



MULTILINGUAL AI DATA SERVICES

CUSTOM-BUILT SPEECH, TEXT & CONVERSATIONAL DATASETS AT SCALE

ABOUT US



CCC is a multilingual language service provider specializing in translation and localization, with 30+ supported languages. Building on this linguistic foundation, we provide audio, transcription, and translation datasets for AI, including culturally nuanced and code-switched languages such as *Taglish*, custom-built to each client’s use case, workflow, and scale.



OUR GOAL

To bridge human language and AI by transforming CCC’s expertise in translation, localization, and multilingual communication into high-quality, bespoke datasets that help AI systems understand diverse and underrepresented languages, cultures, and real-world contexts more naturally and accurately.

OUR LANGUAGE DATA SERVICES



SPEECH RECORDING SERVICES

Scripted, conversational, and meeting-style recordings for AI pretraining and fine-tuning, including podcast-style audio, Zoom sessions, virtual meetings, and studio recordings. For fine-tuning, we can produce spontaneous or scripted recordings, including high-quality 48 kHz audio.



AUDIO DATA TRANSCRIPTION AND QA

Human transcription, evaluation, and post-editing of AI-generated transcripts for existing audio datasets. Human intervention helps accurately capture slang, jargon, code-switching, code-mixing, accents, and context-dependent expressions that automated systems may miss or misinterpret. We also support annotation for speaker ID, emotion, noise, and other metadata.



CONVERSATIONAL AND CONTEXT-SPECIFIC TEXT GENERATION

Human-generated text datasets—including messages, emails, customer support scenarios, dialogues, and speech-recording scripts—for use cases requiring natural language, cultural nuance, and reliable ground truth. This data is especially valuable for multilingual and underrepresented languages, code-switching, model evaluation, and scenarios where synthetic data alone may not reflect real-world communication.



MACHINE TRANSLATION POST-EDITING AND EVALUATION

Human post-editing and evaluation of machine-translated and LLM-generated multilingual content for accuracy, fluency, terminology, cultural appropriateness, and instruction adherence. Our linguists can support preference ranking, error annotation, quality scoring, and human correction, providing reliable feedback data for model evaluation and improvement.

OUR SUPPORTED LANGUAGES



★ Asia-Pacific

Japanese
Korean
Tagalog
Cebuano
Thai
Bahasa Melayu
Bahasa Indonesia
Chinese (Traditional)
Chinese (Simplified)
Khmer
Nepali
Mongolian

Vietnamese
Burmese
Hindi
Tamil
Urdu
Bengali
★ Europe
Spanish (Spain)
Portuguese (Portugal)
Italian
French
German
Ukrainian

Russian
Greek
Belarusian
Slovak
Polish
Hungarian
Slovenian
Czech
English (UK)
★ Americas
Spanish (LatAm)
Portuguese (Brazil)
English (US)

★ Middle East, Caucasus & Africa

Armenian
Georgian
Persian
Turkish
Arabic
Afrikaans
Swahili
Kinyarwanda
Zulu

If your desired language is not listed, please feel free to contact us.

FLEXIBLE PRICING FOR EVERY PROJECT



From pilot datasets to large-scale multilingual programs, we tailor project scope and pricing to your requirements.

Projects are customized based on **language, volume, complexity**, recording specifications, annotation requirements, and turnaround time. Contact us for a FREE quotation.



FOR FURTHER INQUIRIES

www.cccjph.com | inquiry@cccjph.com | (082) 233 0032

Create a new story with us.

