The corpus is available all languages we consider except for Hindi since it is not considered one of the official languages of the UN.

- USER REVIEWS: User text reviews for products, movies, books, hotels, and restaurants, are sampled from open-source datasets in each language when available. Most these datasets are used in sentiment analysis research.

- DIALOGUE: Conversational text data is collected from three different types of open-source dialogue datasets: **Open-domain** dialogue datasets which focus on open-ended general conversation (Naous et al., 2021; Li et al., 2017; Zhang et al., 2022), **Task-oriented** datasets that are design to train human-assistance or customer support dialogue models (van der Goot et al., 2021; Malviya et al., 2021), and **Negotiation** dialogues that are used in developing automated sales dialogue agents with negotiation capabilities (He et al., 2018).

- FINANCE: We leverage the Financial Phrasebank dataset (Malo et al., 2014) which provides English sentences with financial references and content collected from finance-focused news, and the CoFiF corpus (Daudert and Ahmadi, 2019) which provides financial reports in French.

- FORUMS: We collect text from several online forums. These include:
  Reddit: Reddit is a popular platform where online communities discuss common interests and passions. We used the latest version of the Reddit dump available at the time of this study to sample user posts. We filtered posts for language using the fasttext language identification model with a confidence > 0.9. NSFW and Over 18 content were automatically filtered before sampling. Further, any sampled sentence that still contained sexual or offensive content was manually removed.

  QA Websites: We collected questions and answers from QA websites using publicly available datasets for Question Answering research (Nakov et al., 2016; Quora.com, 2017; Howard et al., 2021; d'Hoffschmidt et al., 2020; Efimov et al., 2020).

  StackOverflow: Sentences were collected from the StackOverflow NER dataset (Tabassum et al., 2020) which contains user posts that describe what the user is trying to accomplish, a problem they are facing, or questions to seek advice from the community.

- SOCIAL MEDIA: We sample tweets from the the Stanceosaurus dataset (Zheng et al., 2022) which provides thousands of tweets in English, Arabic, and Hindi that discuss recent region-specific rumors. French tweets were sampled from the dataset of Kozlowski et al. (2020) built to detect crisis messages in French tweets, while Russian tweets were sampled from the RuSentiTweet dataset (Smetanin, 2022) for sentiment analysis in Russian. Tweets that include offensive or hate speech were manually omitted.

- POLICIES: We group under "Policies" several type of documents that delineate plans of what to do in a particular situation. This includes text extracted from: freely available **contract** templates for apartment/house leasing and job employment, **Special Olympics rules** which are available in multiple languages among which are but not in Hindi, and online **codes of conduct** of different organizations that we identify.

- GUIDES: Several domains that aim at providing instructions to the reader are grouped under "Guides". We extract data from Samsung Smartphones **User Manuals** which are available in a variety of languages. Another source is **Online Tutorials** which we collect from WikiHow that provides how-to articles in multiple languages. We also manually collect **Recipe Instructions** from multiple online cooking resources for each language. Additionally, we collect **Code Documentation** sentences from documentation of different functions of the Matlab software[10].

- CAPTIONS: We collect four different types of captions: image and video captions from various public datasets used in automatic captioning research, movie subtitles from the OpenSubtitles (Lison and Tiedemann, 2016) dataset used in machine translation research, and YouTube captions that we manually collect from video released under a Creative Commons license. While high-quality YouTube captions are easy to find for English, we could not find any high-

[10] mathworks.com

quality YouTube captions for non-English languages.

- MEDICAL TEXT: We use clinical reports written by medical professionals from the i2b2/VA dataset (Uzuner et al., 2011). We could not find similar high-quality medical resources for non-English languages.

- DICTIONARIES: We manually collect sentence examples from Arabic and English dictionaries using words that have appeared in the Word of the Day. No similar resource under a Creative Commons license was found for Hindi, French, and Russian.

- ENTERTAINMENT: We use Humour detection datasets to collect jokes (Al-Khalifa et al., 2022; Weller and Seppi, 2019; Jokes). Hindi jokes were manually collected.

- SPEECH: Two types of sources for speech data are used: **publicly available presidential speeches** that are usually posted on governmental websites. We used speeches by the United States President that are posted on the department of state's website. These speeches are also professionally translated to Arabic. We also collect sentences from **TED Talk transcriptions**, which are professionally translated from English to multiple languages.

- STATEMENTS: Two different types of standalone sentences that we group under "statements" were identified which are: Rumours, and quotes. We collect rumours in Arabic, English, and Hindi from the Stanceosaurus dataset (Zheng et al., 2022) used in misinformation detection. The rumours/claims are collected from various fact-checking websites in the Arab World, India, and the U.S. We also manually collected quotes in the three languages from various online resources. We did not collect mere translations of famous English quotes to other languages but focused on quotes by old scholars and thinkers of the Arab World, France, Russia and India for more cultural representation.

- POETRY: Poetry lines are extracted from English, Arabic, and Hindi poems, some of which date back several centuries ago. To have culture specific samples, we focus on non-English poems from original Arab, French, Indian, and Russian authors, and not poems translated from English.

- LETTERS: English letters were collected from online archives of historic letters. No high-quality authentic letters were found in Arabic or Hindi.

A.2 Domain Distribution

Table 6 shows the distribution of the domains in each readability level for each language. Basic readability levels (A1, A2) mostly contains sentences from domains that have text that is straightforward to read and contains day-to-day vocabulary such as Captions, Dialogue, User Reviews, User Guides. Intermediate readability levels (B1, B2) largely contain sentences from domains that present factual content such as books, Wikipedia articles, policy documents, news articles, etc. Proficient levels (C1, C2) contain domains that are scientific and technical such as finance, medical, legal documents, or highly literary text such as Arabic Poetry. We show the distribution of readability levels per domain in Figure 8.

A.3 Sentence Examples

Example sentences from various domains are shown in Table 13 for English, Table 14 for Arabic, Figure 13 for Hindi, Figure 14 for French, and Figure 15 for Russian.

B CEFR Levels Descriptors

The CEFR levels descriptors are provided in Table 7. Each level is described by specific capabilities of a language learner which we used to familiarize annotators with the intuition behind the scale being used prior to labeling.

Lang	Readability Level	Distribution (>5%)
ar	A1	Captions (50.62%) Dialogue (28.4%) Reviews (7.41%)
	A2	Reviews (19.44%) Dialogue (18.65%) Guides (17.46%) Captions (12.7%) Social Media (5.45%) Literature (5.95%)
	B1	Wikipedia (22.37%) Reviews (15.76%) Guides (13.23%) News (10.12%) Speech (6.03%) Legal (5.84%)
	B2	News (21.59%) Wikipedia (21.06%) Reviews (6.9%) Entertainment (6.73%) Legal (6.55%) Policies (6.37%) Speech (5.31%)
	C1	Wikipedia (40.29%) Research (14.53%) Literature (13.43%) Textbooks (5.71%)
	C2	Poetry (24.04%) Wikipedia (26.23%) Novels (18.58%) Dictionaries (9.84%) Quotes (6.01%)
fr	A1	Captions (44.29%) Dialogue (9.29%) Twitter (8.57%) Poetry (7.86%) Quotes (5%)
	A2	Recipes (9.02%) Dialogue (12.02%) Twitter (7.1%) Quotes (7.1%) QA Websites (6.28%) Children Stories (5.46%)
	B1	Wikipedia (21.85%) Guides (15.32%) Books (10.36%) Legal (6.98%) Reddit (5.41%)
	B2	Wikipedia (43.47%) Legal (10.51%) Policies (9.66%) Books (7.39%) Guides (6.25%)
	C1	Wikipedia (46.47%) Policies (12.03%) Research (9.96%) Finance (7.74%)
	C2	Research (21.43%) Policies (7.14%) Finance (6.39%)
en	A1	Dialogue (38.25%) Captions (27.87%) Reviews (10.38%) Guides (5.46%)
	A2	Captions (16.74%) Reviews (13.33%) Statements (8.15%) Guides (10.03%) Dialogue (8.74%) Forums (7.41%) Entertainment (5.63%)
	B1	Wikipedia (16.72%) Reviews (13.85%) News (11.74%) Forums (7.8%) Guides (8.12%) Textbooks (7.17%)
	B2	Wikipedia (21.94%) News (11.8%) Research (10.8%) Textbooks (11.03%) Policies (7.83%) Literature (7.39%)
	C1	Wikipedia (24.23%) Research (13.14%) Literature (12.82%) Legal (9.54%) Textbooks (9.28%) Policies (5.67%) News (5.65%)
	C2	Wiki-Natural Sciences (16.25%) Literature (18.75%) Clinical Reports (11.25%) Research (8.7%) Textbooks (7.5%)
hi	A1	Captions (33.09%) Literature (16.91%) Dialogue (12.82%) Jokes (9.56%) Reviews (5.15%)
	A2	Captions (12.88%) Dialogue (12.88%) Forums (7.46%) Statements (7.46%) Children Stories (6.78%) (5.37%) Guides (5.76%)
	B1	Wikipedia (15.02%) Literature (13.31%) Guides (11.26%) Reviews (9.56%) Statements (8.53%) Forums (8.53%)
	B2	Wikipedia (21.27%) Textbooks (9.7%) Literature (9.33%) Poetry (8.96%) Research (7.46%) Policies (7.46%) Quotes (5.6%)
	C1	Wikipedia (31.08%) Textbooks (12.16%) Legal (10.36%) Research (10.36%) Literature (8.53%) Forums (7.21%) Poetry (5.41%)
	C2	Wikipedia (44.25%) Textbooks (10.92%) Legal (10.9%) Research (8.05%)
ru	A1	Reviews (10.7%) Recipes (9.2%) Twitter (9.45%) Dialogue (8.21%) Jokes (7.96%) Captions (5.97%)
	A2	Wikipedia (23.80%) Guides (15.36%) Research (8.19%) Speech (7.14%)
	B1	Wikipedia (32.76%) Guides (6.11%) Policies (5.62%) Legal (5.62%)
	B2	Wikipedia (34.05%) Research (20.86%) Legal (12.88%) Policies (9.51%) Community Websites (6.13%)
	C1	Wikipedia (31.65%) Research (26.16%) Legal (19.38%) Policies (8.81%)
	C2	Legal (28.42%) Research (17.58%) Policies (6.59%)

Table 6: Distribution of domains for each readability level in each language. Only domains that compose more than 5% of the distribution are show.

C Traditional Metrics

ARI and FKGL are statistical formulas based on the number of words, characters, and syllables.

Automated Readability Index (ARI). ARI aims at approximating the grade level needed by an individual to understand a text. It is computed by:

$$\text{ARI} = 4.71\left(\frac{\#\text{Chars}}{\#\text{Words}}\right) + 0.5\left(\frac{\#\text{Words}}{\#\text{Sents}}\right) - 21.43 \tag{3}$$

Flesch-Kincaid Grade Level (FKGL). FKGL also aims at predicting the grade level, but unlike ARI, considers the total number of syllables in the text. It is computed as follows:

$$\text{FKGL} = 0.39\left(\frac{\#\text{Words}}{\#\text{Sents}}\right) + 11.8\left(\frac{\#\text{Sylla}}{\#\text{Words}}\right) - 15.59 \tag{4}$$

Open Source Metric for Measuring Arabic Narratives (OSMAN). OSMAN is computed according to the following formula:

$$\text{OSMAN} = 200.791 - 1.015\left(\frac{A}{B}\right) + 24.181\left(\frac{C}{A} + \frac{D}{A} + \frac{G}{A} + \frac{H}{A}\right) \tag{5}$$

where *A* is the number of words, *B* is the number of sentences, *C* is the number of words with more than 5 letters, *D* is the number of syllables, *G* is the number of words with more than four syllabus, and *H* is the number of "Faseeh" words, which contain any of the letters (ظ، ذ، ؤ، ئ، ء) or end with (ون، وا).

D Experimental Details

D.1 Language Models

The details of the pre-trained LMs used in our experiments are provided in Table 8, including the number of parameters and pre-training data sources. The majority of models have been pre-trained using CommonCrawl data. Aya is based on mT5_{XXL} and further instruction-tuned using the Aya dataset (Singh et al., 2024). Training was performed using