TALOS Software

"Text File Analyser"

Installation and Usage Guide

Contents

1. Software Identification

- **Name**: Text File Analyzer (TFA)
- **Program**: Talos_Text_Analyzer.py
- **Version**: 2.0
- **Date**: 14/08/2025
- **Language**: Python 3.12
- **Type**: Software (standalone)
- **Topic**: Textual Analysis Tool
- **Web site**: http://talos-ai4ssh.eu/File_Analyser/
- **Author**: Christophe Roche

2. General Description

Text File Analyser (TFA) is a standalone text analysis tool written in Python. It is designed to extract statistical and linguistic information (tokens and named entities) from textual documents and to export the results in Excel or CSV formats.

3. Core Functionalities

- Word (token) extraction and frequency analysis
- Named-Entity Recognition extraction and frequency analysis
- Lexico-Syntactic Pattern extraction
- Export in CSV or Excel format

4. Installation & Launch

4.1. System Requirements

- **Operating System**: Windows 10/11, macOS, or Linux
- **RAM**: Minimum 4GB, recommended 8GB+
- **Disk Space**: At least 2GB free space for Python and libraries
- **Internet Connection**: Required for initial setup and model downloads

4.2 Install Python

Windows:

1. Download Python 3.8+ from https://www.python.org/downloads/
2. **IMPORTANT**: During installation, check **"Add Python to PATH"**
3. Verify installation by opening Command Prompt and typing: python --version

macOS:

1. Install using Homebrew (recommended): `brew install python`
   Or download from https://www.python.org/downloads/

Linux (Ubuntu/Debian):

```
sudo apt update
sudo apt install python3 python3-pip
```

4.3 Install Required Python Libraries

Open Terminal/Command Prompt and run the following commands:

Core Libraries (Required)

`tkinter`: for Python 3 `tkinter` should already be included with your Python installation. However, if you don't have tkinter installed, the easiest way is to reinstall Python from the official website and ensure that the tkinter package is included during installation.

```
pip install pandas openpyxl spacy langdetect
```

Install spaCy Language Models

```
# English (Essential for TALOS - always install this one)
python -m spacy download en_core_web_sm
# Greek (Essential for TALOS - always install this one)
python -m spacy download el_core_news_sm    # Greek

# Optional - Install other languages as needed:
python -m spacy download fr_core_news_sm    # French
python -m spacy download es_core_news_sm    # Spanish
python -m spacy download de_core_news_sm    # German
python -m spacy download it_core_news_sm    # Italian
python -m spacy download pt_core_news_sm    # Portuguese
python -m spacy download nl_core_news_sm    # Dutch
```

Install Stanza (for Ancient Greek)

Text_File_Analyser uses **Stanza** for processing Ancient Greek texts.

First install Stanza:

```
pip install stanza
```

Then, open Python and download the Ancient Greek (grc) model (only once):

```
python
>>> import stanza
>>> stanza.download("grc")
```

This will install the **Ancient Greek Universal Dependencies models** (grc_perseus and grc_proiel). After this initial download, TALOS will automatically load the Ancient Greek model when needed.

4.4 Download the Program TALOS_Text_Analyzer.py

- Save the Talos_Text_Analyzer.py file to your computer
- Choose a dedicated folder (e.g., Documents/TALOS/Text_Analyser)

4.5 Launch the Program TALOS_Text_Analyzer.py

Method 1 - Command Line:

```
cd /path/to/your/TALOS/Text_Analyser
python Talos_Text_Analyzer.py
```

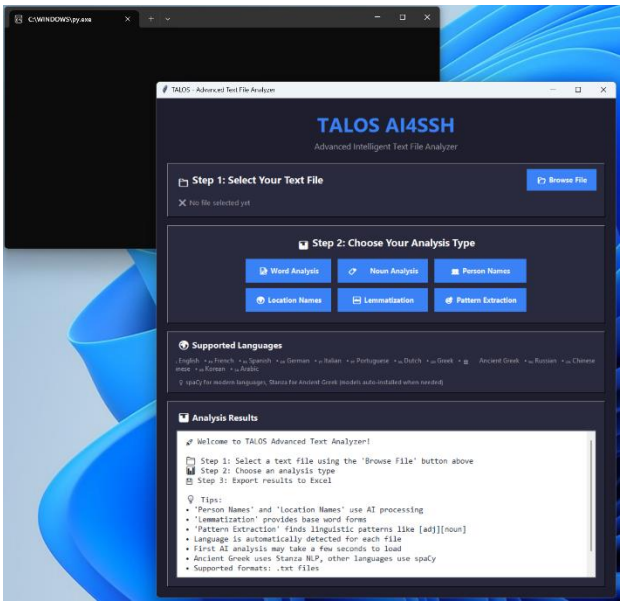
Method 2 - Double-click (Windows):

- Right-click on Talos_Text_Analyzer.py
- Select "Open with" → "Python"

Method 3 - Create a Shortcut (Windows):

- Create a .bat file with:

```
@echo off
cd "C:\path\to\your\TALOS\folder"
python TALOS_Text_Analyzer.py
pause
```



The Windows console opens, and after some time the application window is displayed.

The first analyses of the program can take a long time, corresponding to the loading of the language models into memory.

4.5 Troubleshooting

Common Issues:

1. "Python is not recognized" - Solution: Reinstall Python and check "Add Python to PATH"

2. "Module not found" errors - Solution: Run `pip install [module_name]` for each missing module

3. "spaCy model not found" - Solution: Run `python -m spacy download en_core_web_sm`

4. tkinter not found (Linux) - Solution: `sudo apt-get install python3-tk`

5. Export issues - Ensure you have write permissions in the export folder - Try exporting to Desktop first

6. Check Dependencies: Ensure all required libraries are installed

7. Update Libraries: Run pip install --upgrade [library_name]
8. File Format: Ensure your text file is properly encoded (UTF-8)
9. Memory: Try with smaller files first to test the installation

Performance Tips:

- **First AI Analysis:** The first time you use Person Names, Location Names, or Lemmatization, it may take 10-30 seconds to load the AI model
- **Large Files:** Files over 10MB may take longer to process
- **Memory:** Close other applications when processing very large text files

5. Updates

To update Talos_Text_Analyzer.py :

1. Replace the Talos_Text_Analyzer.py file with the new version
2. Update dependencies: pip install --upgrade pandas openpyxl spacy langdetect
3. Update language models: python -m spacy download en_core_web_sm --upgrade

6. How to Use TALOS_Text_Analyzer.py

The next chapter, dedicated to a use case, details each of the software's functions

6.1 Language Support

TALOS automatically detects the language of your text files and supports:

- US **English** (always available)
- FR French
- GR Greek
- DE German
- IT Italian
- ES Spanish
- NL Dutch
- PT Portuguese
- RU Russian
- CN Chinese
- JP Japanese
- KR Korean
- SA Arabic

Note: Advanced AI features (Person Names, Location Names, Lemmatization) require language-specific models. Install them as needed using the spaCy commands above.

6.2 Basic Workflow:

1. 📁 **Step 1:** Click "🔍 Browse File" to select your .txt file
2. 📊 **Step 2:** Choose your analysis type:
   - 📝 **Word Analysis:** Frequency of all words
   - 🏷️ **Noun Analysis:** Extract only nouns
   - 👥 **Person Names:** AI-powered person name recognition
   - 📍 **Location Names:** AI-powered location identification
   - 🔤 **Lemmatization:** Base forms of words
   - 🎯 **Pattern Extraction:** Custom linguistic patterns
3. 💾 **Step 3:** Export results to Excel or CSV

6.3 Supported File Formats:

- **.txt files** (UTF-8 encoding recommended)
- Maximum recommended size: 50MB

6.4 Export Options:

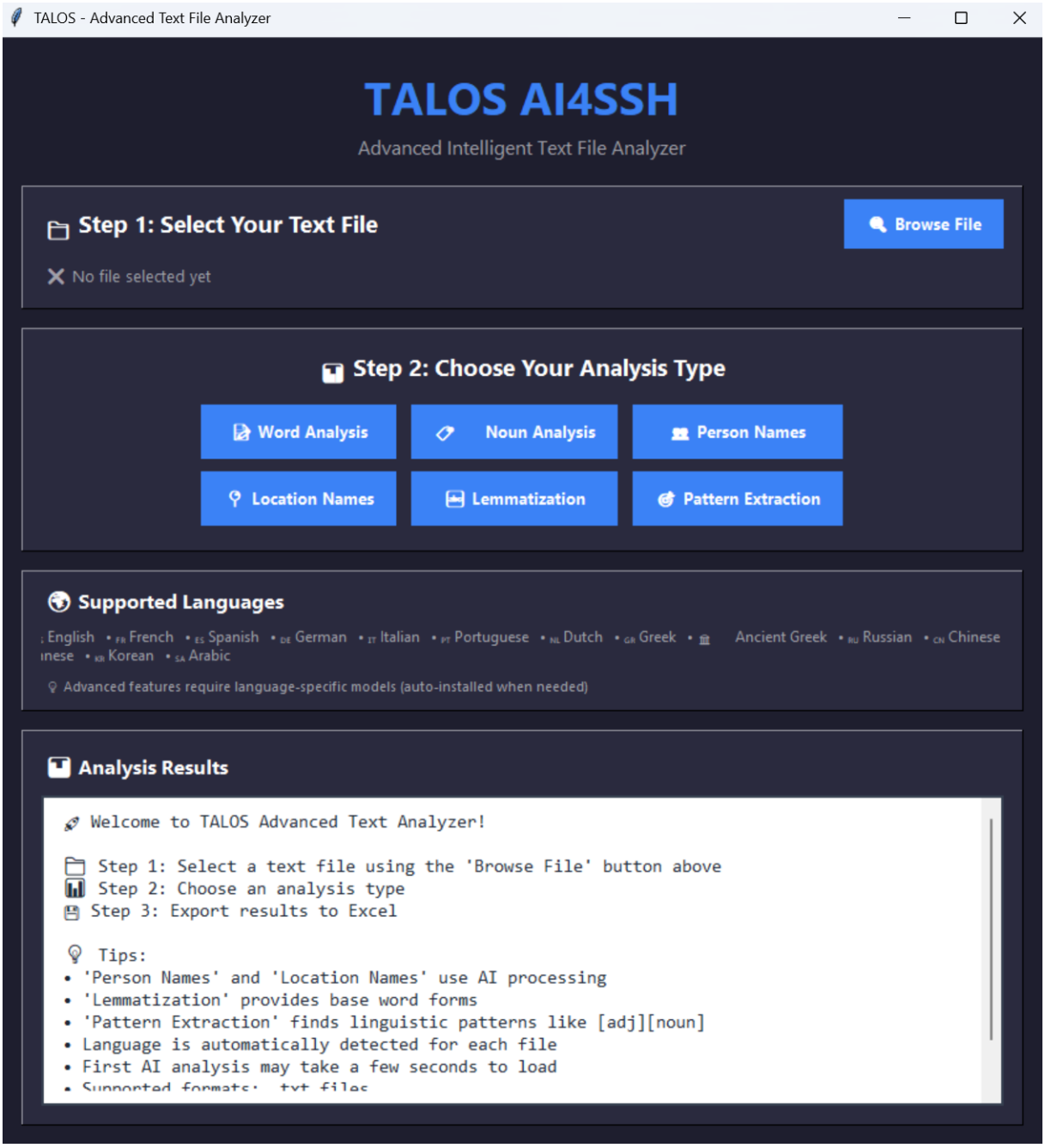
- 📊 **Excel (.xlsx):** Multi-sheet with statistics
- 📄 **CSV (.csv):** Simple comma-separated values

7. Use Cases

The first time you use Person Names, Location Names, or Lemmatization, it may take 10-30 seconds to load the AI model

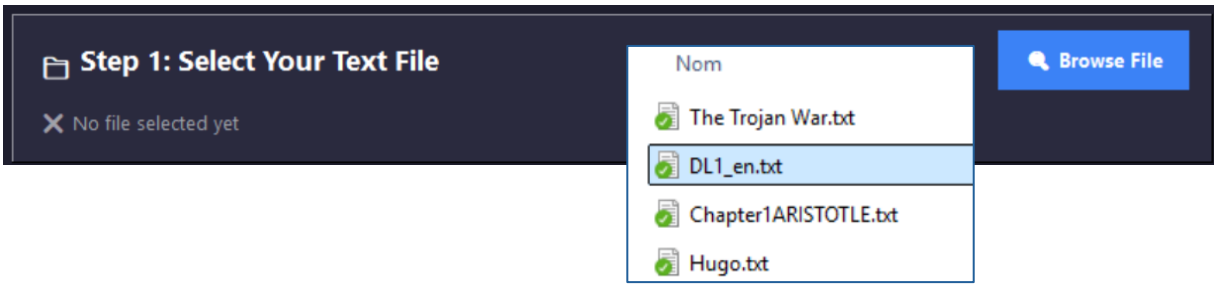
7.1 Interface

The main interface is divided into 3 main panels corresponding to the selection of the file to analyse (Step 1), the 6 analysis types proposed by the software (Step 2) and the Analysis Results panel.



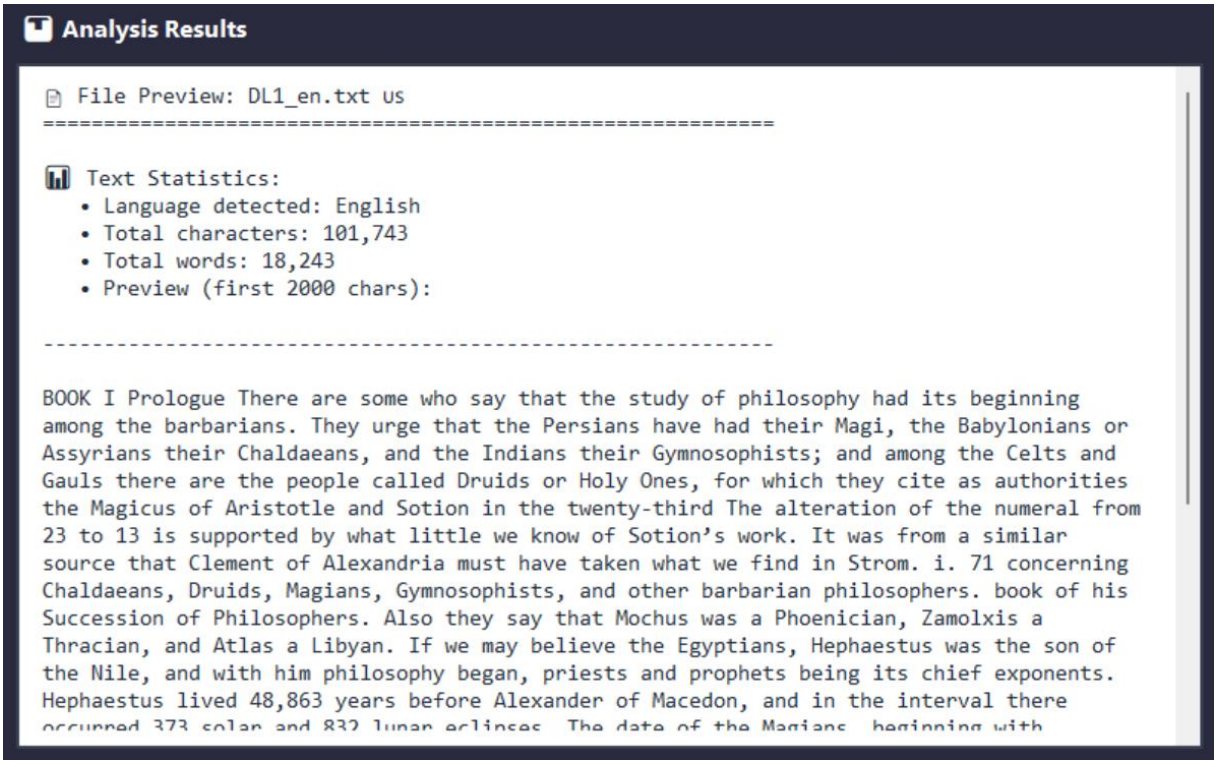
7.2 Selecting a text to analyse

The first step consists of selecting the file to analyse.



The "Select Your Text File" panel is updated to display information about the uploaded file.

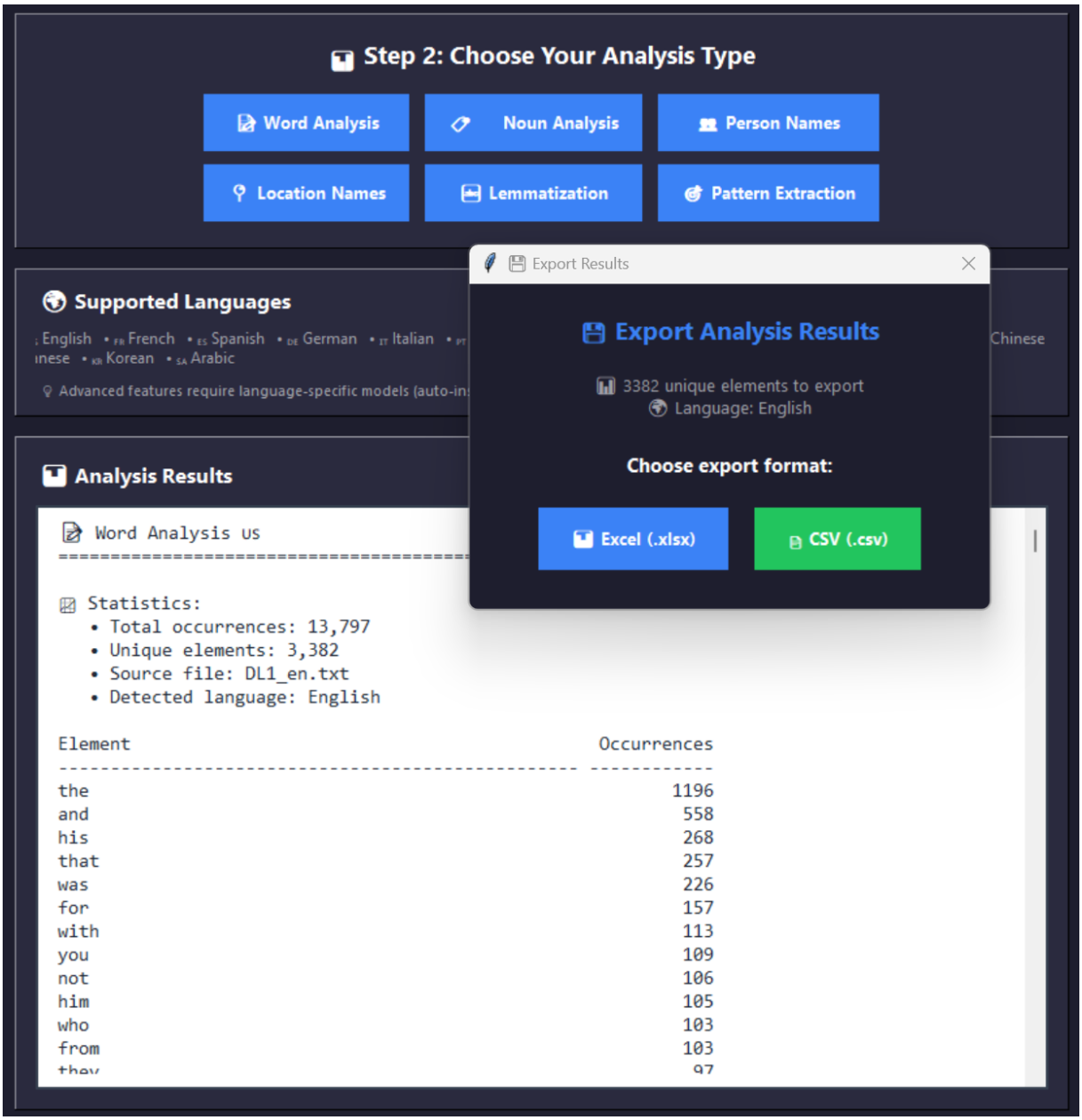
The Analysis Results panel displays some statistics and a preview of the uploaded file.



7.3 Word Analysis

The Word Analysis function counts all words in the text, listing each unique word and the number of times it appears.

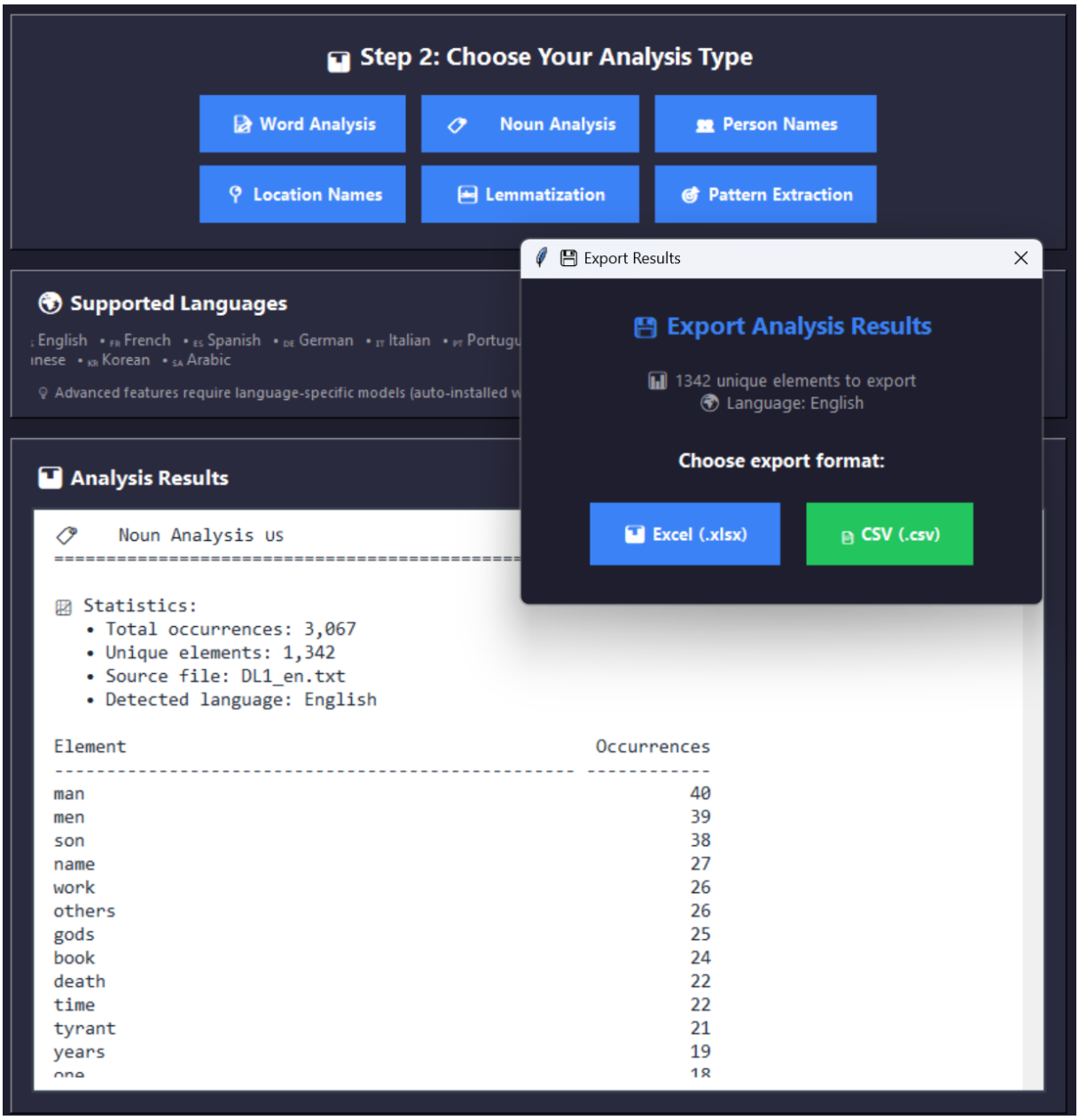
It ignores very short tokens (less than 3 letters) and normalizes text to lowercase for consistent counting. This provides a full statistical overview of the vocabulary distribution. The results can be exported in Excel or CSV formats.



7.4 Noun Analysis

The Noun Analysis function identifies and counts only common nouns, showing their frequency in the text.

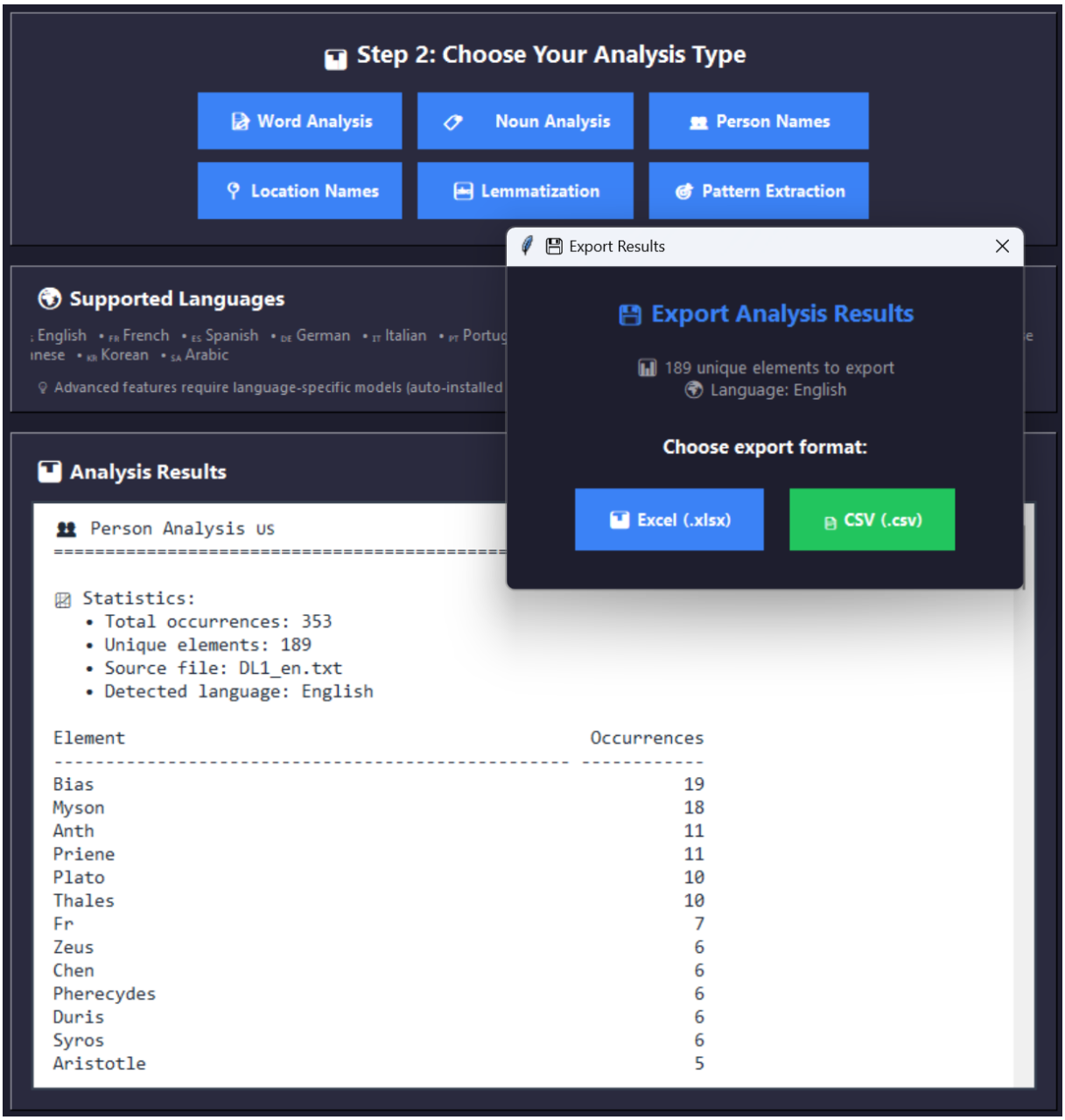
It uses part-of-speech tagging to ensure only grammatical nouns are included, filtering out proper names and other word types. This helps focus on key objects, concepts, or entities in the corpus. The results can be exported in Excel or CSV formats.



7.5 Person Names

The Person Names analysis detects and counts named entities of type “PERSON” (people’s names) using NLP.

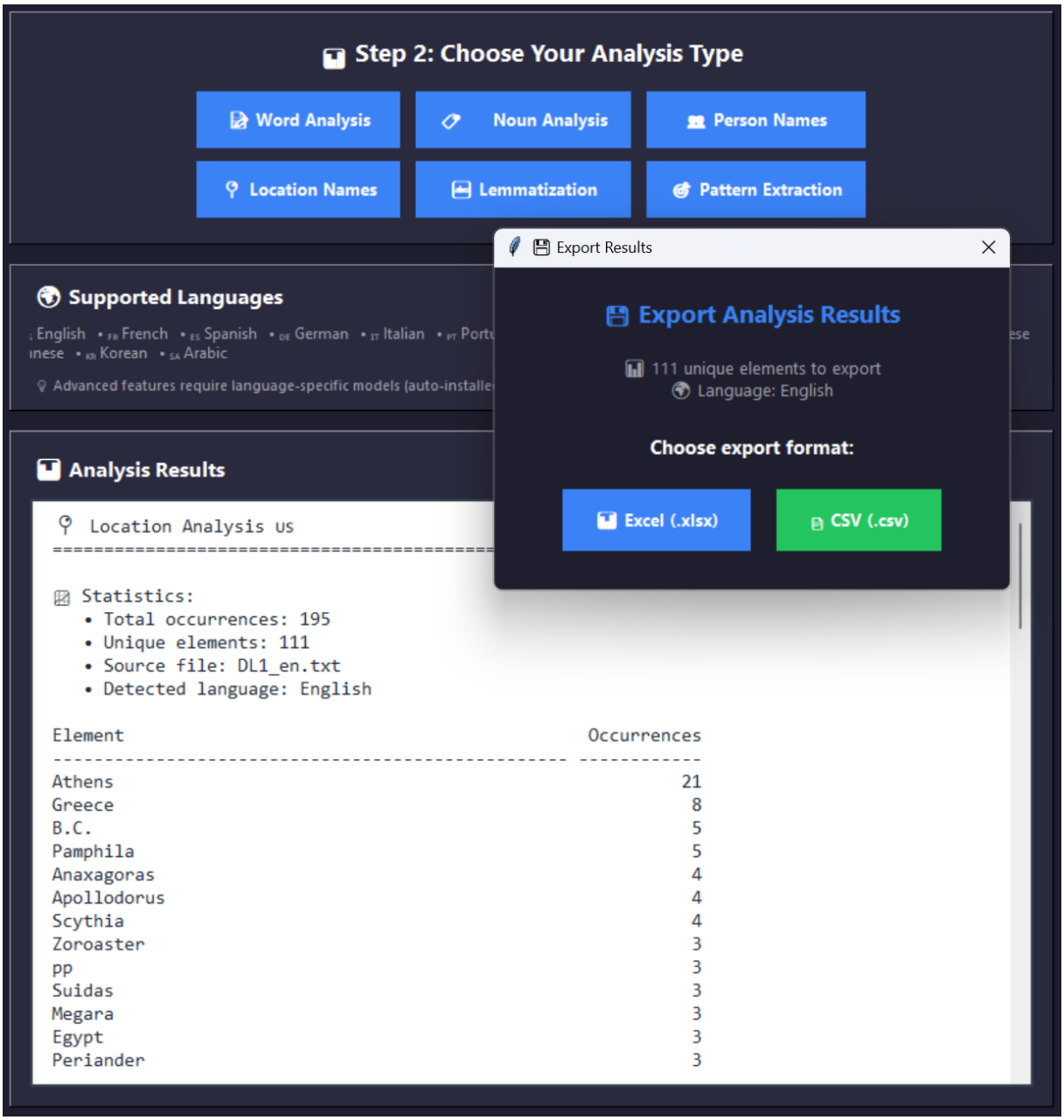
It leverages language-specific models to recognize personal names regardless of their grammatical form. This is useful for prosopographical studies or identifying main actors in a narrative. The results can be exported in Excel or CSV formats.



7.6 Location Names

The Location Names function detects and counts named entities of type “GPE” (geopolitical entities such as cities, countries).

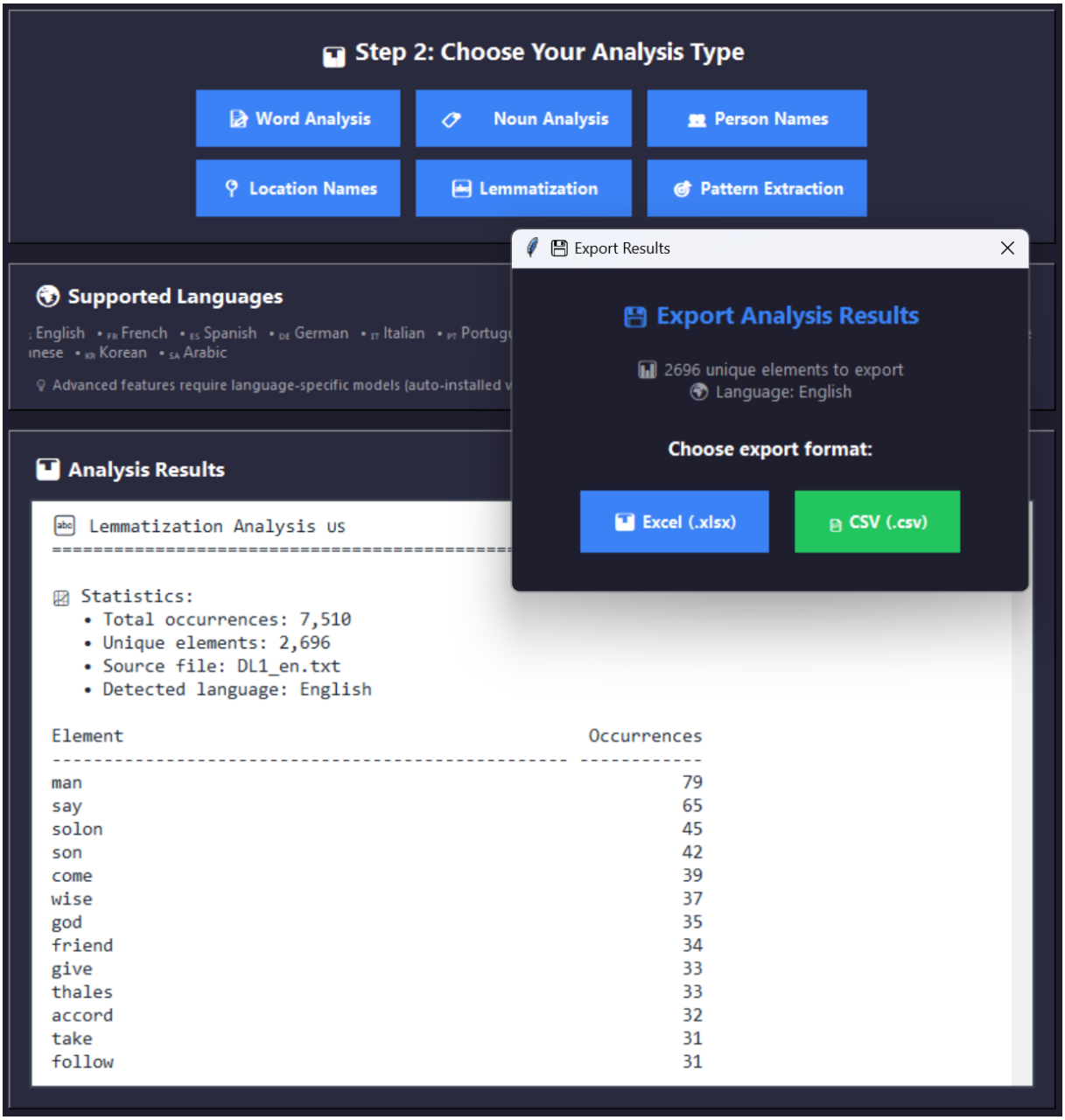
It distinguishes place names from other nouns and adapts to the detected language to improve accuracy. This is particularly relevant for geographical analysis or mapping historical references.



7.7 Lemmatization

Lemmatization reduces words to their base form (lemmas) and counts their frequency.

It removes grammatical inflections, allowing related forms of a word (e.g., “running”, “ran”) to be counted together. This is essential for linguistic analysis, topic modeling, or building lexicons.

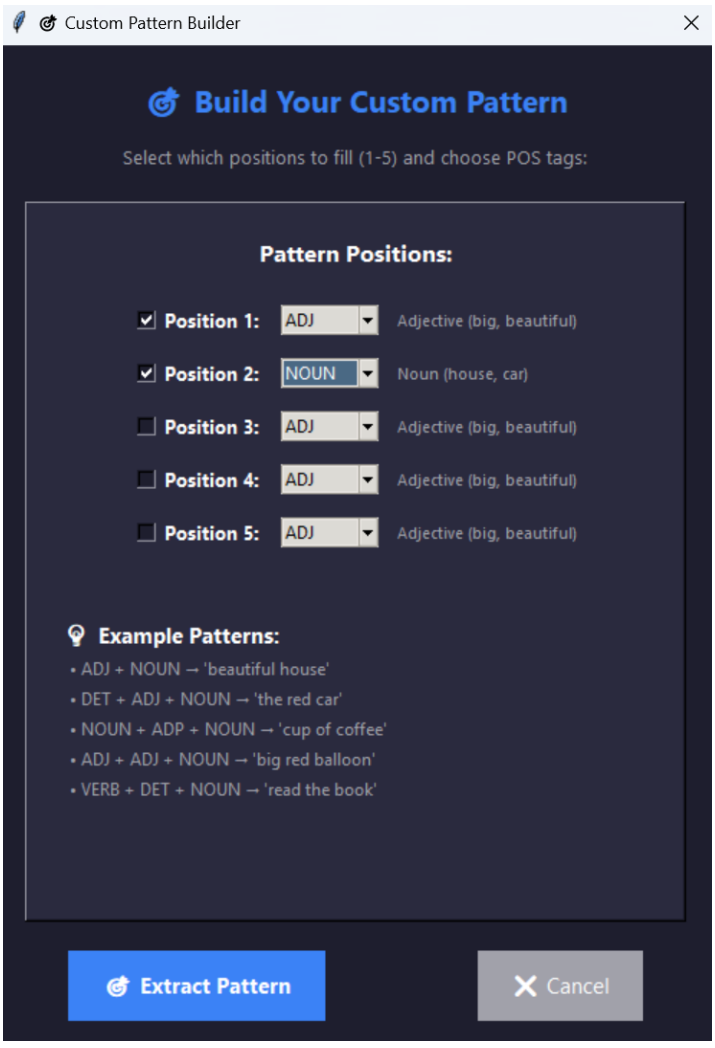


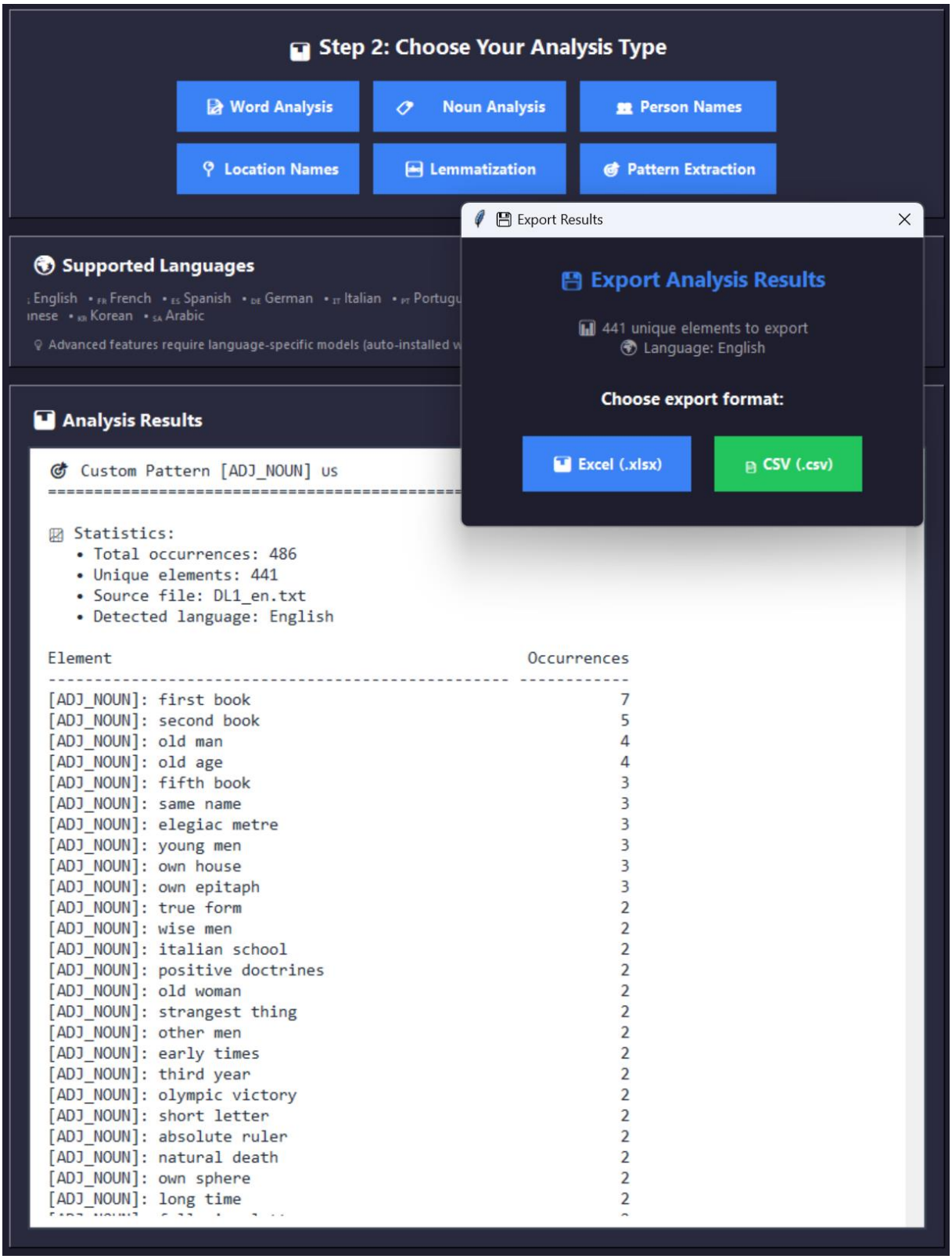
7.8 Pattern Extraction

Pattern Extraction finds repeated lexical-syntactic patterns (e.g., adjective + noun) or custom POS patterns chosen by the user.

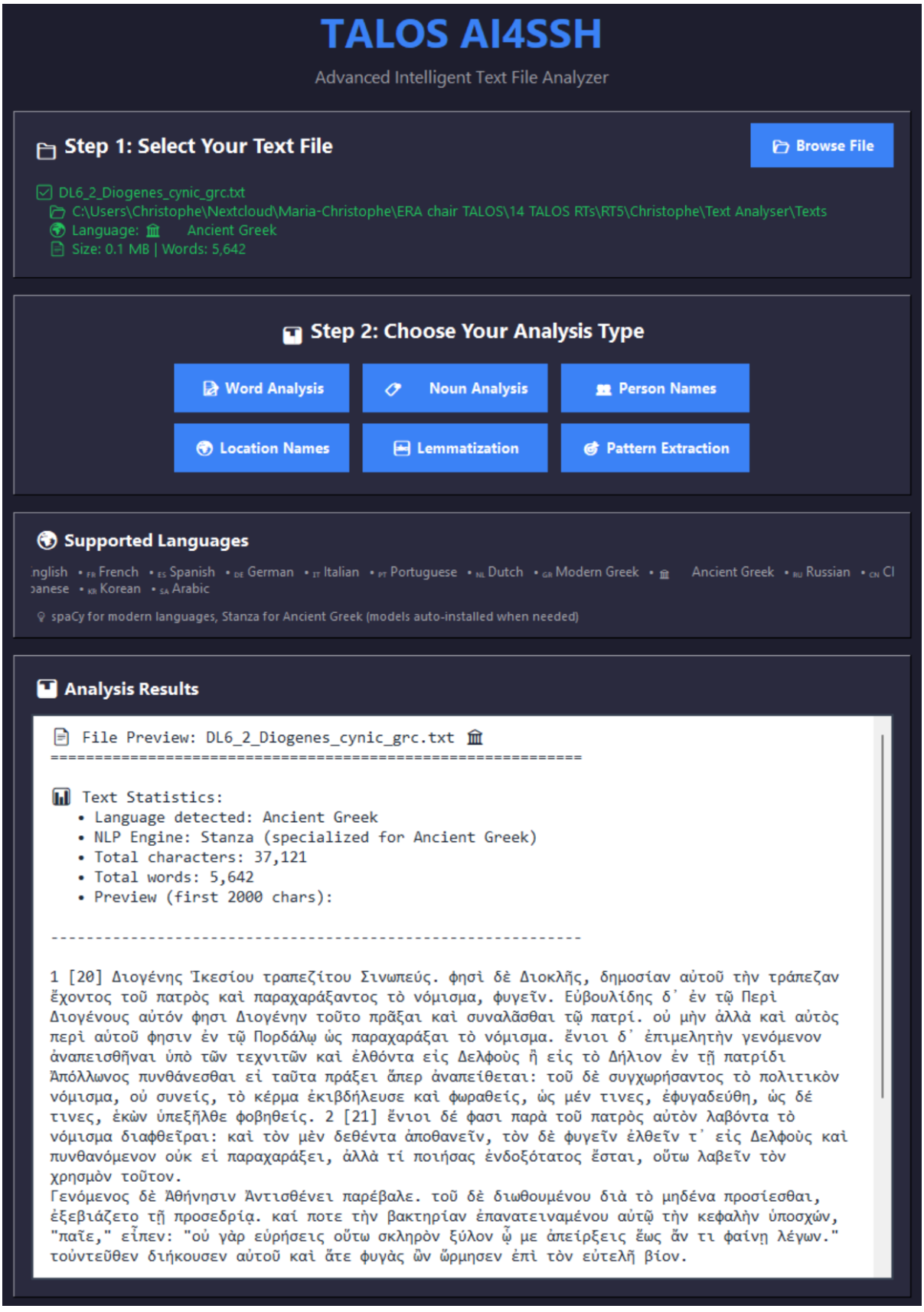
It scans the text for predefined structures but also lets the user define up to 5-word custom patterns with optional wildcards. This is valuable for studying collocations, stylistic tendencies, or formulaic expressions.

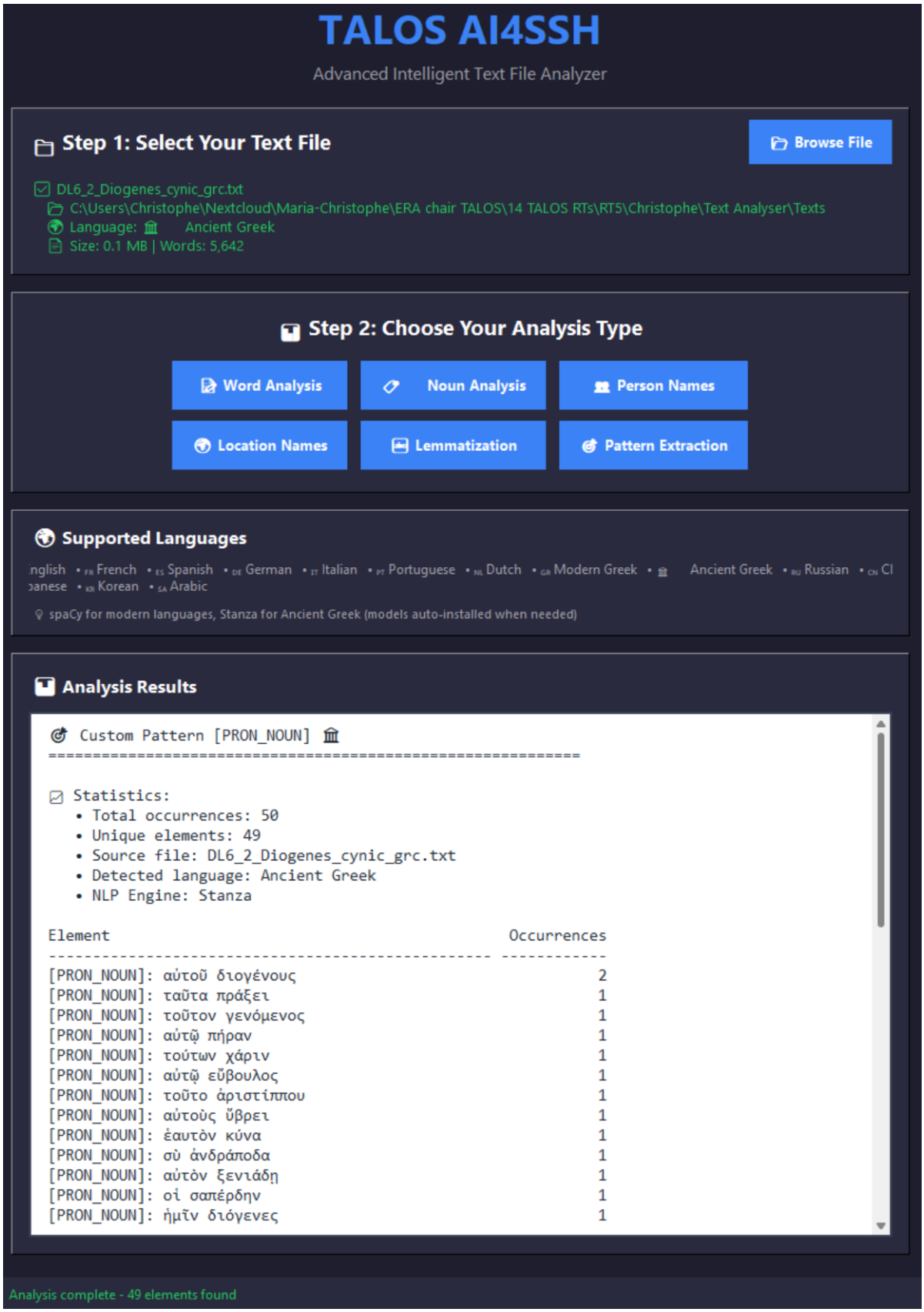
The custom pattern builder interface allows the user to define lexical-syntactic patterns by selecting up to five consecutive positions and assigning each one a part-of-speech tag (e.g., ADJ, NOUN, VERB) or a wildcard.




Step 2: Choose Your Analysis Type
Word Analysis
Noun Analysis
Person Names
Location Names
Lemmatization
Pattern Extraction
Supported Languages
English • FR French • ES Spanish • DE German • IT Italian • PT Portugu
inese • KR Korean • SA Arabic
Advanced features require language-specific models (auto-installed w
Export Results
Export Analysis Results
441 unique elements to export
Language: English
Choose export format:
Excel (.xlsx)
CSV (.csv)
Analysis Results
Custom Pattern [ADJ_NOUN] us
Statistics:
• Total occurrences: 486
• Unique elements: 441
• Source file: DL1_en.txt
• Detected language: English
Element Occurrences
[ADJ_NOUN]: first book 7
[ADJ_NOUN]: second book 5
[ADJ_NOUN]: old man 4
[ADJ_NOUN]: old age 4
[ADJ_NOUN]: fifth book 3
[ADJ_NOUN]: same name 3
[ADJ_NOUN]: elegiac metre 3
[ADJ_NOUN]: young men 3
[ADJ_NOUN]: own house 3
[ADJ_NOUN]: own epitaph 3
[ADJ_NOUN]: true form 2
[ADJ_NOUN]: wise men 2
[ADJ_NOUN]: italian school 2
[ADJ_NOUN]: positive doctrines 2
[ADJ_NOUN]: old woman 2
[ADJ_NOUN]: strangest thing 2
[ADJ_NOUN]: other men 2
[ADJ_NOUN]: early times 2
[ADJ_NOUN]: third year 2
[ADJ_NOUN]: olympic victory 2
[ADJ_NOUN]: short letter 2
[ADJ_NOUN]: absolute ruler 2
[ADJ_NOUN]: natural death 2
[ADJ_NOUN]: own sphere 2
[ADJ_NOUN]: long time 2

7.9 Example with an Ancient Greek Text




TALOS AI4SSH
Advanced Intelligent Text File Analyzer
Step 1: Select Your Text File
Browse File
DL6_2_Diogenes_cynic_grc.txt
C:\Users\Christophe\Nextcloud\Maria-Christophe\ERA chair TALOS\14 TALOS RTs\RT5\Christophe\Text Analyser\Texts
Language: Ancient Greek
Size: 0.1 MB | Words: 5,642
Step 2: Choose Your Analysis Type
Word Analysis
Noun Analysis
Person Names
Location Names
Lemmatization
Pattern Extraction
Supported Languages
nglish • FR French • ES Spanish • DE German • IT Italian • PT Portuguese • NL Dutch • GR Modern Greek • Ancient Greek • RU Russian • CN Cl
panese • KR Korean • SA Arabic
spaCy for modern languages, Stanza for Ancient Greek (models auto-installed when needed)
Analysis Results
Custom Pattern [PRON_NOUN]
Statistics:
• Total occurrences: 50
• Unique elements: 49
• Source file: DL6_2_Diogenes_cynic_grc.txt
• Detected language: Ancient Greek
• NLP Engine: Stanza
Element Occurrences
[PRON_NOUN]: αὐτοῦ διογένους 2
[PRON_NOUN]: ταῦτα πράξει 1
[PRON_NOUN]: τοῦτον γενόμενος 1
[PRON_NOUN]: αὐτῷ πήραν 1
[PRON_NOUN]: τούτων χάριν 1
[PRON_NOUN]: αὐτῷ εὔβουλος 1
[PRON_NOUN]: τοῦτο ἀριστίππου 1
[PRON_NOUN]: αὐτοὺς ὕβρει 1
[PRON_NOUN]: ἑαυτὸν κύνα 1
[PRON_NOUN]: σὺ ἀνδράποδα 1
[PRON_NOUN]: αὐτὸν ξενιάδῃ 1
[PRON_NOUN]: οἱ σαπέρδην 1
[PRON_NOUN]: ἡμῖν διόγενες 1
Analysis complete - 49 elements found

8. Developer

- Prof Christophe Roche – TALOS ERA Chair Holder – University of Crete
  https://talos-ai4ssh.uoc.gr/
- Contact: roche.university@gmail.com

9. License

- Creative Commons CC-BY-NC[1] 

10. Download links

- Web site: http://talos-ai4ssh.eu/Text_Analyser/
- Source code: https://talos-ai4ssh.uoc.eu/Text_Analyser/Talos_Text_Analyser.zip
- Documentation:
  https://talos-ai4ssh.uoc.gr/File_Analyser/Talos_Text_Analyser_Documentation.pdf

11. Support

- Prof Christophe Roche – TALOS ERA Chair Holder – University of Crete
  https://talos-ai4ssh.uoc.gr/
- Contact: roche.university@gmail.com

End of document

[1] This license enables reusers to distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only, and only so long as attribution is given to the creator.
https://creativecommons.org/share-your-work/cclicenses/