

translate.py - Text-Extraktion und -Übersetzung für DXF-Dateien

Übersicht

translate.py ist ein Python-Tool zur Extraktion, Filterung und Übersetzung von Texten aus DXF/DWG-Dateien. Es unterstützt mehrere Arbeitsweisen und kann Texte aus verschiedenen Quellen verarbeiten.

Hauptfunktionen

- **Text-Extraktion:** Extrahiert TEXT, MTEXT und Block-Attribute aus DXF-Dateien
- **Text-Filterung:** Filtert Texte basierend auf konfigurierbaren Mustern (Wildcards und Regex)
- **Automatische Übersetzung:** Übersetzt Texte mittels konfigurierbarer Wörterbücher (Multi-Phrasen, Trigramme, Bigramme, Einzelwörter)
- **DXF-zu-DXF Übersetzung:** Erstellt übersetzte DXF-Dateien mit mehrsprachigen Attributen
- **Export:** Exportiert Texte in verschiedene Formate (Excel, JSON, Text)

Konfigurationsdateien

translator.cfg

Hauptkonfigurationsdatei im **cfg/** Verzeichnis:

```
[symbol_attribute]
# Zuordnung von Sprachcodes zu Block-Attributnamen
DE = TEXT-D
EN = TEXT-E
ES = TEXT-ES
CS = TEXT-CS
FR = TEXT-F

[symbol_layer]
# Zuordnung von Sprachcodes zu Layer-Namen
DE = 0-0_TEXT-D
EN = 0-0_TEXT-E
ES = 0-0_TEXT-ES
CS = 0-0_TEXT-CS
FR = 0-0_TEXT-F

[ignore_pattern_wildcard]
# Wildcard-Muster für zu ignorierende Texte
layer = LAYER=*
coord_x = X=*
# ... weitere Muster

[ignore_pattern_regex]
# Regex-Muster für zu ignorierende Texte
```

```
decimal_number = ^-?\d+([\.,]\d+)?$  
# ... weitere Muster
```

Übersetzungs-Configs

Sprachspezifische Konfigurationsdateien im [translation/](#) Verzeichnis (z.B. [CS.cfg](#), [EN.cfg](#)):

```
[multi]  
# Phrasen mit 4+ Wörtern  
Main distribution panel = Hlavní rozváděč  
  
[trigramme]  
# 3-Wort-Phrasen  
Power supply unit = Napájecí zdroj  
  
[bigramme]  
# 2-Wort-Phrasen  
Emergency stop = Nouzové zastavení  
  
[single]  
# Einzelwörter  
Weight = Hmotnost  
Cable = Kabel
```

Verwendung

1. Text-Extraktion aus DXF

Extrahiert Texte aus einer DXF-Datei und speichert sie in verschiedenen Formaten.

```
# Extraktion in JSON (Standard)  
bin\translate.bat --filename file.dxf --extract --outname output.json  
  
# Extraktion in Excel  
bin\translate.bat --filename file.dxf --extract --export-type excel --outname  
output.xlsx  
  
# Extraktion in Textdatei  
bin\translate.bat --filename file.dxf --extract --export-type text --outname  
output.txt  
  
# Mehrere Formate gleichzeitig (komma-separiert)  
bin\translate.bat --filename file.dxf --extract --export-type excel,json,text --  
outname output  
# Erzeugt: output.xlsx, output.json, output.txt  
  
# Ohne Angabe von --outname: Automatische Namensvergabe  
bin\translate.bat --filename easy.dxf --extract --export-type excel,json  
# Erzeugt: easy_texts.xlsx, easy_texts.json
```

```
# Mit Quellsprache (Standard: DE -> TEXT-D)
bin\translate.bat --filename file.dxf --extract --lang_source EN
```

Extrahierte Entitäten:

- TEXT-Objekte im Modellbereich
- MTEXT-Objekte im Modellbereich
- Block-Attribute (nur Blöcke mit IO, ID, VERW, BEZEICHNUNG, KENNZEICHNUNG)

2. Text-Extraktion mit Übersetzung

Extrahiert und übersetzt Texte automatisch mittels Übersetzungs-Config.

```
# Extraktion mit automatischer Übersetzung nach CS (Tschechisch)
bin\translate.bat --filename file.dxf --extract --translate CS --outname
output.json

# Extraktion mit Übersetzung nach EN (Englisch)
bin\translate.bat --filename file.dxf --extract --translate EN --export-type excel
```

Output-Format (JSON):

```
{
  "translations": [
    {"source": "Weight", "target": "Hmotnost"}
  ],
  "untranslated": [
    {"source": "Custom text", "target": ""}
  ],
  "ignored": [
    "X=123.45",
    "LAYER=0"
  ]
}
```

3. DXF-zu-DXF Übersetzung

Erstellt eine neue DXF-Datei mit übersetzten Texten in mehrsprachigen Attributen.

```
# Übersetzung TEXT-D -> TEXT-CS (verwendet automatisch file_texts.json)
bin\translate.bat --filename file.dxf --todxf file_cs.dxf --translate CS

# Mit expliziter JSON-Datei
bin\translate.bat --filename file.dxf --todxf output.dxf --translate CS --
translation-json custom.json
```

```
# Mit alternativer Quellsprache
bin\translate.bat --filename file.dxf --todxf output.dxf --lang_source EN --
translate CS
```

Verhalten:

- **TEXT/MTEXT:** Direkt im Entity übersetzt
- **Block-Attribute:**
 - Liest aus Quell-Attribut (z.B. TEXT-D)
 - Schreibt in Ziel-Attribut (z.B. TEXT-CS)
 - Quell-Attribut bleibt unverändert
 - Erstellt fehlende ATTDEF im Block
 - Erstellt fehlende Layer (aus `[symbol_layer]`)

4. Text-Filterung aus TXT-Datei

Verarbeitet bereits extrahierte Texte aus einer Textdatei.

```
bin\translate.bat --fromtxt texts.txt --outname filtered.json
bin\translate.bat --fromtxt texts.txt --translate CS --outname translated.xlsx
```

5. JSON-Reprocessing

Übersetzt den `untranslated` Block einer JSON-Datei erneut.

```
bin\translate.bat --fromjson output.json --translate CS
```

Parameter-Referenz

Eingabe-Modi (mutual exclusive)

- `--extract` / `-e`: Extrahiere Texte aus DXF/DWG (benötigt `--filename`)
- `--fromtxt FILE.txt`: Filtere Texte aus Textdatei
- `--fromjson FILE.json`: Re-prozessiere JSON (benötigt `--translate`)
- `--todxf OUTPUT.dxf`: DXF-zu-DXF Übersetzung (benötigt `--filename`)

Hauptparameter

- `--filename FILE` / `-f FILE`: Eingabe-DXF/DWG-Datei
- `--outname NAME` / `-o NAME`: Name der Ausgabedatei (ohne Endung bei mehreren Formaten)
- `--translate LANG` / `-t LANG`: Übersetzungssprache (CS, EN, FR, etc.)
- `--lang_source LANG`: Quellsprache für Block-Attribute (Standard: DE)
- `--export-type TYPE`: Exportformat(e): `json` (Standard), `excel`, `text` oder komma-separiert (z.B. `excel,json,text`)
- `--translation-json FILE`: JSON-Datei mit Übersetzungen für `--todxf`

- `--debug` / `-d`: Debug-Ausgabe für Pattern-Matching

Arbeitsablauf-Beispiele

Workflow 1: Extraktion und manuelle Übersetzung

```
# 1. Texte extrahieren
bin\translate.bat --filename project.dxf --extract --outname project_texts.json

# 2. JSON manuell bearbeiten (untranslated -> translations verschieben)

# 3. Übersetzte DXF erstellen
bin\translate.bat --filename project.dxf --todxf project_cs.dxf --translate CS
```

Workflow 2: Automatische Übersetzung

```
# 1. Texte extrahieren und automatisch übersetzen
bin\translate.bat --filename project.dxf --extract --translate CS --outname
project_texts.json

# 2. Nicht-übersetzte Texte in JSON prüfen und manuell ergänzen

# 3. Übersetzte DXF erstellen
bin\translate.bat --filename project.dxf --todxf project_cs.dxf --translate CS
```

Workflow 3: Mehrsprachige Dokumente

```
# Tschechische Version
bin\translate.bat --filename project.dxf --todxf project_cs.dxf --translate CS

# Englische Version
bin\translate.bat --filename project.dxf --todxf project_en.dxf --translate EN

# Französische Version
bin\translate.bat --filename project.dxf --todxf project_fr.dxf --translate FR
```

Block-Attribute und Layer

Automatische ATTDEF-Erstellung

Wenn ein Block noch kein ATTDEF für das Ziel-Attribut hat:

1. **ATTDEF wird erstellt** im Block-Definition
2. **Layer wird erstellt** (falls nicht vorhanden, aus `[symbol_layer]`)
3. **Eigenschaften werden übernommen** vom Quell-ATTDEF:
 - Position (`insert`)

- Höhe (**height**)
- Stil (**style**)
- Rotation (**rotation**)
- Layer (aus Config oder vom Quell-Attribut)

Erforderliche Block-Attribute

Blöcke müssen folgende Attribute haben, um verarbeitet zu werden:

- **IO**
- **ID**
- **VERW**
- **BEZEICHNUNG**
- **KENNZEICHNUNG**
- Quell-Textattribut (z.B. **TEXT-D**)

Umgebungsvariablen

Das Programm benötigt folgende Umgebungsvariablen (werden in **bin/setenv.bat** gesetzt):

- **PROJECT_WORK**: Arbeitsverzeichnis für Ein-/Ausgabe-Dateien
- **PROJECT_CFG**: Konfigurationsverzeichnis (**cfg/**)
- **PROJECT_TRANSLATION**: Übersetzungs-Konfigurationsverzeichnis (**translation/**)

Fehlerbehandlung

Das Programm gibt detaillierte Statusmeldungen aus:

```
Anzahl gefundener TEXT Objekte: 150
Anzahl gefundener MTEXT Objekte: 25
Anzahl gefundener TEXT-D in Blöcken: 340
Duplikate entfernt: 75
Eindeutige Texte extrahiert: 440

Filterung abgeschlossen:
  Akzeptierte Texte: 380
  Ignorierte Texte: 60

DXF-Übersetzung abgeschlossen:
  TEXT übersetzt: 140, übersprungen: 10, ignoriert: 0
  MTEXT übersetzt: 20, übersprungen: 5, ignoriert: 0
  TEXT-CS (Blöcke) erstellt/aktualisiert: 310, übersprungen: 30, ignoriert: 0
  Gesamt: 470 übersetzt, 45 übersprungen, 0 ignoriert
```

Hinweise

- Texte werden als Set verarbeitet → Duplikate werden automatisch entfernt
- Leere Texte werden ignoriert
- Übersetzungen erfolgen in der Reihenfolge: Multi → Trigramme → Bigramme → Einzelwörter

- Bei DXF-zu-DXF Übersetzung bleibt das Original-Attribut immer unverändert
- Layer und ATTDEF werden nur bei Bedarf erstellt (lazy creation)
- Alle Ausgabedateien landen im `PROJECT_WORK` Verzeichnis

Siehe auch

- [CLAUDE.md](#) - Allgemeine Projektdokumentation
- `cfg/translator.cfg` - Hauptkonfiguration
- `translation/*.cfg` - Sprachspezifische Übersetzungs-Configs