The folder’s content and instructions

This folder contains the resources needed to automatically recreate the results (the automatic portion) and the file with the manual annotations.

**STEP 1.** Make sure you have in the same folder with the scripts:

1. a) To reproduce same reproduce same results:  
    Dataset with the translated sentences “data\_master\_text\_clean\_2.csv”

b) For your own dataset upload your dataset and in the script “Create resouces, automatic annotation for domain,rare and comparison.ipynb” change the name of the csv file above to your own file name and change the name of the columns in the script based on your file]

1. Download the file from “<https://www.kaggle.com/datasets/rtatman/english-word-frequency>" in the same folder as the script
2. The list of domain-specific words present here “List of domain specific (reference set and the extended one).xlsx”. For another domain other than political conflict, use your own reference list.

**STEP 2.** Run the script “Create list of the most common lemmas of English (non-mod,stanza).ipynb” [This script will lemmatize the English Word Frequency list, will sum the frequency for the same lemmas in one, eliminate the duplicates, will calculate the times per million according to OED frequency bands and will keep only the most common lemmas and their times per million. The list will be used by the other script “Create resouces, automatic annotation for domain,rare and comparison.ipynb” to annotate the rarity feature of the lemmas]. It will create the file “Most common English lemmas.xlsx”.

**STEP 3.** Run the script “Create resouces, automatic annotation for domain,rare and comparison.ipynb”

[The script opens **data\_master\_text\_clean\_2.csv** and selects the columns en, es\_en\_DEEP, es\_en\_DEEPL, es\_en\_GOOGLE, es\_en\_TRANSFOMERS, ar\_en\_DEEP, ar\_en\_DEEPL, ar\_en\_GOOGLE, and ar\_en\_TRANSFORMERS and:

* The script runs the Stanza tokenizer and lemmatizer on each cell. It lists the unique words for every column together with their frequencies. It also lists the unique lemmas for every column and their frequencies. It then creates the column **lost lemmas\_en**, which holds the words that appear only in the en column and their frequencies.
* Next: annotation of “rare” feature. The script opens **Most common English lemmas.xlsx** and checks every lemma list. It writes “yes” for lemmas that appear in the common list. It writes “no,” followed by the “times per million” value, for lemmas that do not appear.
* Next: annotation of “domain-specific” feature. The script then opens **List of domain specific (reference set and the extended one).xlsx** and compares every lemma list with the domain list. It adds a domain-specific column for every original column and writes “yes” or “no” for each lemma. Finally, it saves the full table as **Automatic annotations\_rare\_domain-specific\_stanza.csv and Automatic annotations\_rare\_domain-specific\_spacy.csv** (in two versions, one uses stanza tokenizer/lemmatizer and one uses SpaCy).
* It will also create two table with statistics and graphs for summary\_statistics\_stanza.csv and summary\_statistics\_spacy.csv
* It creates statistics at the end for nr of unique words/lemmas/lost lemmas/domain-specific/rare for each language pair & machine translation tool
* And separately statistics for the lost lemmas only to see their compoundness related to rare and domain-specific
* Compares the results between two different pipelines
* Compares the results between the automatic annotation and the manual annotation (table\_1\_manual\_annotations.csv)

Other files:

* Process description on investigating rarity, domain specificity, relationship mt-rt.docx
* Data\_master\_text\_clean\_2.csv (Dataset of translated sentences)
* Manual\_annotations\_lost\_lemmas\_rarity\_domain-specificity\_relationships.xlsx (List of
* Create list of the most common lemmas of English.ipynb
* Most common English lemmas.xlsx (list created with the script “Create list of the most common lemmas of English.ipynb”
* Manual annotations of relationships lost rare words

Manual of relationships (rare lemmas) (here we should leave only the four columns [lemma/relationship/rt translation unit/rare (maybe this one is implied)]

* es\_en\_DEEP
* es\_en\_DEEPL
* es\_en\_GOOGLE
* es\_en\_TRANSFORMERS
* Domain-specific lemmas (manual reference set and Conflibert)