

Creating Multilingual Contact Center Datasets with Synthetic PII, Machine Translation, and Expert QA

An ongoing project that's allowed us to stretch our skills was working with a large e-tailer looking to create synthetic multilingual datasets via machine translation. The goal was to produce a dataset of naturalistic contact center transcripts, covering both online chat and telephone calls in multiple European and Asian languages. The transcripts needed to include fabricated yet have realistic personally identifiable information (PII) that has to be marked and categorized by type, for example: names, addresses, phone numbers, etc.

The Challenge **Does this sound right in that language?**

To help us meet this challenge, of crafting custom transcripts, we were provided with a set of machine-translated conversations.

The original dialogues were in Spanish and were accompanied by corresponding English reference translations. Making the text seem natural, accurate, and with a target language native fluency – which includes dialect – made for an interesting challenge. Each PII instance would need to be tagged and documented to maintain data consistency and traceability across the dataset.

Solution **AIDA Hub (our proprietary Data Platform) and humans working together**

We configured our AIDA Hub to support the following high-level workflow components:

- **Conversation Editing:** Linguists edited each conversation to ensure it sounded natural and realistic within the target language customer service context, tailored to the relevant domain e.g. retail, tech support, finance, etc.
- **Localized PII Generation:** Artificial, culturally appropriate PII was generated to replace the non-localized PII data in each conversation.
- **PII Tagging and Labeling:** All PII spans were annotated and labeled by category based on provided annotation guidelines.

To ensure quality, we handpicked a highly skilled team of freelancers from our pool, rigorously tested and onboarded them using trial tasks. From there we expanded our team to include up to 20 expert linguists per language across 8 languages. This ensured we would meet the customer's turnaround time requirements.

In addition, we:

- Ensured multi-layered QA, with both human oversight and a full review by a second annotator for maximum accuracy.
- Developed and implemented a suite of real-time and offline automated validation tools to ensure data quality and integrity at every stage of the process.
- To support the team, we created training videos covering tips and tricks and also showed the team how to correctly leverage the pre-generated fictitious PII for their respective market.
- Our Python-based automation framework included:
 - Custom language-specific validation scripts to ensure fake PII data adhered to local market conventions. These were typically executed before final post-processing and delivery.
 - Tailored scripts to detect and fix spacing issues, particularly in Asian languages.
 - Consistency checks for speech and chat datasets, ensuring numerical representations e.g. “9” in chat vs. “Nine” in speech, matched expected formats.
 - Over 100 real-time quality control checks, covering a wide range of validations including:
 - Proper formatting of synthetic PII data
 - Detection of content inconsistencies
 - Integration of customer-supplied JSON validation scripts to enforce structural integrity and flag issues such as extra spaces, unauthorized tags, or invalid attribute values prior to delivery.

Result

A successful first project

The initial language batches were successfully delivered with exceptional quality and rapid turnaround, which led to expansion into additional languages and similarly scoped projects.

About us

Our Data Services

We specialize in:

- Data collection & enhancement
- Annotation & labeling
- Evaluation
- Data management
- Benchmarking & model evaluation
- Model training & fine tuning

Exceptional AI outcomes begin with dependable, high-quality data. Backed by years of experience and a diverse global talent network, our AI data services are designed to give you confidence.

Learn more about our services on our Data Services page. Or contact us below to see how you can be our next success story.

datamundi.ai/ai-data-services