

## Hybrid Written Language Translator for Offline Text Handling (Language A => English)

**Version 1.0** | December 2025

Github: [@profdilley](#) | Created by Prof LC Dilley, PhD with assistance from Claude Opus 4.5 [v. Desktop, Thinking Mode) and Claude Code, & Perplexity — 2025-12-02

THOTH is a powerful, privacy-first translation tool designed for professionals who need to translate CSV data containing multiple languages—completely offline, with no data ever leaving your machine.

Named in honour of the Egyptian god of writing, language, and knowledge, THOTH brings enterprise-grade translation capabilities and dual state-of-the-art translation engines to your environment for local compute.

Feature	Description
100% Offline	All translation happens locally. No cloud services, no API calls, no data transmission.
38 Languages	Full support for Russian, Ukrainian, Baltic, Balkan, European, Nordic, East Asian, and Middle Eastern languages.
Dual Translation Engines	Two industry-leading engines for maximum quality and coverage.
Smart Language Detection	Automatic per-cell language detection—no manual configuration required.
Mixed-Language Support	A single column can contain text in multiple languages; THOTH handles each cell independently.
Adjacent Column Output	Translated columns appear immediately next to their source columns for easy comparison.

THOTH provides two translation engines, giving you flexibility and redundancy:

🌟 Engine #1: NLLB-200 (Default) 🌟

- **Model:** Meta's No Language Left Behind (NLLB-200-distilled-600M)
- **Coverage:** 200 languages

- **Strengths:** State-of-the-art neural translation, excellent for low-resource languages (Baltic, Balkan), superior handling of Cyrillic scripts
- **Size:** ~2.5 GB

### ✨ Engine #2: Argos Translate (Alternative) ✨

- **Model:** Open-source neural machine translation
- **Coverage:** 38 language pairs to English
- **Strengths:** Lightweight, fast inference, strong Western European performance
- **Size:** ~1.5 GB (all language packs)

### Switch between engines anytime:

```
python thoth.py input.csv --engine nllb    # Default
python thoth.py input.csv --engine argos   # Alternative
```

---

## ✨✨✨🌍 Supported Languages ✨✨✨

### Slavic Languages

Russian, Ukrainian, Belarusian, Polish, Czech, Slovak, Bulgarian, Serbian, Croatian, Bosnian, Slovenian, Macedonian

### Baltic Languages

Lithuanian, Latvian, Estonian

### Nordic Languages

Swedish, Norwegian, Danish, Finnish, Icelandic

### Western European Languages

German, French, Spanish, Portuguese, Italian, Dutch, Greek, Romanian, Hungarian

### East Asian Languages

Mandarin Chinese, Cantonese, Japanese, Korean

### Middle Eastern Languages

Arabic, Hebrew, Turkish, Persian (Farsi)

### Additional Languages

Albanian, and more...

---

```

////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////
////////
////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////
////////
////
////
////
////                               ✨ ✨ < QUICK START < ✨ ✨
////
////
////
////
////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////
////////
////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////
////////

```

Step 1: Create Virtual Environment

```

cd thoth-translator
python3 -m venv venv
source venv/bin/activate

```

Step 2: Install Dependencies

```

pip install -r requirements.txt

```

Step 3: Download Translation Models

```

python -m translator.setup --download-models

```

What gets downloaded:

Component	Size	Purpose
fastText LID218	~130 MB	Language detection model
NLLB-200-distilled-600M	~2.5 GB	Primary translation engine
Argos language packs	~1.5 GB	Secondary translation engine

Total download: approximately 4 GB. Requires internet connection for this step only.

Step 4: Translate Your Data

```

python thoth.py your_file.csv --columns "column1,column2,column3"

```

**Output:** `your_file_translated.csv` with `_en` columns adjacent to originals.

////////////////////////////////////  
/////  
////////////////////////////////////  
/////

---

🌟🌟🌟 📖 Usage Examples 🌟🌟🌟

Basic Translation (Auto-detect all text columns)

python thoth.py data.csv

Specify Columns to Translate

python thoth.py data.csv --columns "description,notes,comments"

Use Alternative Engine

python thoth.py data.csv --engine argos

Force Source Language (Skip auto-detection)

python thoth.py data.csv --force-lang rus\_Cyrl

Custom Output Filename

python thoth.py data.csv --output translated\_data.csv

---

🌟🌟🌟 🎯 Smart Features 🌟🌟🌟

Per-Cell Language Detection

Unlike traditional translation tools that require one language per column, THOTH analyzes **each cell independently**.

Example input:

id	comment
1	Отличный продукт!
2	Чудовий сервіс
3	Świetna obsługa
4	素晴らしい品質

THOTH output:

id	comment	comment_en
1	Отличный продукт!	Excellent product!
2	Чудовий сервіс	Excellent service
3	Świetna obsługa	Great service
4	素晴らしい品質	Excellent quality

Russian, Ukrainian, Polish, and Japanese—all handled automatically in a single column.

Adjacent Column Layout

Translated columns are inserted immediately after their source columns:

✔ THOTH output:

id | description | description\_en | notes | notes\_en | country

✖ Other tools:

id | description | notes | country | description\_en | notes\_en

🌟📁 Supported File Formats 🌟🌟🌟

Format	Extension	Notes
CSV	.csv	UTF-8, UTF-8-BOM, Latin-1, CP1252 auto-detected
Excel	.xlsx	Full support
Excel (Legacy)	.xls	Full support

🌟🌟🌟⚙️ Configuration 🌟🌟🌟

For advanced users, THOTH supports YAML configuration:

```
# config.yaml
translation:
  default_engine: nllb
  target_language: eng_Latn

detection:
  confidence_threshold: 0.7
  fallback_language: eng_Latn

column_defaults:
  skip_numeric: true
  skip_dates: true
  skip_english: true
  skip_empty: true
  auto_select_foreign_text: true
```

---

### 🌟🌟🌟🔧 Verify Installation 🌟🌟🌟

Run the test suite to confirm everything is working:

```
python thoth.py --test
```

Expected output: 21 passed

---

### 🌟🌟🌟🔒 Privacy & Security 🌟🌟🌟

- **Zero network transmission:** After model download, THOTH never connects to the internet
- **No telemetry:** No usage data, analytics, or logging to external services
- **Local processing:** All translation happens on your CPU/GPU
- **Open source:** Full source code included for audit

---

### 🌟🌟🌟💻 System Requirements 🌟🌟🌟

Requirement	Minimum	Recommended
Python	3.10+	3.11+
RAM	8 GB	16 GB
Disk Space	6 GB	10 GB
OS	macOS, Linux, Windows	Apple Silicon optimized

---

### 🌟🌟🌟🆘 Troubleshooting 🌟🌟🌟

"No module named '\_tkinter'" (GUI mode)

The GUI requires tkinter. Use CLI mode instead:

```
python thoth.py input.csv --columns "col1,col2"
```

Model download stalls

Cancel (Ctrl+C) and restart:

```
python -m translator.setup --download-models
```

Completed downloads are cached and won't re-download.

NumPy compatibility error

```
pip install "numpy<2.0"
```

Memory errors on large files

Reduce batch size in config.yaml:

```
performance:
  batch_size: 8 # Default is 16
```

---

🌟🌟🌟 🇮🇹 Performance Benchmarks 🌟🌟🌟

*Tested on Apple M3, 16 GB RAM*

Dataset Size	Columns	Time (NLLB)	Time (Argos)
1,000 rows	2 columns	~30 sec	~20 sec
10,000 rows	2 columns	~5 min	~3 min
10,000 rows	10 columns	~25 min	~15 min

---

🌟🌟🌟 📄 License 🌟🌟🌟

Open source for private use.

---

🌟🌟🌟 🙏 Acknowledgments 🌟🌟🌟

🌟🌟🔧 **THOTH Language Translator** 🌟🌟 is built on the shoulders of giants:

- **Meta AI** — NLLB-200 translation model
  - **Argos Open Tech** — Argos Translate
  - **Facebook Research** — fastText language identification
  - **Hugging Face** — Transformers library
  - **Perplexity.AI** — Tech SOTA Consulting
  - **Prof George Lakoff** — Language Cognitive Neuroscientist & Influencer on [@profdilley](#)
  - **Prof Claude Shannon** — Father of Information Theory & Influencer on [@profdilley](#)
  - **Claude by Anthropic** — AI Collaborator with [@profdilley](#) on 🌟🌟🔧 **THOTH Language Translator**  
🌟🌟
- 

*Built with care by and for 🌟🌟 **professionals** 🌟🌟 who value 🌟 **privacy** 🌟 and 🌟 **precision**. 🌟*

Contact Author **Prof LC Dilley, PhD** on Github: [@profdilley](#)

🌟🌟🌟🔧 **THOTH Language Translator** 🌟🌟🌟 | 🌟 "Your words, your machine, your control." 🌟