GESIS Leibniz-Institut für Sozialwissenschaften

Heinrich Heine Universität Düsseldorf

elfen: A Python Package for Efficient Linguistic Feature Extraction for Natural Language Datasets

elfen is a fast, easy to use, multilingual linguistic feature extraction tool

Multilingual

Depending on backend:

SpaCy: 24 languages

Stanza: 80 languages

Fast Extraction

Dataset	Size	Preprocessing	Extraction
MMLU-Pro	12,032	86.07s	812.86s
BigBench Hard	6,511	67.25s	453.96s

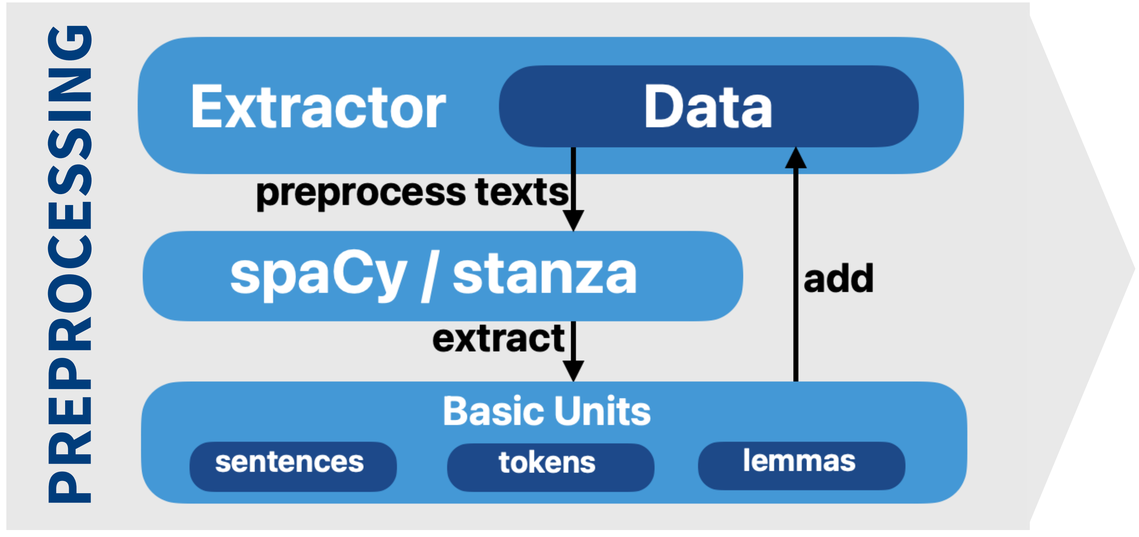
Extraction times for all features in elfen on an M4 Macbook Pro

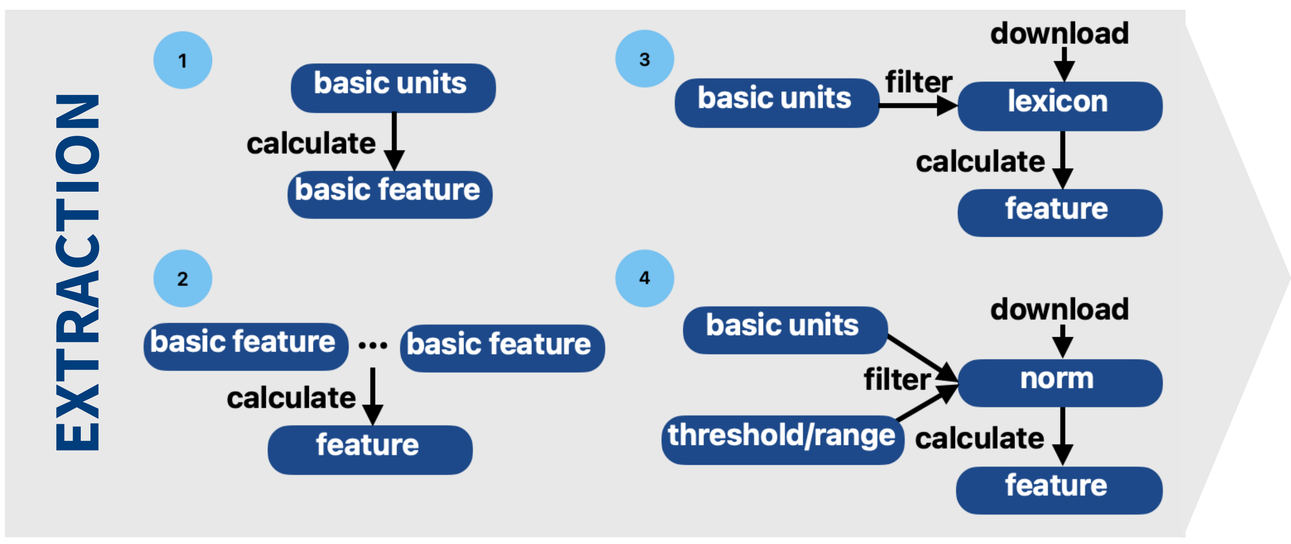
Straight-forward usage

```
# initializing extractor
extractor = elfen.Extractor(
                        data = df,
                        language = "en",
                        text_column = "text")
# extracting a single feature: ttr
extractor.extract("ttr")
# extracting a feature area/group: readability
extractor.extract_feature_group("readability")
# extracting all available features
extractor.extract_features()
```

>1000 Features in Eleven Feature Areas

Surface · Psycholinguistics · Named Entities · Semantics · Lexical Richness · Information · POS · Dependencies · Emotion · Morphology · Readability





elfen facilitates research on

LLM-generated and human-written text

Explainable Authorship Verification

Multi-Agent Dynamics

Annotation Patterns of Models and Humans

Gendered Perception of Language

For more information, check out

GitHub Repo

Documentation

Code Examples

GESIS MethodsHub

Maximilian Maurer
Contact: maximilian.maurer@gesis.org

Leibniz Gemeinschaft