✨✨🔧 THOTH Language Translator ✨

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*

✨ About 🔧 THOTH Language Translator

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.

✨ Key Features

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.

🖥️ Two Ways to Use THOTH

Use THOTH with a Graphical User Interface (GUI) or Command Line Interface (CLI).

Mode	Command	Best For
GUI	`python thoth.py --gui`	Visual column selection, previewing translations

Mode	Command	Best For
CLI	`python thoth.py input.csv`	Scripting, automation, batch processing

Both modes offer the same translation quality — choose based on your workflow.

✨ 🔧 Dual-Engine Architecture

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
```

When to Use Each Engine (Updated with Benchmark Data)

Engine	Best For	Benchmark Result	Speed
NLLB (default)	All languages, especially Ukrainian, Japanese, Korean, Arabic, Spanish, German, and all Baltic/Balkan	chrF 61.15 (mean)	Baseline
Argos	Russian, French, Polish, Czech, Greek, Chinese when speed matters	chrF 56.77 (mean)	14× faster

Critical: Argos does **not** support Lithuanian, Latvian, Estonian, Serbian, Bulgarian, or Norwegian. Use NLLB for these languages.

✨ 🌍 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...

✨ 📋 Language Codes Reference

Language	NLLB Code	Argos Code
Arabic	ara_Arab	ar
Bulgarian	bul_Cyrl	bg
Chinese (Simplified)	zho_Hans	zh
Croatian	hrv_Latn	hr
Czech	ces_Latn	cs
Danish	dan_Latn	da
Dutch	nld_Latn	nl
English	eng_Latn	en
Estonian	est_Latn	et
Finnish	fin_Latn	fi
French	fra_Latn	fr

Language	NLLB Code	Argos Code
German	deu_Latn	de
Greek	ell_Grek	el
Hebrew	heb_Hebr	he
Hungarian	hun_Latn	hu
Icelandic	isl_Latn	—
Italian	ita_Latn	it
Japanese	jpn_Jpan	ja
Korean	kor_Hang	ko
Latvian	lvs_Latn	lv
Lithuanian	lit_Latn	lt
Norwegian	nob_Latn	nb
Persian	pes_Arab	fa
Polish	pol_Latn	pl
Portuguese	por_Latn	pt
Romanian	ron_Latn	ro
Russian	rus_Cyrl	ru
Serbian	srp_Cyrl	sr
Slovak	slk_Latn	sk
Slovenian	slv_Latn	sl
Spanish	spa_Latn	es
Swedish	swe_Latn	sv
Turkish	tur_Latn	tr
Ukrainian	ukr_Cyrl	uk

...

✨✨✨✨ ⚡ 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
```

📦 Batch Processing

Translate multiple files at once:

All Files in Current Directory

```
python thoth.py --batch --columns "description,notes"
```

All Files in a Specific Directory

```
python thoth.py --batch-dir /path/to/data --columns "description,notes"
```

All Files in Directory and Subdirectories (Recursive)

```
python thoth.py --batch-recursive --columns "description,notes"
```

Batch with Target Language

```
python thoth.py --batch-dir ./data --columns "description" --target-lang
fra_Latn
```

Notes:

- Supports `.csv`, `.xlsx`, and `.xls` files
- Files with `_translated` in the name are automatically skipped
- Each output file is saved alongside its source with `_translated` suffix

---

✨ 🎯 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:

```yaml
# 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.

Option #1: Install it:

- **macOS**: `brew install python-tk@3.13` (match your Python version)
- **Ubuntu/Debian**: `sudo apt-get install python3-tk`
- **Windows**: Reinstall Python and check "tcl/tk" option

Option #2: 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

✨ 🔬 Translation Quality Benchmarks (FLORES+ Validation)

THOTH's translation engines were rigorously evaluated against the **FLORES+** benchmark dataset—the gold standard for multilingual translation evaluation used by Meta to benchmark NLLB-200.

Evaluation Summary

Metric	NLLB-200	Argos Translate
Languages Evaluated	18	12
Mean chrF Score	**61.15**	56.77
Mean BLEU Score	**34.95**	28.81
Translation Speed	Baseline	14× faster

chrF (character n-gram F-score) is the primary metric—higher is better. Tested on 200 sentences per language from FLORES+ devtest split.

Results by Language Family

Slavic Languages

Language	NLLB chrF	Argos chrF	Recommended Engine
Russian	60.3	**62.5**	Either (Argos slightly better)
Ukrainian	**62.1**	50.7	**NLLB** (+11.4)

Language	NLLB chrF	Argos chrF	Recommended Engine
Polish	56.6	56.0	Either (similar quality)
Czech	63.9	62.9	Either (similar quality)
Bulgarian	**66.1**	N/A	**NLLB** (Argos unavailable)
Serbian	**65.6**	N/A	**NLLB** (Argos unavailable)

Baltic Languages

Language	NLLB chrF	Argos chrF	Recommended Engine
Lithuanian	**57.6**	N/A	**NLLB** (Argos unavailable)
Latvian	**58.7**	N/A	**NLLB** (Argos unavailable)
Estonian	**61.9**	N/A	**NLLB** (Argos unavailable)

East Asian Languages

Language	NLLB chrF	Argos chrF	Recommended Engine
Chinese (Simplified)	56.1	54.6	Either (similar quality)
Japanese	**53.0**	46.4	**NLLB** (+6.6)
Korean	**55.4**	45.6	**NLLB** (+9.8)

Western European & Other

Language	NLLB chrF	Argos chrF	Recommended Engine
French	69.4	69.2	Either (similar quality)
German	**67.0**	64.2	**NLLB** (+2.8)
Spanish	**59.9**	53.9	**NLLB** (+6.0)
Arabic	**63.8**	57.2	**NLLB** (+6.6)
Greek	59.5	58.1	Either (similar quality)
Norwegian	**63.9**	N/A	**NLLB** (Argos unavailable)

Key Findings

1. **NLLB is the recommended default engine** — Higher quality across 17 of 18 languages tested
2. **Argos coverage gaps** — Does not support Baltic (LT, LV, ET), most Balkan (SR, BG), or Norwegian
3. **Argos wins only for Russian** — Slight advantage (+2.2 chrF)

4. **Speed vs Quality trade-off** — Argos is 14× faster; acceptable for FR, PL, CZ, EL, ZH when speed matters

Engine Selection Guide

```
┌─────────────────────────────────────────────────────────────────┐
│                    WHICH ENGINE SHOULD I USE?                   │
├─────────────────────────────────────────────────────────────────┤
│                                                                 │
│  Is your source language Baltic, Balkan, or Norwegian?          │
│     YES → Use NLLB (Argos doesn't support these)                │
│     NO  ↓                                                       │
│                                                                 │
│  Is your source language Japanese, Korean, Arabic, or Spanish?  │
│     YES → Use NLLB (significantly better quality)               │
│     NO  ↓                                                       │
│                                                                 │
│  Is your source language Ukrainian or German?                   │
│     YES → Use NLLB (better quality)                             │
│     NO  ↓                                                       │
│                                                                 │
│  Is speed critical and source is FR/PL/CZ/EL/ZH/RU?             │
│     YES → Argos acceptable (14× faster, similar quality)        │
│     NO  → Use NLLB (default, best overall)                      │
│                                                                 │
└─────────────────────────────────────────────────────────────────┘

---

## ✨ 📝 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](https://github.com/profdilley)
- **Prof Claude Shannon** — Father of Information Theory & Influencer on
[@profdilley](https://github.com/profdilley)
- **Claude by Anthropic** — AI Collaborator with [@profdilley]
(https://github.com/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]
(https://github.com/profdilley)

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