a specific situation.

**Background** There is a large body of work in creating datasets for MT evaluation (e.g. WMT International Evaluation Campaigns (Deutsch et al., 2025)). However, the vast majority are limited in languages. we next discuss the main efforts to build massively multilingual MT benchmarks and one representation of multi-domain dataset.

**FLORES+** FLORES+ (Maillard et al., 2024) is the largest multilingual extension of FLORES-200 (Goyal et al., 2022) and it covers the largest multi-way parallel dataset in terms of languages in 3 domains (Wikipedia, News, Travel guides). Even if FLORES+ has paragraph information, the translation has been done at the level of sentence without showing context to the annotators.

**NTREX-128** Similarly to FLORES+ NTREX-128 covers a multi-way parallel dataset but for 128 languages. Unlike FLORES-200, translators had the full context of the document available when translating sentences, but the authors did not know if (or to what extent) they used this information (Federmann et al., 2022).

**NLLB-MD** was motivated to complement FLORES-200 in terms of domains in the context of the NLLB (NLLBTeam, 2024) project. It covers chat, news and health domains in 6 languages and it includes a much larger number of sentences.

All these datasets are English-localised and English-centric, meaning that all languages have been translated from the original source English. They cover limited amount of domains (a maximum of 4) and do not differentiate among registers.

## 3 Dataset: Source-BOUQuET

In this section, we describe the creation criteria that have been followed to design Source-BOUQuET, as well as the languages it includes.

### 3.1 Main characteristics

As described in greater detail next, the Source-BOUQuET dataset is mainly characterized by its non-English-centric focus, its diverse range of registers and domains (which are complementary to FLORES-200), its manual and original composition, and its built-in dynamic extensibility. Table 2 provides a comparison of several relevant statistics from BOUQuET and the closest related datasets covered in the previous section.

**Non-English-centric focus** Source-BOUQuET is handcrafted by proficient speakers of Egyptian Arabic and Modern Standard Arabic (MSA),[3] French, German, Hindi, Indonesian, Mandarin Chinese, Russian, and Spanish. Each of these languages contributes the same number of sentences to the final dataset. The languages for Source-BOUQuET (see Table 1) are all part of the top 20 languages in the world in terms of user population, as listed in Eberhard et al. (2024). In addition, they are also used by a large number of non-native

[3] Speakers of Egyptian Arabic typically use it in informal contexts and switch to MSA for more formal communication. For constructing Source-BOUQuET in Egyptian Arabic, we reproduce this code switching.

speakers, which makes them good candidates for what we refer to as *pivot* languages; i.e., higher-resource languages that can facilitate—as source languages—the translation of datasets into lower-resource languages. English is often used as such a pivot language, since numerous people have a high degree of proficiency in English as a second language. English is not the only language in this situation, however, and is not always the best pivot language option. For example, it is much easier to find Guarani-Spanish bilingual speakers than it is to find Guarani-English bilingual speakers. What is more, cultural proximity may also make translation slightly easier.

| ISO 6393 | ISO 15924 | LANGUAGE | FAMILY | SUBGROUP1 |
|---|---|---|---|---|
| arb | Arab | Modern Standard Arabic | Afro-Asiatic | West Semitic |
| cmn | Hans | Mandarin Chinese | Sino-Tibetan | Sinitic |
| deu | Latn | German | Indo-European | West Germanic |
| fra | Latn | French | Indo-European | Italic |
| hin | Deva | Hindi | Indo-European | Indo-Aryan |
| ind | Latn | Indonesian | Austronesian | Malayic |
| rus | Cyrl | Russian | Indo-European | Balto-Slavic |
| spa | Latn | Spanish | Indo-European | Italic |

Table 1: Source-BOUQuET Languages

**Diverse registers and domains** Registers derive from communicative purposes and, as such, are related to domains. However, the relationship between registers and domains is not one to one. See the register and domain correspondence in Figure 5 (Appendix B). For example, if we take a domain such as TV news, we can identify at least 3 registers: (1) the register used by the news anchor, which is represented by fully scripted language that is read from a teleprompter with a very specific and unnatural form of diction (e.g., hypercorrect enunciation, unnatural intonation, homogeneous pace); (2) The register produced by communication specialists (i.e., people who have been trained to be spokespersons or surrogates). The points they make have been scripted and rehearsed to the point of being known by heart. It sounds spontaneous but it is not structured like informal language; (3) the register represented in person-in-the-street segments, which is more informal and spontaneous (possibly colloquial). This example is taken from a domain where both speech and writing are used but the situation is not significantly different in the written modality only. Language users all commonly shift between registers, which is typically referred to as style-shifting. Style-shifting (i.e., register-shifting) occurs within domains; so the domain itself is not a fool-proof way of getting a specific register. Although the norms of the domain can impose the degree of formality and of lexical specialization, it is often the register (which derives from the communicative purpose), not the domain, that determines many aspects of linguistic structure (e.g., lexical density, pronoun use, syntax, etc.).

**Manual construction and original composition (not crawled) with accurate revisions** To develop Source-BOUQuET, we set a variety of linguistic criteria that need to be covered, including both unmarked and marked structures (e.g., expected and unexpected number agreement between subject and verb). Guidelines are then shared with linguists who manually craft sentences covering examples of these linguistic criteria and compose paragraphs ranging from 3 to 6 sentences in length. These paragraphs are then manually translated across all pivot languages.

The main strategies for open collaboration are to design contribution guidelines and build an annotation tool that enables the free collection of translations in any language. BOUQuET is shared in a repository (Section 5) that allows language community to easily add a new language by translating it from one of the 8 pivot languages or the English translation.

**Language coverage extensibility** Using both private and community-driven initiatives, we could potentially support any written language, as long as there is specific interest in contributing to multilingual advancements.

**Dynamic in nature** Since BOUQuET includes the community, it can continuously evolve by constantly engaging it.

### 3.2 Creation criteria

For the design of the creation guidelines, detailed in Appendix A, we prepared a list of linguistic coverage requirements along with some statistical information.

**Linguistic coverage requirements** In order for BOUQuET to be representative of various linguistic phenomena, linguistic coverage requirements are defined (as listed in Table 3), which are to be included in sentences that form paragraphs. Sentences are assigned a unique identifier that com-

| Dataset | Split | #Parag. | #Sent | Avg. Wrd. Parag/Sent | Reg. | Dom. | Lang. | Dyn. |
|---|---|---|---|---|---|---|---|---|
| FLORES+ | Dev<br>Devtest<br>Eval | × | 997<br>1,012<br>992 | 25 | × | Wikipedia, News, Travel guides | 220 | ✓ |
| NTREX-128 | Test | 123 | 1,997 | 389/24 | × | News | 128 | × |
| NLLB-MD | Dev<br>Devtest<br>Eval | × | 6,000<br>1,310<br>1,500 | 25 | × | Chat, News, Health | 6 | × |
| BOUQuET | Dev<br>Devtest<br>Eval | 120<br>200<br>144 | 504<br>864<br>628 | 55/15 | ✓ | Fiction, Conversation, Social media posts/comments, Tutorials, Website, Reflection pieces, Miscellaneous | 55+[a] | ✓ |

[a] See Appendix D for language coverage details

Table 2: Main statistics from MT evaluation datasets including BOUQuET: number of sentences, number of paragraphs, average word per paragraph (or sentence), register information, domains, languages, dynamism.

bines a unique paragraph ID number with a serial sentence number. Thus, paragraphs can be retrieved by concatenating sentences that share the same paragraph ID.

| Phenomena |
|---|
| Paragraph-like continuity |
| Variation in sentence lengths |
| Dominant (unmarked) and non-dominant (marked) word orders |
| Different emphasis or topicalization |
| Different sentence structures (affirmation, interrogation, negation, subordination, coordination) |
| Different verb moods, tenses, and aspects |
| Different morphosyntactic options |
| Different grammatical persons (1st, 2nd, 3rd, singular, plural) |
| Different grammatical genders |
| Different grammatical number agreement |
| Different grammatical case or forms of inflection |
| Most frequent words used in various registers |
| Presence of named entities, numbers, slang, and emojis |

Table 3: Source-BOUQuET Linguistic Requirements

**Variety of domains** Source-BOUQuET is intended to cover 8 domains: narration (as in fiction writing), dialog, social media posts, social media comments, how-to manuals and instructions, miscellaneous website content (excluding social media or news), opinion pieces, and other miscellaneous (such as written speeches or signage). The choice of these domains optimizes for variety and popular usefulness and complementary to FLORES-200.

**Variety of registers** Source-BOUQuET is built with register variety in mind, differently from FLORES-200, which covers a few different domains but remains largely within similar registers. We characterize the registers through 3 main features (connectedness, preparedness, and social differential). Connectedness attempts to describe the type of interaction typically available in a given domain. Preparedness aims to gauge how much time is typically used to produce or edit language content. Social differential describes the relationship between the interlocutors involved in a given social situation (e.g., writer and reader, characters in a dialog, etc.). Each individual domain can present different combinations of features but become differentiated at the level of the sentence. There are a variety of feature combinations, which are mentioned in Figure 5 and defined in Appendix B.

By including new registers and domains, the new dataset is likely to be more generalizable to different contexts and applications.

**Statistical guidance for domain representation.** In order to appropriately cover linguistic requirements and adequately represent domains, we performed a statistical analysis to understand the linguistic characteristics of each domain before creating BOUQuET. In particular, our analysis covers most domains that we are including in Source-BOUQuET by using diverse public datasets: narration (Books3, Gutenberg library (Gerlach and Font-Clos, 2018)); Social media posts (Reddit (Baumgartner et al., 2020)); Social media comments (Wikipedia comments[4]); Conversations / Dialogues (dialogsum (Chen et al., 2021), Open Orca (Lian et al., 2023)); Tutorials/how-to articles (how-to Wikipedia-lingua [5]); Website content (C4 (Raffel et al., 2020)); News / Reflection pieces (CNN-DailyMail (Nallapati et al., 2016), XSum (Narayan et al., 2018)) and Miscellaneous (Wikipedia). Note that we collect information from public data that do not always accurately match our categories but constitute a proxy. For each of these domains, we have analyzed dimensionality: characters per token; to-

[4] https://www.kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification

[5] https://huggingface.co/datasets/GEM/wiki_lingua

kens per sentence and sentences per paragraph; and linguistic complexity with CEFR levels[6].

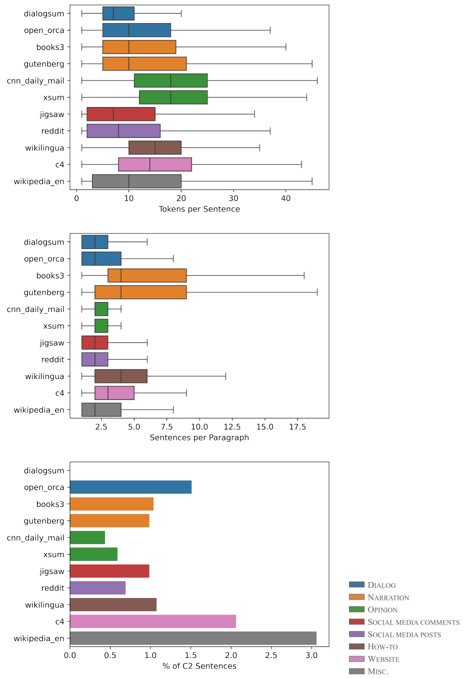


Figure 1: (Top) Tokens per sentence and (Middle) sentences per paragraph (Bottom) CEFR per dataset representative of BOUQuET domains.

Regarding tokens per sentence (Figure 1 left), we can see correlations between different domains, and clear differences in length, especially in dialogs which tend to be much shorter. Regarding sentences per paragraph (Figure 1 middle), we can find a correlation between different datasets representing the same domain, where fiction writing paragraphs tend to be much longer (averaging 5 but reaching up to 20), dialogs and news articles are much shorter (barely reaching 3-4 sentences in a paragraph), and the rest of the categories are somewhere in between (normally staying between 1-5 but reaching up to 10 in some cases).

To guide BOUQuET creators on the linguistic complexity required for each domain, we have assessed complexity using the distribution of CEFR levels as a proxy. This includes the % of C2 scores at the sentence level for each dataset (see the bottom part of Figure 1). These scores were labeled by a SONAR-based model (Duquenne et al., 2023) trained on CEFR-SP text classification dataset (Arase et al., 2022), which significantly outperformed LLAMA-3 (Touvron et al., 2023). Wikipedia seems to be the only dataset with a more considerable share of C2 sentences, with some others like dialogues having no samples scored as such.

**Annotations and Quality Checks** Each entry of Source-BOUQuET includes the source text (in one of the 8 pivot languages of Table 1) and its translation into English, domain information and contextual information for better translation accuracy. To double-check that Source-BOUQuET does not contain repeated sentences, we explored the similarity across English sentences. For each English sentence, we computed SONAR embeddings (Duquenne et al., 2023) and we computed the cosine distance on the vectors. There were only 14 pairs out of 2000 sentences with a cosine distance below 0.3 (as approximately threshold for similarity). These pairs are reported in Appendix E.

## 3.3 Languages

BOUQuET is created in 8 non-English languages (Table 1) and comprises 2,000 original sentences: 250 in each language. Although the overall structure of each section is identical (same 8 domains represented, 58 paragraphs per section, 250 sentences distributed into paragraphs in the same way), the sets of 250 sentences are not translations of each other and are not translated from an English source. For example, Paragraph 001 (the first paragraph created in Egyptian Arabic) is made up of 3 sentences originally created with the purpose of representing the style of Egyptian Arabic How-to articles; Paragraph 059 (the first paragraph created in Mandarin Chinese) is also made up of 3 sentences, which is also originally created with the purpose of representing the style of Mandarin Chinese How-to articles. However, the 3 sentences in Paragraph 001 are not the same sentences as those in Paragraph 059. They do not convey the same message. Paragraph 001 deals with a recipe for a well-known Egyptian dish, while Paragraph 059 provides instructions on how to set up a screen lock on a smart phone.

By writing 250 original sentences in each of the 8 non-English languages, we obtain a 2,000-sentence dataset that is not initially created in English and is not grounded in English-language culture. In addition, since the creators of the dataset also happen to be linguists who are fluent in English, we have asked them to produce gold-level English translations for each of the 250 sentences they have created in their respective languages. At

[6] https://www.coe.int/en/web/common-european-framework-reference-languages/

the end of this initial phase, we had 8 sections, each in a different language, of 250 sentences each, as well as English translations.

### 3.4 Multi-way extension to Source-BOUQuET languages

**Details** Source-BOUQuET creators composed 250 sentences for each of the 8 pivot languages plus the corresponding English translation. To get to 2,000 sentences in all 8 languages (= 16,000 sentences) while keeping the dataset parallel, linguists have used the English version of the 1,750 sentences that had not been created in their respective languages in order to produce the missing parallel sentences in their respective languages. At the end of this second phase, we then have a parallel dataset comprising 2,000 sentences in each of the 9 languages (8 initial languages + English), or 18,000 sentences total.

**Quality checks** Since multi-way parallel data is created from English, we manually checked that translations did not lose the linguistic information when translating from English. While translating BOUQuET, we had to make sure that the contextual information which was applicable to the whole paragraph was taken into consideration by the translators. To ensure this, we used a number of QA strategies reported in Appendix C.

**Additional contextual information** The multi-centric nature of BOUQuET is also a reminder that English is not morphologically rich (e.g., it doesn't mark grammatical gender agreement between nouns, adjectives, and verbs) and displays relatively little information about formality in its written form (e.g., it uses only one second-person singular pronoun, regardless of who is addressing whom). As such, English isn't an ideal source language for translation purposes unless translators can be provided with additional contextual information. The BOUQuET dataset includes such additional information; for example, the grammatical gender of the first and second person (when this isn't obvious) or the linguistic markedness of some words or phrases (e.g., literary or archaic verb tenses, use of slang, infrequently used level of formality).

### 3.5 Overall Statistics

In total, BOUQuET contains 2,000 sentences. These sentences are split by making a stratified paragraph-level selection among source languages and domains into development, test and evaluation sets. Initially, the evaluation set (632/144 sentences/paragraphs) is intended to be kept hidden. Figure 2 shows the representation of registers (top) and domains (bottom) in the non-hidden splits. Labels for each of the combinations of register options are created by concatenating the lowercase letters used as unique identifiers (see details of these register options in the Appendix B). For example, a register characterized as impersonal (in connectedness), composed (in preparedness), and equal-assumed (in social differential) is labeled `ica`.

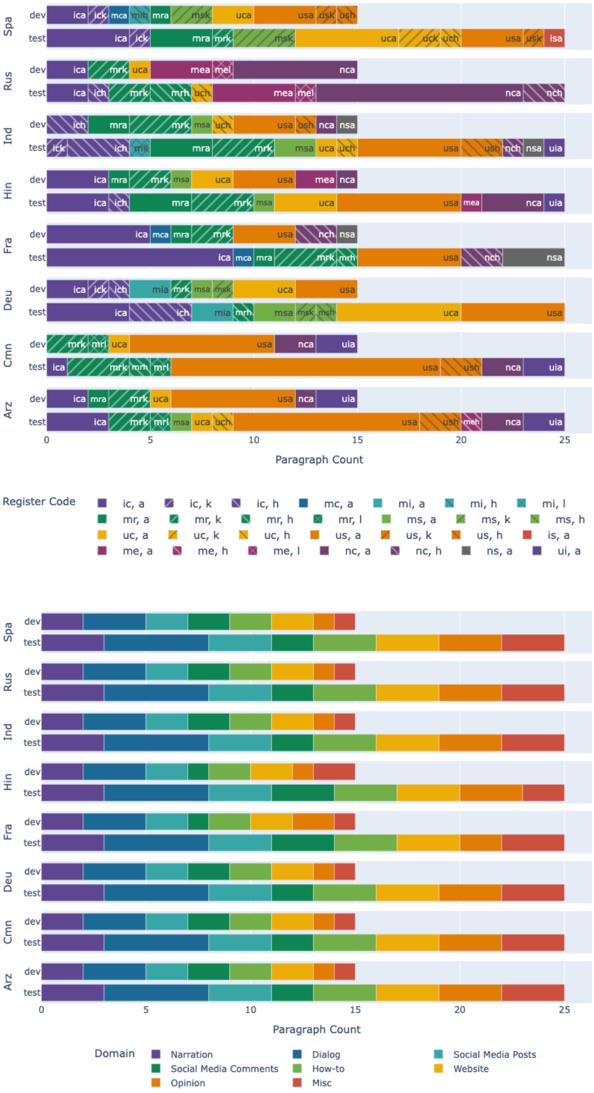


Figure 2: Registers (top) and domains (bottom) representations in development and test partitions.

The results in Section 4 are presented with the test split of 864/200 sentences/paragraphs.

### 3.6 Dataset difficulty and diversity

We intend BOUQuET to be a difficult benchmark not due to inherently difficult content (see Proietti et al. (2025) for a possible formalization), but by virtue of being extended into lower-resourced languages with which MT models struggle. With the goal of extending BOUQuET to potentially any language, we aim at making it less difficult to translate by non-experts than other massively parallel datasets. To demonstrate it, we measure the difficulty of the English side of BOUQuET and other datasets by CEFR levels.

Despite BOUQuET texts being simple, they still express a variety of linguistic phenomena. While we are not aware of generally accepted measures of linguistic diversity of corpora, we show that BOUQuET is grammatically diverse compared to most other parallel datasets, as measured by the entropy of the distribution of morphological features of words on the English side.[7]

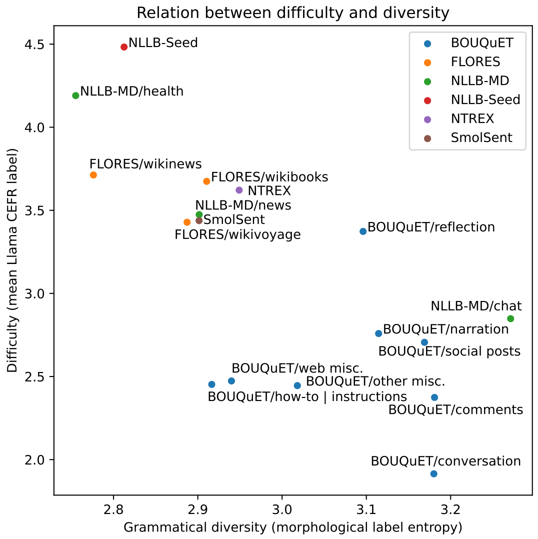


Figure 3: Mean difficulty (on the scale of 1-6: A1 to C2) and grammatical diversity (entropy of morphological labels) of the major parallel datasets and their domains.

Figure 3 illustrates the relationship between the difficulty and grammatical diversity across the major parallel datasets (split by domain, where relevant). Perhaps surprisingly, the dependency is inverse: conversational and literary subcorpora are the most grammatically diverse while being the easiest, whereas the informational texts (news and encyclopedia), while being difficult (potentially due to the density of specialized terms and named entities), grammatically are more uniform.

## 4 Benchmark

We benchmark BOUQuET in two dimensions: domain representation and machine translation. The former quantifies how representative BOUQuET is of public datasets of multiple domains compared to other evaluation datasets. The latter addresses how several MT systems are ranked with BOUQuET compared to other evaluation datasets.

**Domain representation** The performance of the model in a new or unseen dataset depends on the similarity between the dataset that was used to fit the model and the new dataset. We compare the domain coverage of BOUQuET with that of FLORES+, NTREX-128, and NLLB-MD. To do this comparison, we take a random sample of 2,000 sentences (which seems to be a sufficiently large sample of the embedding space for score stability)[8] from each of the domain datasets from Figure 1; as well as 2000 from each alternative dataset FLORES+, NTREX-128, NLLB-MD, and BOUQuET. We create vector representations of each sentence in these datasets with SONAR (Duquenne et al., 2023). We measure the overlap between each parallel dataset with each of the domains using the Wasserstein distance (implemented with the POT library[9]). The Wasserstein Distance (WD), also known as the Earth Mover's Distance (EMD), is a metric that measures the ”effort” required to transform one probability distribution into another. Lower values indicate a higher similarity between clusters. Figure 4 shows that the lowest consistent results are obtained for all domains with BOUQuET. Appendix F provides a more direct visualization of the distribution overlap, leading to the same conclusion.

**Machine Translation** To help the reader understand why the dataset is useful, we present preliminary results to demonstrate its use for its intended purpose: MT benchmarking. We evaluate 14 translation systems: LLAMA-3 (Llama3.1-8B, Llama3.2-3B, Llama3.3-70B) (Touvron et al., 2023), Tower (TowerInstruct-7B-v0.2) (Rei et al.,

[7] We annotated difficulty by prompting annotated by prompting Llama 3.3 70B, and moprhology, by a Spacy model; more details are in Appendix G.

[8] Some domain sets and evaluation sets were several orders of magnitude larger than each other. Downsampling them to the same size (2,000) makes the metric computable in a reasonable amount of time and removes any sensitivity to class imbalance in the distribution distance metric.

[9] https://pythonot.github.io/

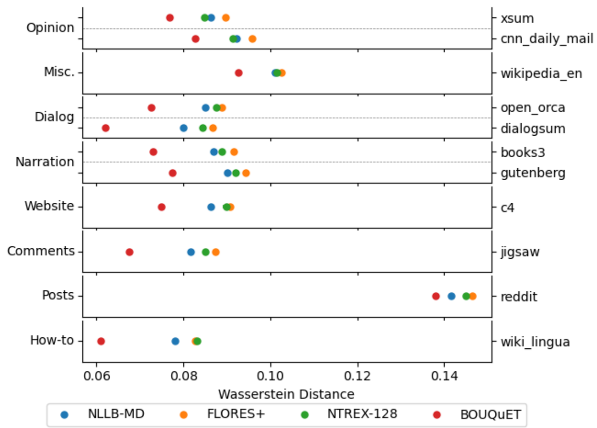


Figure 4: Wasserstein Distance (WD) for each domain and dataset. Lower WD indicate better representation of the domain.

2024), Aya (Aya101-13B, Aya-Expanse-8B) (Dang et al., 2024), Babel (Babel-9B-Chat) (Zhao et al., 2025), Cohere (CohereLabs-command-r7b-12-2024), Eurollm (EuroLLM-9B-Instruct) (Martins et al., 2024), MADLAD (MADLAD-3B-MT and MADLAD-10B-MT) (Kudugunta et al., 2023), Mistral (Mistral-7B-Instruct-v0.3)[10], Qwen (Qwen2.5-7B-Instruct) (Bai et al., 2023) and NLLB (NLLB-3.3B) (NLLBTeam, 2024). We select the models as ones with open weights, focusing primarily on moderate sizes (about 10B) and variety of architectures. Following the official evaluation metrics of WMT 2024 (Kocmi et al., 2024), we use two automatic metrics: CometKiwi (CometKiwi-da-xl, range 0-1 and ↑ better, COM) (Chimoto and Bassett, 2022) and MetricX (MetricX-24-hybrid-xl-v2p6, range 0-25 and ↓ better, MetX) (Juraska et al., 2024). We include in the benchmarking datasets that cover Source-BOUQuET languages (FLORES+ and NTREX-128).

Table 4 shows that BOUQuET scores consistently higher than other datasets on average, suggesting that BOUQuET is easier to translate. This is an advantage for the open initiative, since the complexity of current MT test sets makes it harder to ask the community to participate in translations as it requires a high-level of expertise.

Rankings across models and datasets is not preserved, which hints that all datasets may be posing different challenges to the models. Rankings is computed as counting when a system is similar in the same position according to CometKiwi. This ranking and Pearson correlation on the CometKiwi is dissimilar for datasets evaluated at the sentence-level, with BOUQuET being the most different. This difference is enlarged when evaluating at the paragraph-level where number of swaps increases and, coherently, Pearson correlation decreases, meaning that datasets pose different challenges to models. We need to further investigate which linguistic challenges BOUQuET is adding. However, best two systems are consistent across datasets and level of evaluation (sentence and paragraphs) being those the largest model (Llama3.3-70B) and Aya-e-8B.

NLLB-3.3B has the highest variation between being evaluated at the sentence or paragraph-level, which makes sense since it is the only model trained exclusively with sentence-level data.

Appendix H reports more detailed results on BOUQuET by languages and domains. We plan to release an open leaderboard[11] so that developers of MT models or multilingual LLMs will be able to compare the BOUQuET evaluation scores for their systems.

## 5 Beyond commissioning translations: Open initiative

Source-BOUQuET is intended to be translated into any written language. For this, we have commissioned an initial set of priority languages covering a variety of high and low-resource languages representing different geographical regions, linguistic families and scripts. See the list of languages currently covered by BOUQuET in Appendix D.

However, it would be challenging to achieve our language coverage target to any language. This ambition can only be achieved with the support of the community. For this, we have organized an open collaborative effort which involves language communities that are interested in contributing to this effort.

The purpose of this open initiative is to collect translations from Source-BOUQuET. To facilitate this, we have developed a publicly accessible tool that enables crowd-sourcing of translations from anyone interested in contributing. The Source-BOUQuET data has been uploaded to this tool, along with the annotation guidelines from Section 3.4, which very much resemble those from FLORES-200 (NLLBTeam, 2024) and which are

[10] https://docs.mistral.ai/getting-started/models/models_overview/

[11] https://huggingface.co/spaces/facebook/bouquet.

| Model | BOUQUET | | FLORES | | NTREX | | BOUQUETP | | NTREXP | |
|---|---|---|---|---|---|---|---|---|---|---|
| | COM | METX | COM | METX | COM | METX | COM | METX | COM | METX |
| NLLB-3B | 0.68 | 2.1 | 0.66 | 2.56 | 0.65 | 2.97 | 0.59 | 3.71 | 0.29 | 14.1 |
| aya101-13B | 0.67 | 2.02 | 0.63 | 2.65 | 0.63 | 3.14 | 0.58 | 3.29 | 0.24 | 13.17 |
| aya-e-8B | 0.69 | **1.75** | 0.65 | 2.9 | **0.67** | **2.45** | 0.64 | **2.42** | 0.34 | **8.7** |
| babel-9B | 0.67 | 2.33 | 0.65 | 2.66 | 0.63 | 3.36 | 0.61 | 3.4 | 0.32 | 10.39 |
| cohere-7B | 0.67 | 2.15 | 0.65 | 2.89 | 0.64 | 3.01 | 0.61 | 3.2 | 0.32 | 9.61 |
| eurollm-9B | 0.67 | 2.33 | 0.65 | 2.89 | 0.61 | 3.64 | 0.61 | 3.64 | 0.31 | 10.08 |
| madlad-10B | 0.63 | 2.74 | 0.64 | 2.72 | 0.63 | 3.35 | 0.41 | 6.76 | 0.15 | 15.99 |
| madlad-3B | 0.63 | 2.85 | 0.63 | 2.94 | 0.61 | 3.67 | 0.37 | 6.71 | 0.49 | 5.29 |
| mistral-7B | 0.54 | 4.29 | 0.51 | 5.69 | 0.49 | 6.12 | 0.49 | 6.64 | 0.24 | 10.96 |
| qwen-7B | 0.59 | 3.25 | 0.6 | 3.75 | 0.59 | 4.21 | 0.57 | 4.5 | 0.52 | 4.93 |
| Llama3.1-8B | 0.66 | 2.36 | 0.64 | 2.82 | 0.63 | 3.27 | 0.6 | 3.33 | 0.32 | 10.17 |
| Llama3.2-3B | 0.59 | 3.59 | 0.57 | 4.34 | 0.55 | 4.89 | 0.52 | 5.52 | 0.27 | 12.67 |
| Llama3.3-70B | **0.7** | 1.85 | **0.68** | **2.21** | **0.67** | 2.59 | **0.63** | 2.72 | **0.35** | 9.76 |
| Tower-7B | 0.58 | 3.69 | 0.56 | 4.19 | 0.56 | 4.35 | 0.49 | 5.65 | 0.28 | 12.22 |
| | BOUQUET-FLORES | | FLORES-NTREX | | NTREX-BOUQUET | | BOUQUETP-NTREXP | | | |
| Swaps | 3 | | 4 | | 7 | | 11 | | | |
| Pearson Cor. | 0.95 | | 0.99 | | 0.95 | | 0.92 | | | |

Table 4: Averaged Results XX-to-XX 9 Source-BOUQuET (8 pivot plus English) languages for BOUQuET, FLORES+, NTREX-128 at the level of sentence 2 columns on the left and at the level of paragraph 3 columns on the right. Number of ranking swaps (a system not being in the same position according to CometKiwi) from each dataset compared to the other two (in similar sentence or paragraph-level) and Pearson correlation indicate that while datasets report similar results at sentence-level, being BOUQuET the most different, it is not the case for paragraph-level where the ranking of systems varies by a larger amount.

available in the 9 BOUQuET languages. A key advantage of the tool is that annotators can select the source language from any of the Source-BOUQuET languages. These languages were carefully chosen to cover a wide range of speakers, making the process more accessible by reducing reliance on English bilingual speakers. We have also implemented a translation validation stage, making this a multi-stage crowd-sourcing pipeline, similar to the workflow supported by Mozilla Common Voice[12]. Additionally, before releasing data in new languages, we plan to introduce further quality checks, based on techniques such as cross-lingual sentence representation similarity and automatic language and script identification (more details in Appendix I). Any data subsets flagged as potentially invalid will undergo additional review by our team. These measures are designed to ensure that only high-quality data is published.

We recognize the importance of acknowledging contributors to open initiatives like BOUQuET. To support this, we encourage dataset contributors to submit papers describing their work to the Open Language Data Initiative[13] – which in the past have been published as WMT systems papers – following the example of other community-led extensions of open source datasets such as FLORES and NLLB-Seed. In addition, we are committed to establishing meaningful rewards for contributors to these important resources which benefit the entire research community. For instance, this year, we are pleased to sponsor EMNLP/WMT conference registration fees for a select group of accepted OLDI task participants. This open initiative is available at https://bouquet.metademolab.com/.

## 6 Conclusions and Next Steps

In this paper, we have presented the Source-BOUQuET dataset and the attached open initiative. We have shown consistent gains in domain diversity in two different metrics while keeping complexity lower than its competitors. The latter is particularly relevant to simplify the translation for non-experts that may join the open initiative. We also provide MT evaluation results for the 9 languages in which Source-BOUQuET has been created. Although BOUQuET is currently totally completed for 55 languages (see list 6), this number is only a fraction of the language coverage ambition that we are pursuing by launching the open initiative for community efforts. Please join us in making Universal Quality Evaluation in Translation available in any language.

Beyond increasing in number of languages, BOUQuET is actively evolving, and we are currently working on designing quality control for each of the contributions and adding new languages to the incremental releases of BOUQuET and extending the benchmarking by further showing the capabilities of BOUQuET, e.g. increasing the evaluation of linguistic signals over its alternatives.

[12] https://commonvoice.mozilla.org/

[13] OLDI, https://oldi.org

## Limitations and Ethical Considerations

The BOUQuET dataset is still limited in the number of languages and translations.The benchmarking is quite complete (4 datasets comparison, 14 models and 2 metrics) but it can also be extended in several axes (number of languages, model diversity, linguistic analysis, analysis of MT errors). However, the entire purpose of this work is to describe the dataset and open-initiative, while providing a minimal benchmarking. Authors expect the community to extend the benchmarking by further using this dataset for further exploration. Creators and commissioned translation's annotators are paid a fair rate.

## References

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *CoRR*, abs/2001.08435.

Yulong Chen, Yang Liu, and Yue Zhang. 2021. DialogSum challenge: Summarizing real-life scenario dialogues. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 308–313, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Everlyn Chimoto and Bruce Bassett. 2022. COMET-QE and active learning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4735–4740, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marta Costa-jussà, Mariano Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alexandre Mourachko, Christophe Ropers, and Carleigh Wood. 2024. MuTox: Universal MUltilingual audio-based TOXicity dataset and zero-shot detector. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5725–5734, Bangkok, Thailand. Association for Computational Linguistics.

Marta R. Costa-jussà, Bokai Yu, Pierre Andrews, Belen Alastruey, Necati Cihan Camgoz, Joe Chuang, Jean Maillard, Christophe Ropers, Arina Turkantenko, and Carleigh Wood. 2024. 2m-belebele: Highly multilingual speech and american sign language comprehension dataset. *Preprint*, arXiv:2412.08274.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pages arXiv–2307.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *Preprint*, arXiv:2502.12404.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: Sentence-level multimodal and language-agnostic representations. *Preprint*, arXiv:2308.11466.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. Ethnologue: Languages of the world. twenty-seventh edition. https://www.ethnologue.com/. Last accessed on 2025-02-03.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Martin Gerlach and Francesc Font-Clos. 2018. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *CoRR*, abs/1812.08092.

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

M. A. K. Halliday and C. M. I. Matthiessen. 2004. *An Introduction to Functional Grammar*. Routledge.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Preprint*, arXiv:2309.04662.

William Labov. 1991. *Sociolinguistic patterns*. University of Pennsylvania Press.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Openorca: An open dataset of gpt augmented flan reasoning traces. `https://https://huggingface.co/Open-Orca/OpenOrca`.

Jean Maillard, Laurie Burchell, Antonios Anastasopoulos, Christian Federmann, Philipp Koehn, and Skyler Wang. 2024. Findings of the WMT 2024 shared task of the open language data initiative. In *Proceedings of the Ninth Conference on Machine Translation*, pages 110–117, Miami, Florida, USA. Association for Computational Linguistics.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gul̇çehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

NLLBTeam. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630:841–846.

Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. Estimating machine translation difficulty. *Preprint*, arXiv:2508.10175.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa

Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Shuo Sun and Kevin Duh. 2020. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Ioannis Tsiamas, David Dale, and Marta R. Costa-jussà. 2025. Improving language and modality transfer in translation by character-level modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20171–20187, Vienna, Austria. Association for Computational Linguistics.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The bitter lesson learned from 2,000+ multilingual benchmarks. *Preprint*, arXiv:2504.15521.

Xinyan Yu, Trina Chatterjee, Akari Asai, Junjie Hu, and Eunsol Choi. 2022. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3725–3743, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yiran Zhao, Chaoqun Liu, Yue Deng, Jiahao Ying, Mahani Aljunied, Zhaodonghui Li, Lidong Bing, Hou Pong Chan, Yu Rong, Deli Zhao, and Wenxuan Zhang. 2025. Babel: Open multilingual large language models serving over 90% of global speakers. *Preprint*, arXiv:2503.00865.

# A Specific guidance for paragraph and sentence creation

## A.1 Overview

The Bouquet-source dataset comprises 250 unique sentences in each of its source languages. This means that each linguist created (i.e., wrote from scratch, did not copy; see Section 2.4 above) 250 original sentences. These sentences were requested to be:

- Organized in logically structured paragraphs (see the Paragraphs section below)
- Representative of the linguistic structures and features most frequently used in specific domains (see the Domains section below)
- Representative of the most common register of language used in similar situations (see the Registers section below)
- Accompanied by a gold-standard (i.e., best in class) human translation into English.

## A.2 Paragraphs

The linguist received a template in the form of a spreadsheet, in which paragraph structures were designed and laid out. The template specified the exact number of paragraphs and the exact number of sentences for each of the paragraphs. Each paragraph was given a unique paragraph ID (e.g., P01, P02, P15). Each sentence within each paragraph was also given a serial, non-unique ID (e.g., S1, S2, S3).

## A.3 Domains

The template was divided into 8 domains:

1. How-to, written tutorials or instructions
2. Conversations (dialogues)
3. Narration (creative writing that doesn't include dialogues)
4. Social media posts
5. Social media comments (reactive)
6. Other web content
7. Reflective piece
8. Miscellaneous (address to a nation, disaster response, etc.)

The creators had to produce the set number of sentences for each of the domains; the structure of the template (domain / paragraph / sentence) could not be changed.

### A.4 Language Register Information

When creating sentences, the creators had to make sure that the register of language being used was representative of the most expected and appropriate register for the situation. When several registers were possible, the creators were asked to use discretion when selecting a register, while making sure that the chosen register was among the most expected and appropriate. To help them make a determination, we defined 3 main functional areas of language register:

- Connectedness: What type of connection do language users who initiate the text have with other language users?
- Preparedness: How much time do language users who initiate the text had or took to prepare the text?
- Social differential: What is the relative social status of the language users who initiate the text towards other language users?

### A.5 Linguistic Features

One of the main reasons for dividing the dataset into sections that correspond to domains is to attempt to cover as many registers and aspects of language as possible. For example, we know that:

- Some pro-drop languages may drop the subject pronouns more often in some situations than in others.
- Some case-marking languages may use some cases in specific situations but avoid them in others.
- In English, lexical density increases when the level of formality increases.
- Some languages use a specific past verb tense in storytelling, which stands out from other past verb tenses used in casual conversations or other situations.
- Some languages use specific verb moods in some situations but avoid them in others.

### A.6 Violating Content

While creating sentences, the creators were asked to avoid inserting violating content. Violating content is language that can fall under one (or more) of the below categories:

- Toxicity
- Illegal activities
- Stereotypes and biases

### A.7 Step-by-Step Description of Tasks

Please refer to Table 5 for the step-by-step description of the tasks.

### A.8 Additional Guidance on Domain-Specific Content

Dialogues, especially those inserted in long creative writing (such as novels), often include the name of the speaker or a cue mark (e.g., — ), and sometimes quotation marks. When creating sentences for conversations, the creators were let free to invent names for speakers or to label speaker turns (e.g., A, B); they were also asked to place the names or speaker reference in markup tags, similarly to this: <Name:>or <A:>.

**Emojis**: As there are emojis frequently in some social media and messaging domains, some representation was also expected from the creators. However, the creators were asked to keep this representation very limited, as there are no real agreed ways to translate them across hundreds of languages.

**Social media comments**: The creators were told that they could keep the structure of those comments flat, and that including tags was not absolutely necessary, though it was permitted (even expected).

**Disfluencies in informal conversations**: Disfluencies were permitted provided they were representative of conversations and they could be translated (i.e., there is some consensus on how to write them in the language — ah, oh, um).

## B Registers Details

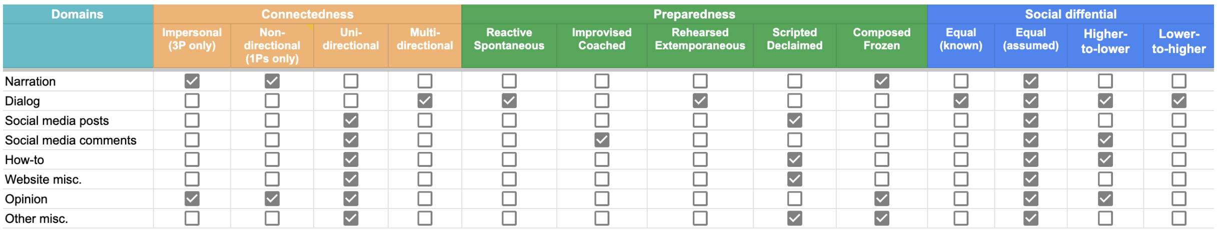
We provide non-exclusive options for each of the 3 functional areas that characterize registers described in Section 3.2 and mentioned in Figure 5. By non-exclusive, we mean that a domain may be characterized by more than one option. The functional area / option breakdown can be described as follows (the bold lowercase letters in square brackets represent a unique identifier for each option):

### Connectedness

- Impersonal [**i**]: For example a text written for the purpose of giving definitions or explanations with no specific readership in mind;

| | |
|---|---|
| **Column A: Lang-ID** | This column should have the same 3-lowercase-letter code representing the source language of the sentences being created followed by an underscore character ( _) and a 4-letter code representing the script. |
| **Column B: Domain** | This is 1 of the 8 domains represented in the dataset (see Section 3.1). |
| **Column C: Subdomain** | Please insert your description of the subdomain or topic. |
| **Column D: P-ID** | This is the unique code identifying a paragraph (e.g., P01, P02, . . . , P58). |
| **Column E: S-ID** | This is the non-unique code identifying the sequential place of the sentence within a paragraph. |
| **Column F: Sentence** | In this cell, please type a sentence you created. |
| **Column G: Translation into English** | After entering a sentence in your language in Column F, please provide a gold-standard human translation in this cell. |
| **Column H: S-Nchars** | This represents a count of the number of characters in the sentence. |
| **Column I: S Comment_src_lang** | To help other linguists expand this dataset by translating your sentences into their own languages, please add any comments that bring more context about the sentence. |
| **Column J: S Comment_English** | Please provide an English translation of the comment your inserted in Column I. |
| **Column K: Linguistic features** | Please list the register- or domain-specific linguistic features you tried to showcase in the sentence. |
| **Column L: Connectedness** | Please use any of the options best describing the register area of Correctedness. |
| **Column M: Preparedness** | Please use 1 of the options best describing the register area of Preparedness. |
| **Column N: Social differential** | Please use any of the options best describing the register area of Social differential. |
| **Column O: Formality** | Please indicate the level of formality best characterizing the sentence. |
| **Column P: Relationship** | Please insert the intended relationship between the language users involved in the situation. |
| **Column Q: Idea origin** | Please insert the name of the media type or platform that inspired the sentence. |
| **Column R: P Comment_src_lang** | To help other linguists expand this dataset by translating your sentences into their own languages, please add any comments that bring more context about the entire paragraph. |
| **Column S: P Comment_English** | Please provide a translation into English for the comment you inserted in Column R. |
| **Column T: P-Nchars** | This represents a count of the number of characters in the current paragraph. |
| **Column U: Creator_Translator-ID** | Please insert your ID here, if it isn't pre-populated. |

Table 5: Step-by-step guidance.



| Domains | Connectedness | | | | Preparedness | | | | | Social diffential | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Impersonal (3P only) | Non-directional (1Ps only) | Uni-directional | Multi-directional | Reactive Spontaneous | Improvised Coached | Rehearsed Extemporaneous | Scripted Declaimed | Composed Frozen | Equal (known) | Equal (assumed) | Higher-to-lower | Lower-to-higher |
| Narration | ☑ | ☑ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☑ | ☐ | ☑ | ☐ | ☐ |
| Dialog | ☐ | ☐ | ☐ | ☑ | ☑ | ☐ | ☑ | ☐ | ☐ | ☑ | ☑ | ☑ | ☑ |
| Social media posts | ☐ | ☐ | ☑ | ☐ | ☐ | ☐ | ☐ | ☑ | ☐ | ☐ | ☑ | ☐ | ☐ |
| Social media comments | ☐ | ☐ | ☑ | ☐ | ☐ | ☑ | ☐ | ☐ | ☐ | ☐ | ☑ | ☑ | ☐ |
| How-to | ☐ | ☐ | ☑ | ☐ | ☐ | ☐ | ☐ | ☑ | ☐ | ☐ | ☑ | ☑ | ☐ |
| Website misc. | ☐ | ☐ | ☑ | ☐ | ☐ | ☐ | ☐ | ☑ | ☐ | ☐ | ☑ | ☐ | ☐ |
| Opinion | ☑ | ☑ | ☑ | ☐ | ☐ | ☐ | ☐ | ☐ | ☑ | ☐ | ☑ | ☑ | ☐ |
| Other misc. | ☐ | ☐ | ☑ | ☐ | ☐ | ☐ | ☐ | ☑ | ☑ | ☐ | ☑ | ☐ | ☐ |

Figure 5: Register functional areas and breakdowns within each functional areas and their representations across domains.

typically written in the third person only (e.g., a contract).

- Non-directional [**n**]: A text written with a readership in mind but that doesn't address the readership specifically (e.g., an author recounting a story)

- Uni-directional [**u**]: A text addressing a readership who either cannot respond or is asked to refrain from responding at a given time (e.g., the transcription of a presentation, such as a TED Talk)

- Multi-directional [**m**]: A text addressing a readership who can respond (e.g., SMS, DM) or representing the transcription of a dialogue involving 2 or more language users.

**Preparedness**

- Reactive (spontaneous) [**r**]: The production is immediate either because it needs to be or because the user wants it to be

- Improvised (coached) [**i**]: The production appears spontaneous but takes place after a period of general training or coaching (e.g., spokespeople who answer questions live but have had time to prepare and choose vocabulary to use or to avoid)

- Rehearsed (extemporaneous) [**e**]: The production is live but its overall structure has been carefully crafted and rehearsed (e.g., transcriptions of 20-minute presentations or speeches that aren't fully scripted and given from notes).

- Scripted (declaimed) [**s**]: The production may or may not be live and has been fully scripted (e.g., transcriptions of speeches used in teleprompters)

- Composed (frozen) [**c**]: The production is completely offline, and goes through iterations of reviewing and editing (e.g., the text of a novel).

**Social differential**

- Equal (known) [**k**]: The readership or addressees are known to be peers; this can include a very informal or colloquial attitude

- Equal (assumed) [**a**]: The readership or addressees are not known but assumed to be peers; this can include a casual or informal attitude but likely excludes a very colloquial one
- Higher-to-lower [**h**]: The readership or addressees are considered to be at a lower social level than the producer (e.g., the producer is arrogant or assumes a position of higher authority)
- Lower-to-higher [**l**]: The readership or addressees are considered to be at a higher social level than the producer (e.g., the producer wants to express deference, respect, or admiration)

## C Quality Checks Details in Multi-way extension

In order to make sure that the BOUQuET contextual information was taken into accound while translating BOUQuET, we used the following QA strategies:

1. Checking the correct co-referencing. The Bouquet dataset is a representation of natural language, and the usage of personal and possessive pronouns as a substitute for the nouns is a typical occurrence. If the internal co-referencing in the paragraph is broken (the wrong pronoun is used or the noun is repeated where the noun should be), it indicates that the paragraph was treated as a collection of sentences not linked to each other, rather than a paragraph of text.
2. Checking the lexical consistency. We made sure to check that vocabulary used to translate word denoting objects or events is appropriate in tone, style and register and is used consistently throughout each paragraph. For example, when checking, we found out that translations from Indonesian into Russian did not keep consistency for "potato fritters" ("perkedel kentang"), using three different ways to translate it in P-292. We later applied the necessary corrections.
3. Checking the grammatical consistency. Since the Bouquet dataset contains examples of different domains, we needed to check whether the verb tenses and syntax were appropriate for a given domain and used consistently throughout each paragraph. For example, when checking translated into German paragraphs which imitate fiction narration, we made sure that German Pratäritum tense is used appropriately, not Perfekt.
4. Checking the special symbols such as emojis and numbers.

## D Priority Languages

Table 6 shows the languages in which BOUQuET exists at the time of submission of this paper (May 2025).

## E Dataset Examples

Table 7 reports the sentences with highest similarity score computed with cosine distance of SONAR vectors across all 2,000 Source-BOUQuET English sentences.

Table 8 shows complete entries examples of the Source-BOUQuET dataset.

## F Domain representation details

To directly visualize the similarity of distributions of SONAR vectors (Section 4), we do a PCA-dimensionality reduction, fitted upon the combined multi-domain set, see Figure 6. The public domains from Figure 1 are represented in grey; alternatives evaluation datasets are represented in blue and BOUQuET is represented in red. Figure 6, from top to down, compares BOUQuET against FLORES-200, NTREX-128, NLLB-MD, respectively. We qualitatively observe that BOUQuET covers a wider range of domains.

Figures 6 shows the domain representation and overlap across datasets.

## G Difficulty and grammatical diversity analysis

**Difficulty.** We used a prompt with detailed instructions and few-shot examples (see Listing 1) to make Llama-3.3-70B[14] estimate the difficulty of English sentences using the CEFR labels. The few-shot examples were selected from the CEFR-SP dataset (Arase et al., 2022) of English sentences labeled by two human annotators with CEFR labels. On the development split of this dataset, our

[14] https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct

| ISO 639-3 | ISO 15924 | Language | Family | Subgroup | Class |
|---|---|---|---|---|---|
| arz<br>(+ arb | Arab | Egyptian Arabic<br>+Modern Stan. Arabic) | Afro-Asiatic | Central Semitic | Pivot |
| arz | Latn | Romanized Egyptian Arabic | Afro-Asiatic | Semitic | P1-HR |
| aar | Latn | Afar | Afro-Asiatic | Cushitic | P1-LR |
| agr | Latn | Aguaruna | Chicham | – | P1-LR |
| ami | Latn | Amis | Austronesian | East Formosan | P1-LR |
| ben | Beng | Bengali | Indo-European | Indo-Aryan | P1-HR |
| cmn | Hans | Mandarin Chinese | Sino-Tibetan | Sinitic | Pivot |
| ces | Latn | Czech | Indo-European | Balto-Slavic | P1-HR |
| crk | Cans | Plains Cree | Algic | Algonquian | P1-LR |
| deu | Latn | German | Indo-European | West Germanic | Pivot |
| dje | Arab, Latn | Zarma | Songhay | Eastern Songhay | P1-LR |
| ell | Grek | Modern Greek | Indo-European | Hellenic | P1-HR |
| fra | Latn | French | Indo-European | Italic | Pivot |
| gaz | Latn | West Central Oromo | Afro-Asiatic | Cushitic | P1-LR |
| gil | Latn | Gilbertese | Austronesian | Micronesian | P1-LR |
| guc | Latn | Wayuu | Arawakan | Caribbean Arawakan | P1-LR |
| hin | Deva | Hindi | Indo-European | Indo-Aryan | Pivot |
| hin | Latn | Romanized Hindi | Indo-European | Indo-Aryan | P1-HR |
| hrv | Latn | Croatian | Indo-European | Balto-Slavic | P1-HR |
| hun | Latn | Hungarian | Uralic | Hungaric | P1-HR |
| ind | Latn | Indonesian | Austronesian | Malayic | Pivot |
| ita | Latn | Italian | Indo-European | Italic | P1-HR |
| jav | Latn | Javanese | Austronesian | Javanesic | P1-HR |
| jpn | Jpan | Japanese | Japonic |  | P1-HR |
| kaa | Cyrl | Karakalpak | Turkic | Kipchak | P1-LR |
| kal | Latn | Kalaallisut | Eskimo-Aleut | Eskimo | P1-LR |
| khm | Khmr | Central Khmer | Austroasiatic | Mon-Khmer | P1-HR |
| kor | Kore | Korean | Korean | Koreanic | P1-HR |
| kru | Deva | Kurukh | Dravidian | North Dravidian | P1-LR |
| lij | Latn | Ligurian | Indo-European | Italic | P1-LR |
| lin | Latn | Kinshasa Lingala | Atlantic-Congo | Central West. Bantu | P1-LR |
| mya | Mymr | Burmese | Sino-Tibetan | Burmo-Qiangic | P1-LR |
| nld | Latn | Standard Dutch | Indo-European | West Germanic | P1-HR |
| pes | Arab | Western Persian | Indo-European | Iranian | P1-HR |
| pol | Latn | Polish | Indo-European | Balto-Slavic | P1-HR |
| rus | Cyrl | Russian | Indo-European | Balto-Slavic | Pivot |
| ron | Latn | Romanian | Indo-European | Italic | P1-HR |
| sba | Latn | Ngambay | Central Sudanic | Sara-Bongo-Bagirmi | P1-LR |
| spa | Latn | Spanish | Indo-European | Italic | Pivot |
| por | Latn | Portuguese (Brazilian) | Indo-European | Italic | P1-HR |
| swe | Latn | Swedish | Indo-European | North Germanic | P1-HR |
| swh | Latn | Coastal Swahili | Atlantic-Congo | N.E. Coastal Bantu | P1-HR |
| tha | Thai | Thai | Tai-Kadai | Southwestern Tai | P1-HR |
| tir | Ethi | Tigrinya | Afro-Asiatic | Semitic | P1-LR |
| tgl | Latn | Tagalog | Austronesian | Greater Central Philippine | P1-HR |
| tur | Latn | Turkish | Turkic | Oghuz | P1-HR |
| ukr | Cyrl | Ukrainian | Indo-European | Balto-Slavic | P1-HR |
| urd | Arab | Urdu | Indo-European | Indo-Aryan | P1-HR |
| vie | Latn | Vietnamese | Austroasiatic | Vietic | P1-HR |
| yor | Latn | Yoruba | Atlantic-Congo | Defoid | P1-LR |
| zlm<br>+zsm | Latn | Colloquial Malay<br>+ Standard Malay | Austronesian | Malayic | P1-HR |

Table 6: Source-BOUQuET Languages (Pivot) and Priority languages (P) both high-resource (HR) and low-resource (LR) included in BOUQuET at the time of submission. Note that these languages have been commissioned, we do not include updates in annotations collected from the open-initiative, which we will include in later versions of the paper.

Llama-based labels demonstrate high Spearman correlation with the average of the human labels (75%, getting close to the 79% correlation that the human labels have with each other).

**Grammatical diversity.** To evaluate how diverse each dataset is in terms of grammar, we computed the morphological features of each word (e.g. `Aspect=Perf|Tense=Past|VerbForm=Part` for the word "happened") using a Spacy

| cosinedist | lang-A | lang-B | Domain-A | Text-A | UNIQID-A | DomainB | Text-B | UNIQID-B |
|---|---|---|---|---|---|---|---|---|
| 0.19 | ind | arz | conversation | What time do we meet? | P304-S4 | conversation | when will we meet? | P017-S1 |
| 0.20 | fra | cmn | web misc. | About us | P220-S1 | web misc. | About our team | P098-S1 |
| 0.22 | rus | deu | conversation | <B:> Which one? | P363-S2 | conversation | <B:> When and where? | P134-S2 |
| 0.24 | rus | fra | conversation | <B:> Nah, I am sick | P360-S2 | conversation | <B:>You're sick? | P185-S4 |
| 0.25 | rus | fra | conversation | <A:> You know what I mean! | P362-S4 | conversation | <A:>Did you hear? | P183-S1 |
| 0.28 | fra | arz | web misc. | Send us your résumé and motivation letter at the below address. | P215-S6 | web misc. | Please sendyour CV with letters of recommendation to this email address | P043-S5 |
| 0.28 | spa | rus | comments | <B:> WHAT IS THIS??? | P443-S2 | conversation | <B:> What do you mean? | P362-S2 |
| 0.29 | rus | fra | conversation | <A:> Get well soon | P360-S3 | conversation | <A:>Not doing very well. | P185-S3 |
| 0.29 | rus | fra | conversation | <B:> What do you mean? | P362-S2 | conversation | <A:>Did you hear? | P183-S1 |
| 0.29 | rus | fra | conversation | <B:> What do you mean? | P362-S2 | conversation | <B:>You're sick? | P185-S4 |
| 0.29 | rus | fra | conversation | <B:> Nothing is working for me. | P366-S2 | conversation | <A:>Not doing very well. | P185-S3 |

Table 7: Source-BOUQuET sentences with closest similarity score (cosine distance lower than 0.3)

| LangID | Domain | Subdomain | PID | SID | Sentence | English | Linguistic label | Reg. |
|---|---|---|---|---|---|---|---|---|
| spa_Latn | conversation | text message chain | P417 | S1 | <Guillermo:> Habéis cenado ya? | <Guillermo:> Have you had dinner already? | word:named-entity | mrk |
| spa_Latn | conversation | text message chain | P417 | S2 | <Jaime:> No, estábamos pensando en salir ahora, te apuntas? | <Jaime:> No, we were thinking about going out now. Are you in? | word:named-entity | mrk |
| spa_Latn | conversation | text message chain | P417 | S3 | <Guillermo:> Sí, me estoy muriendo de hambre. | <Guillermo:> Yes, I'm starving. | word:named-entity, miscellaneous:collocation | mrk |
| spa_Latn | conversation | text message chain | P417 | S4 | <Jaime:> Guai, salimos en cinco, te esperamos en la parada del metro. | <Jaime:> Cool, we're leaving in five, we'll wait for you at the metro station. | word:named-entity, word:slang | mrk |
| spa_Latn | conversation | text message chain | P417 | S5 | <Guillermo:> Perfecto, me cambio y salgo. | <Guillermo:> Perfect, I'll change and head out. | word:named-entity | mrk |
| fra_Latn | social posts | Integrity | P204 | S1 | Choses que j'aurais aimé savoir plus tôt | Things I wish I had known earlier | sentence:fragment | usa |
| fra_Latn | social posts | Integrity | P204 | S2 | Si tu ne prends pas de décision pour toi-même, d'autres les prendront pour toi. | If you don't make decisions for yourself, others will take them for you. | word:impersonal-pronoun | usa |
| fra_Latn | social posts | Integrity | P204 | S3 | Quand on te submerge de généralités, demande plusieurs exemples spécifiques. | When you are getting submerged in generalities, request several specific examples. | word:impersonal-pronoun | usa |
| ind_Latn | narration | Folklore / Fable | P310 | S1 | Pada suatu masa, hiduplah sepasang suami istri di sebuah pedesaan. | Once upon a time, there lived a husband and wife in a village. | Third person, impersonal, narration | ica |
| ind_Latn | narration | Folklore / Fable | P310 | S2 | Mereka belum juga dikarunia anak setelah sekian lama menikah. | They have not yet been blessed with children after being married for so long. | Third person, impersonal, narration | ica |
| ind_Latn | narration | Folklore / Fable | P310 | S3 | Keduanya bermimpi bahwa mereka harus menanam timun, jika mereka ingin memiliki anak. | Both of them dreamed that they had to plan cucumbers, if they wanted to have a child. | Third person, impersonal, narration | ica |
| ind_Latn | narration | Folklore / Fable | P310 | S4 | Kemudian ditanamlah timun-timun itu. | Then they planted the cucumbers. | Third person, impersonal, narration | ica |

Table 8: BOUQuET examples including main fields

en_core_web_md model[15] (skipping the tokens with no morphological labels, such as prepositions and clitics). As a diversity measure, we used the Shannon entropy computed over the distribution of all word morphological labels occurring in it (with higher values indicating more diverse distribution).

As Figure 3 shows, the average difficulty of a dataset appears to correlate negatively with its grammatical diversity. Note that lexical diversity (which we could measure similarly, by computing the entropy of the distribution of word lemmas) behaves differently: it is positively correlated with

[15] https://spacy.io/models/en#en_core_web_md

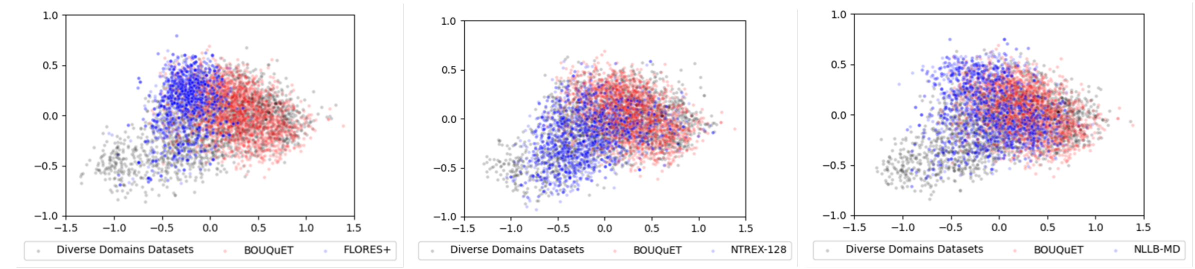


Figure 6: Domain representation and overlap across FLORES+ (left), NTREX-128 (middle), NLLB-MD (right)(in blue) with diverse domains datasets (in grey) and BOUQuET (in red).

difficulty and negatively, with grammatical diversity. We conjecture that the difficulty of translation for humans is mostly driven by difficult words (and by how nested the syntactic constructions are, which can be measured by the depth of the dependency tree or simply by the text length) than by the forms into these words are inflected. The statistics in Table 9 corroborate this hypothesis.

## H Detailed results

Table 10 reports averaged MT results from and into all pivot languages.

Figure 7 shows results of the 3 best systems averaged across language directions, evaluated at the sentence-level, per domains. Worse performing domains are comments, conversations, how-to and narration. Best performing domains are web and other miscellaneous, reflection and social posts.

## I Automated quality checks for the translation contributions

Currently, we are considering analysing the following automated scores for the acceptance of crowd-sourced BOUQuET translations:

- Cosine similarity of SONAR embeddings for detecting translation accuracy problems, such as omissions. To compute the embeddings, we plan to use a text encoder inspired by (Tsiamas et al., 2025) with enhanced zero-shot cross-lingual generalization.

- Back-translation into one of the Source-BOUQuET languages with subsequent translation quality estimation, as an alternative accuracy check.

- Applying language identification (LID) models to validate the language of the contributions.

- Analysis of added toxicity, using the methodology of (NLLBTeam, 2024) or SONAR-based classification, such as MuTox (Costa-jussà et al., 2024).

Based on the analysis of the subsequent contributions, we may revise the list of these checks.

| dataset | domain | grammatical diversity | lexical diversity | tree depth | num words | difficulty |
|---|---|---|---|---|---|---|
| BOUQuET | comments | 3.18 | 5.35 | 3.44 | 13.71 | 2.37 |
| | conversation | 3.18 | 5.26 | 3.06 | 13.58 | 1.91 |
| | how-to — instructions | 2.92 | 5.57 | 4.38 | 17.00 | 2.45 |
| | narration | 3.11 | 5.58 | 4.39 | 16.80 | 2.76 |
| | other misc. | 3.02 | 5.62 | 4.52 | 15.34 | 2.45 |
| | reflection | 3.10 | 5.59 | 4.95 | 18.67 | 3.37 |
| | social posts | 3.17 | 5.63 | 4.18 | 16.12 | 2.71 |
| | web misc. | 2.94 | 5.76 | 4.40 | 14.38 | 2.47 |
| FLORES | wikibooks | 2.91 | 6.05 | 5.95 | 24.42 | 3.67 |
| | wikinews | 2.78 | 6.25 | 5.89 | 23.32 | 3.71 |
| | wikivoyage | 2.89 | 6.10 | 5.92 | 24.35 | 3.43 |
| NLLB-MD | chat | 3.27 | 5.41 | 4.37 | 27.32 | 2.85 |
| | health | 2.76 | 6.11 | 6.37 | 26.91 | 4.19 |
| | news | 2.90 | 5.98 | 5.39 | 22.38 | 3.47 |
| NLLB-Seed | - | 2.81 | 6.55 | 6.29 | 25.50 | 4.48 |
| NTREX | - | 2.95 | 6.51 | 5.93 | 24.45 | 3.62 |
| SmolSent | - | 2.90 | 6.66 | 4.81 | 16.68 | 3.44 |

Table 9: Average values of grammatical diversity (entropy of morphological labels distribution), lexical diversity (entropy of lemmas distribution), syntactic tree depth, number of words, and mean difficulty in each of the studied datasets. The difficulty is computed with Llama-3.3-70B, the other features, with spacy.

| Src-lang | arz-Arab | | cmn-Hans | | deu-Latn | | eng-Latn | | fra-Latn | | hin-Deva | | ind-Latn | | rus-Cyrl | | spa-Latn | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | COM | METX | COM | METX | COM | METX | COM | METX | COM | METX | COM | METX | COM | METX | COM | METX | COM | METX |
| nllb-3B | 0.59 | 2.31 | 0.66 | 1.39 | 0.72 | 2.39 | 0.75 | 1.86 | 0.69 | 2.11 | 0.61 | 2.33 | 0.69 | 2.03 | 0.67 | 2.33 | 0.72 | 2.13 |
| aya101-13B | 0.59 | 2.23 | 0.65 | 1.32 | 0.71 | 2.31 | 0.75 | 1.76 | 0.67 | 2.07 | 0.6 | 2.32 | 0.69 | 1.88 | 0.67 | 2.26 | 0.71 | 1.99 |
| aya-e-8B | 0.6 | 2.04 | 0.68 | **1.09** | 0.73 | **2.06** | **0.77** | **1.47** | 0.7 | **1.77** | 0.62 | **1.94** | 0.7 | **1.69** | 0.69 | **1.94** | 0.73 | **1.76** |
| babel-9B | 0.57 | 2.78 | 0.66 | 1.61 | 0.72 | 2.58 | 0.75 | 1.92 | 0.68 | 2.47 | 0.59 | 2.54 | 0.69 | 2.16 | 0.67 | 2.47 | 0.71 | 2.46 |
| cohere-r7B | 0.58 | 2.38 | 0.66 | 1.28 | 0.72 | 2.45 | 0.75 | 1.87 | 0.66 | 2.52 | 0.61 | 2.18 | 0.7 | 1.86 | 0.66 | 2.65 | 0.71 | 2.16 |
| eurollm-9B | 0.58 | 2.55 | 0.66 | 1.59 | 0.71 | 2.68 | 0.75 | 2.04 | 0.68 | 2.38 | 0.61 | 2.41 | 0.66 | 2.37 | 0.67 | 2.58 | 0.71 | 2.36 |
| madlad-10B | 0.52 | 3.71 | 0.62 | 1.8 | 0.66 | 3.23 | 0.73 | 2.0 | 0.64 | 2.61 | 0.58 | 2.75 | 0.62 | 3.06 | 0.63 | 3.08 | 0.69 | 2.39 |
| madlad-3B | 0.51 | 3.93 | 0.62 | 1.76 | 0.63 | 3.51 | 0.72 | 2.26 | 0.63 | 2.82 | 0.59 | 2.63 | 0.61 | 3.25 | 0.63 | 3.1 | 0.7 | 2.4 |
| mistral-7B | 0.44 | 5.32 | 0.54 | 3.53 | 0.59 | 4.52 | 0.6 | 3.83 | 0.55 | 4.2 | 0.46 | 4.95 | 0.59 | 3.81 | 0.56 | 4.31 | 0.59 | 4.11 |
| qwen-7B | 0.5 | 3.83 | 0.58 | 2.55 | 0.63 | 3.49 | 0.68 | 2.79 | 0.6 | 3.25 | 0.52 | 3.45 | 0.61 | 3.16 | 0.59 | 3.47 | 0.62 | 3.29 |
| Llama-3.1-8B | 0.56 | 2.77 | 0.64 | 1.58 | 0.7 | 2.63 | 0.75 | 1.88 | 0.66 | 2.46 | 0.58 | 2.77 | 0.68 | 2.24 | 0.66 | 2.52 | 0.7 | 2.38 |
| Llama3.2-3B | 0.46 | 4.88 | 0.58 | 2.56 | 0.63 | 3.86 | 0.67 | 3.02 | 0.58 | 3.77 | 0.53 | 3.78 | 0.61 | 3.35 | 0.59 | 3.46 | 0.63 | 3.62 |
| Llama3.3-70B | **0.61** | **1.96** | **0.68** | 1.18 | **0.74** | 2.14 | **0.77** | 1.62 | **0.71** | 1.94 | **0.62** | 2.1 | **0.71** | 1.73 | **0.69** | 2.05 | **0.74** | 1.91 |
| Tower-7B | 0.4 | 6.07 | 0.62 | 1.5 | 0.63 | 3.23 | 0.66 | 3.48 | 0.56 | 4.21 | 0.49 | 4.42 | 0.62 | 3.26 | 0.6 | 3.61 | 0.64 | 3.46 |

| Trg-lang | arz-Arab | | cmn-Hans | | deu-Latn | | eng-Latn | | fra-Latn | | hin-Deva | | ind-Latn | | rus-Cyrl | | spa-Latn | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | COM | METX | COM | METX | COM | METX | COM | METX | COM | METX | COM | METX | COM | METX | COM | METX | COM | METX |
| NLLB-3B | 0.62 | 3.08 | 0.59 | 2.99 | 0.71 | 0.99 | 0.76 | 2.22 | 0.69 | 2.01 | 0.61 | 2.55 | 0.7 | 1.53 | 0.71 | 1.77 | 0.72 | 1.76 |
| aya101-13B | 0.64 | 2.61 | 0.63 | 1.82 | 0.69 | 1.07 | 0.76 | 2.26 | 0.68 | 2.14 | 0.56 | 2.94 | 0.69 | 1.55 | 0.69 | 1.85 | 0.7 | 1.89 |
| aya-e-8B | **0.7** | **1.9** | 0.65 | 1.81 | 0.72 | **0.85** | 0.77 | **2.01** | **0.71** | **1.82** | 0.53 | 3.02 | 0.71 | **1.3** | 0.72 | **1.47** | **0.73** | **1.59** |
| babel-9B | 0.64 | 3.06 | 0.66 | 1.96 | 0.68 | 1.31 | 0.76 | 2.1 | 0.69 | 2.06 | 0.53 | 3.74 | 0.68 | 2.42 | 0.69 | 2.42 | 0.71 | 1.91 |
| cohere-7B | 0.68 | 2.38 | 0.65 | 1.87 | 0.7 | 1.0 | 0.76 | 2.13 | 0.69 | 1.95 | 0.51 | 3.99 | 0.66 | 2.08 | 0.68 | 2.24 | 0.72 | 1.72 |
| eurollm-9B | 0.7 | 1.99 | 0.66 | 1.65 | 0.72 | 0.89 | 0.77 | 2.13 | 0.7 | 1.87 | 0.6 | 2.65 | 0.45 | 6.53 | 0.72 | 1.57 | 0.72 | 1.67 |
| madlad-10B | 0.65 | 2.67 | 0.59 | 2.61 | 0.67 | 1.64 | 0.75 | 2.47 | 0.66 | 2.71 | 0.41 | 5.12 | 0.62 | 2.64 | 0.66 | 2.66 | 0.7 | 2.1 |
| madlad-3B | 0.65 | 2.8 | 0.58 | 2.77 | 0.66 | 1.73 | 0.73 | 2.79 | 0.64 | 2.81 | 0.42 | 5.1 | 0.62 | 2.64 | 0.65 | 2.83 | 0.69 | 2.2 |
| mistral-7B | 0.32 | 8.73 | 0.55 | 3.12 | 0.62 | 1.85 | 0.73 | 2.7 | 0.62 | 2.96 | 0.3 | 8.42 | 0.52 | 4.52 | 0.61 | 3.38 | 0.65 | 2.89 |
| qwen-7B | 0.53 | 4.74 | 0.64 | 2.09 | 0.63 | 1.76 | 0.72 | 2.39 | 0.64 | 2.63 | 0.25 | 7.38 | 0.64 | 2.65 | 0.61 | 3.12 | 0.66 | 2.49 |
| Llama3.1-8B | 0.54 | 4.68 | 0.64 | 1.91 | 0.69 | 1.17 | 0.76 | 2.25 | 0.68 | 2.17 | 0.57 | 3.05 | 0.68 | 1.81 | 0.68 | 2.18 | 0.7 | 2.01 |
| Llama3.2-3B | 0.43 | 6.84 | 0.55 | 3.11 | 0.63 | 1.76 | 0.73 | 2.62 | 0.62 | 2.86 | 0.49 | 4.5 | 0.62 | 2.7 | 0.54 | 5.27 | 0.66 | 2.66 |
| Llama3.3-70B | 0.6 | 3.32 | **0.68** | **1.62** | **0.73** | 0.85 | 0.77 | **2.04** | **0.71** | 1.84 | **0.63** | **2.41** | **0.72** | 1.35 | 0.72 | 1.55 | 0.73 | 1.64 |
| Tower-7B | 0.3 | 7.96 | 0.61 | 2.33 | 0.67 | 1.44 | 0.74 | 2.74 | 0.66 | 2.6 | 0.41 | 6.54 | 0.51 | 4.62 | 0.66 | 2.52 | 0.68 | 2.5 |

Table 10: Averaged results on CometKiwi (COM) and MetricX (etx) at the sentence-level from 9 BOUQuET languages (top) and into (bottom). Best results are in bold (before rounding to 2 decimals). Best results on CometKiwi tend to be with Llama-3.3-70B (the largest model) and best results in MetricX tend to be with Aya-expanse-8B. Best direction is from and into English .

Listing 1: Model prompt for estimating CEFR sentence complexity

~~~~
Please evaluate the difficulty of the following sentence on the CERF scale (from A1
    to C2). Sentence:
‘‘‘
{sentence}
‘‘‘
When evaluating the complexity of the sentences, assume that the person does not
    have expertise in any narrow field (science, law, finance, sports, art, etc).
Therefore, any specialized terminology would imply a high difficulty level.
Also, the person is used to communicate in English mostly orally and in a colloquial
     way, so complex syntactical structures should also imply high level of
    difficulty.
Finally, the person is not a native speaker of English, so idiomatic expressions,
    slang, and rare words may also present a difficulty.

Here are the descriptions of reading skills for each level:
- A1: I can understand familiar names, words and very simple sentences, for example
    on notices and posters or in catalogues.
- A2: I can read very short, simple texts. I can find specific, predictable
    information in simple everyday material such as advertisements, prospectuses,
    menus and timetables and I can understand short simple personal letters.
- B1: I can understand texts that consist mainly of high frequency everyday or job-
    related language. I can understand the description of events, feelings and
    wishes in personal letters.
- B2: I can read articles and reports concerned with contemporary problems in which
    the writers adopt particular attitudes or viewpoints. I can understand
    contemporary literary prose.
- C1: I can understand long and complex factual and literary texts, appreciating
    distinctions of style. I can understand specialised articles and longer
    technical instructions, even when they do not relate to my field.
- C2: I can read with ease virtually all forms of the written language, including
    abstract, structurally or linguistically complex texts such as manuals,
    specialised articles and literary works.

Below are some examples of the sentences of various difficulty levels:
A1:
- No one had money.
- Wass died on 4 January 2017 at the age of 93.
A2:
- If he were my father, I ’d write him.
- Historically, Latin or Romance has been the official language.
B1:
- He has been dieting since his heart attack last spring.
- Color blindness is very sensitive to changes in material.
B2:
- Give your urine sample to the lab technician.
- The Dr. I suffered other deficiencies.
C1:
- The evidence that preventive antibiotics decrease urinary tract infections in
    children is poor.
- The following is a partial list of notable nonfiction works discussing anarcho-
    capitalism.
C2:
- In a straightforward example the two bromine atoms in 3 - tert- butyl- trans - 1,
    2 - dibromohexane mutarotate by heating.
- Arachnids pour digestive juices produced in their stomachs over their prey after
    killing it with their pedipalps and chelicerae.

Now, please tell me what is the difficulty of this sentence:
‘‘‘
{sentence}
‘‘‘
Please make sure the label in your response is one of the following: A1, A2, B1, B2,
     C1, C2. Feel free to assign extreme labels (A1, C2), if needed.
Start with the brief explanation of your observations and logic. Then summarize the
    overall difficulty of the sentence and propose an approproate label.
On the last line, please write once more the label, and the label only, without
    writing anything else (even punctuation).
~~~~

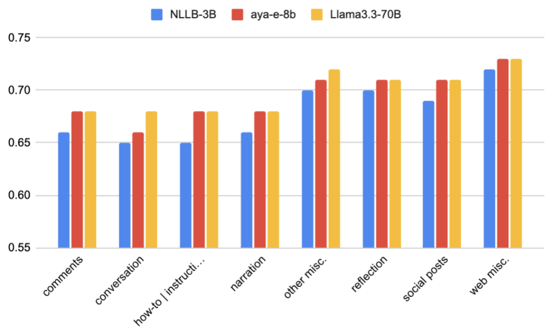


Figure 7: Best performing models and their results in each of the BOUQuET domains .