Prompsit's API and CLI: planet-friendly, privacy-first, open-source translation services for everyone

Lev Nikolaevich Berezhnoy, Gema Ramírez Sánchez
Sergio Ortiz Rojas, Mikel L. Forcada
Prompsit Language Engineering
Edif. Quorum III, Avinguda de la Universitat, s/n, E-03202 Elx
levnikolaevich,gramirez,sergio,mlf@prompsit.com

Abstract

Prompsit is launching an updated API and CLI for its open-source, planet-friendly machine translation services. Operating on a freemium model, the tools offer free limited access alongside tiered pricing for advanced features like MT evaluation, quality estimation, corpus scoring, and multilingual dataset annotation.

1 A classical MT service in 2026?

While large language models (LLMs) offer impressive linguistic nuance, traditional neural machine translation (NMT) remains the proven backbone for professional workflows. NMT is significantly faster —often outperforming LLMs by a large margin— and utilises a transparent, character-based pricing model. By removing the token overhead associated with prompting, NMT provides a more stable and cost-effective solution for both small and large-scale projects. This efficient approach inspires Prompsit's translation services, offering significant sustainability advantages by using purpose-built NMT engines that require a fraction of the computational power and energy needed by general-purpose LLMs. Alongside lean open-source NMT models built from well-curated corpora provided by OPUS (github.com/Helsinki-NLP/Opus-MT) and Mozilla (github.com/mozilla/firefox-translations-models), we provide high-quality Apertium machine translation (apertium.org) and AltLang language variety converters (altlang.net), offering the stable, predictable behaviour of rule-based systems (RBMT) that are even faster and more energy-efficient. Furthermore, we complement these translation engines with automatic evaluation and annotation services.

2 Services available

Translation We offer high-performance NMT and RBMT specializing in low-resource languages and regional variants to ensure contextually accurate output. The API supports text snippet and document translation across a wide range of languages and formats, including robust tag handling, optional quality estimation and leverage from a hierarchy of user's translation memories.

Evaluation Our tools measure translation quality using industry-standard automated metrics, allowing users to audit engines by analysing parallel corpora and model performance. This helps maintain professional standards and linguistic consistency across supported language pairs.

Scoring Parallel segments can be scored for translation likelihood using Prompsit's widely-adopted Bicleaner multilingual models (github.com/bitextor/bicleaner-ai). These scores are used to identify and filter low-quality translations, to help select higher-quality parallel data for model fine-tuning.

Annotation The API provides sophisticated data processing to deduplicate, label, and score multilingual datasets. Documents are enriched with language identification, personally identifiable information (PII) and adult content flagging, encoding fixes, and quality scores. This metadata enrichment is essential for top document selection in model refinement tasks.

Shterionov, Vanmassenhove, De Sisto, Blain, Pourmostafa, Lepp, Manna, Rescigno, Karakanta, Rigouts Terryn, Lardelli, Resende, Murgolo, Hackenbuchner, Zaretskaya, Esplà-Gomis, Etchegoyhen, Gromann, Bawden, Haddow, Szoc, Forcada, Moniz (Eds.)
Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026), p. 49–50
Tilburg, The Netherlands, June 2026.

3 The API and CLI

Access to Prompsit's API is via an access token available to registered users. Here's a curl request to translate a short string:

```
curl -X 'POST' \
'https://edge.prompsit.com/v1/translation?enable
_qe=false' \
  -H 'accept: application/json' \
  -H 'Authorization: Bearer ...auth_token... \
  -H 'Content-Type: application/json' \
  -d '{
  "source_lang": "en",
  "target_lang": "es",
  "texts": [
    "Hello world"
  ]
}'
```

To translate a file, one would use a similar call which would return the URLs needed to check status and to download the result file. Currently customers can directly invoke the API from their internal tools and platforms with simple integration steps, or use the CLI described below. Prompsit plans to offer new CAT connectors to implement the translation services offered in the new API.

The CLI provides easier API access for human users but may also be used to easily script complex translation- related tasks. For instance, the command for the translation query inside the CLI above would simply be `translate "Hello world" -s "en" -t "es"`. From outside the CLI, a script could send `prompsit translate "Hello world" -s "en" -t "es"` and capture the result for further processing. The CLI is available under the Apache 2.0 licence at `github.com/Prompsit/prompsit-cli` for our customers to install locally. The underlying engines and most models are also open-source.

4 A bit of technological detail

Built as a REST application on top of Python 3.13 and FastAPI (`fastapi.tiangolo.com`), our API utilizes a microservice architecture to orchestrate 12 containerised modules that power several translation engines such as Apertium, AltLang, and CTranslate2 (`github.com/opennmt/ctranslate2`). An 8-step pipeline manages tag extraction, 5-level caching, and neural word alignment while a specialized formatting stack (Docling, Okapi, and Tikal) handles over 25 binary and text formats. MetricX (`github.com/google-research/metricx`) and COMET (`unbabel.github.io/COMET`) GPU-based estimation are used to ensure quality while Bicleaner-AI and Monotextor (`github.com/bitextor/monotextor`) provide respectively advanced parallel (sentence pairs) and monolingual corpus (documents) scoring and annotation. Asynchronous job progress is streamed in real-time via server-sent events, all accessible through an open-source CLI.

5 A summary of features

Energy efficiency: quantized NMT engines and microservices save GPU and power usage.

Data privacy: in-memory processing ensures no data storage or use for model training.

Latency: intelligent caching allows for millisecond responses and real-time document progress via streaming.

Language coverage: a selection of NMT and RBMT engines for 17 major and 3 low-resource languages (ca, gl, nn), 11 language varieties for 5 of them (such as fr-CA and fr-FR) in 52 language pairs as of May 2026.

Format support: tag-aware translation for 30 different formats, including Office, PDF, and localization file formats (such as TMX or PO).

Transparency: transparent commands for usage and health monitoring.

6 Use via an AI agent

The CLI repository includes machine-readable skill descriptions that enable most popular AI coding assistants to assist the human user to interact with the CLI programmatically to perform translation, evaluation, scoring, annotation, and initial setup. Skills are bundled with the CLI package and deployed automatically on first launch. This integration is a thin interface layer: the AI assistant interprets user intent and invokes CLI commands. The computational cost of translation services is the same regardless of whether the request originates from a human or an AI assistant.

7 Access and pricing

Visit `prompsit.com/en/contact` for free API access. A secret key will be sent to your email. Install the CLI with `npm install -g prompsit-cli` and authenticate with the provided login, or using your Google address. We offer a freemium pricing model, mostly free (with limits) for MT and paid for additional services.