# About

This preprocessing script was made for the

*Opinion Holder and Target Extraction for German*

Tool, created at Saarland University.

Script and necessary tools were tested under Ubuntu 16.04 LTS, the installation guidlines also assume that the operating system is Ubuntu 16.04 LTS.

Input for the script is either TigerXML or raw text. Raw text has to be formatted one sentence per line tokenized text.

Output of the script are four files in total, located in an output folder named "preprocessing_TIMESTAMP".

The output files are:

- raw_text.txt -> Depending on the input, either the input raw text file, or converted from TigerXML (one sentence per line)
- constituency_parse.xml -> Depending on the input, either the input XML file, or the constituency parse generated by Berkeley Parser (TigerXML)
- dependency_parse.txt -> The dependency parse generated by ParZu Parser (CoNLL2009 format)
- named_entity.txt -> The named entity tagged text generated by GermaNER (one sentence per line, annotated with NE-Tags)

# Installation of the necessary tools

The preprocessing tool uses the following NLP-Tools to create the necessary files:

- ParZu Parser (including CleverTagger) for dependency parsing
- Berkeley Parser for constituency parsing
- GermaNER for named entity recognition
- LinguaAlign for conversion between different parsing formats

If you have not already installed the tools, you can follow following instructions:

## ParZu Parser

**Dependencies:**

- swi-prolog > Version 5.6
- sfst

- git
- python > Version 2.6
- perl > Version 5.10

To install the necessary dependencies, you can use the Ubuntu package manager:

```
sudo apt-get install swi-prolog sfst git perl python2.7
```

### Download:

To download ParZu from github use following command:

```
git clone https://github.com/rsennrich/ParZu ParZu
```

### Installation:

Now open a terminal in the ParZu directory and run:

```
sh install.sh
```

This will install the necessary configuration for the preprocessing script.

## Berkeley Parser

### Dependencies:

- JAVA > Version 8

To install the necessary dependencies, you can use the Ubuntu package manager:

```
sudo apt-get install openjdk-8-jre-headless
```

### Download:

Download the BerkeleyParser-1.7.jar and ger_sm5.gr grammar from **https://github.com/slavpetrov/berkeleyparser**

### Installation:

Create a directory for the Berkeley Parser and move the downloaded files into it.

## GermaNER

We recommend using the 64bit Version, since using the version without freebase features will affect the performance.

### 64bit System

#### Dependencies:

- JAVA > Version 8

To install the necessary dependencies, you can use the Ubuntu package manager:

```
sudo apt-get install openjdk-8-jre-headless
```

#### Download:

Download the GermaNER-09-09-2015.jar from **https://github.com/tudarmstadt-lt/GermaNER**, the file is located in the README.md, under point 1 in "GermaNER in three lines".

#### Installation:

Create a directory for the GermaNER JAR file and move the downloaded file into it.

### 32bit System

#### Dependencies:

- JAVA > Version 8

To install Java you can use the Ubuntu package manager:

```
sudo apt-get install openjdk-8-jre-headless
```

#### Download

Download the GermaNER-nofb-09-09-2015.jar from **https://github.com/tudarmstadt-lt/GermaNER**, the file is located in the README.md, under point 1 in "GermaNER in three lines", it is the file WITHOUT freebase-features.

Download the crfsuite-0.12.tar.gz source package from **www.chokkan.org/software/crfsuite/**.

Download the liblbfgs-1.10.tar.gz source package from

**www.chokkan.org/software/liblbfgs/**.

### Installation

- Create a directory for the GermaNER JAR file and move the downloaded file into it.
- Extract the crfsuite-0.12.tar.gz package into the GermaNER directory.
- Extract the liblbfgs-1.10.tar.gz package into the GermaNER directory.
- Open a terminal in the liblbfgs-1.10 folder and enter following commands:

  ```
  ./configure --prefix=$PWD/../crf

  make

  make install
  ```
- Open a terminal in the crfsuite-0.12 folder and enter following commands:

  ```
  ./configure --prefix=$PWD/../crf --with-liblbfgs=$PWD/../crf

  make

  make install
  ```
- Add the absolute classpath of the /GermaNER/crf/bin folder to your path variable e.g. */home/user/Tools/GermaNER/crf/bin*

## LinguaAlign

### Dependencies:

- perl > Version 5.10

To install the necessary dependencies, you can use the Ubuntu package manager:

```
sudo apt-get install perl
```

### Download:

Download the LinguaAlign-package from **http://search.cpan.org/~tiedemann/Lingua-Align-0.04/bin/convert_treebank**, extract the folder.

Download Munkres-package from **http://search.cpan.org/~tpederse/Algorithm-Munkres-0.08/lib/Algorithm/Munkres.pm**, then extract the Algorithm folder from Munkres/lib into the LinguaAlign/lib folder.

### Installation:

Open an terminal in the LinguaAlign folder, and run the following commands:

```
perl Makefile.PL

make

sudo make install
```

# Running the script

Before you can run the script, the config file has to be edited to set the paths to the external tools.

The paths have to be set including the executable of each tool, for examples you can check the pre-entered paths in the config file.

The default script input is a TigerXML file

To run the script, use the following command:

```
sh preprocess.sh <input file>
```

Furthermore there are following options:

- -k -> keep the temporary preprocessing files
- -r -> use a raw text file instead of TigerXML as input