In [ ]:

```python
import sys
sys.path.append('..')

import time

import numpy as np
import pandas as pd

from src.data_loader import DataLoader
from src.utils import BytePairEncoder
import matplotlib.pyplot as plt
```

Byte Pair Encoding Analysis

In [ ]:

```python
en_data_loader = DataLoader('../data/data_v2/multi30k.en.gz')
de_data_loader = DataLoader('../data/data_v2/multi30k.de.gz')
```

In [ ]:

```python
with open("../results/BPE_results.txt", "w") as file:
    file.write("Start BPE testing...\n\n")
num_operations = [1000,5000,15000]
```

Learning BPE on German and English Data Individually

English Data Set

In [ ]:

```python
vocab_sizes_en = []
fit_time_en = []
for operations in num_operations:
    encoder = BytePairEncoder()
    start_time = time.time()
    encoder.fit(en_data_loader.load_data(),operations=operations)
    end_time = time.time()
    elapsed_time = (end_time - start_time)/60
    # Document the encoded text
    with open("../results/BPE_results.txt", "a") as file:
        file.write(f"Sample of English data set encoded with {operations}-operations BPE \n")
        file.write(' '.join(encoder.encode(en_data_loader.load_data()[:500])))
        file.write("\n\n\n")
    vocab_sizes_en += [len(encoder.get_vocab())]
    fit_time_en += [elapsed_time]
    encoder.save_model(f'../logs/BPE_EN_{operations}')
```

```
Processing Text: 100%|██████████| 1000/1000 [00:08<00:00, 116.68it/s, New Token=lays, New Rule=lay s -> lays]
Processing Text: 100%|██████████| 5000/5000 [00:37<00:00, 132.16it/s, New Token=advice, New Rule=ad vice -> advice]
Processing Text:  39%|████      | 5838/15000 [00:32<00:51, 179.62it/s, New Token=wooden-floored, New Rule=wooden- floored -> wooden-floore
d]
```

German Data Set

In [ ]:

```python
vocab_sizes_de = []
fit_time_de = []
for operations in num_operations:
    encoder = BytePairEncoder()
    start_time = time.time()
    encoder.fit(de_data_loader.load_data(),operations=operations)
    end_time = time.time()
    elapsed_time = (end_time - start_time)/60
    # Document the encoded text
    with open("../results/BPE_results.txt", "a") as file:
        file.write(f"Sample of German data set encoded with {operations}-operations BPE \n")
        file.write(' '.join(encoder.encode(de_data_loader.load_data()[:500])))
        file.write("\n\n\n")
    vocab_sizes_de += [len(encoder.get_vocab())]
    fit_time_de += [elapsed_time]
    encoder.save_model(f'../logs/BPE_DE_{operations}')
```

```
Processing Text: 100%|██████████| 1000/1000 [00:20<00:00, 49.34it/s, New Token=repariert, New Rule=repar iert -> repariert]
Processing Text: 100%|██████████| 5000/5000 [01:08<00:00, 72.80it/s, New Token=wickeln, New Rule=wickel n -> wickeln]
Processing Text:  79%|████████  | 11851/15000 [02:11<00:34, 90.45it/s, New Token=lotsenboots, New Rule=lotsen boots -> lotsenboots]
```

Learning BPE on German and English Data Combined

In [ ]:

```python
vocab_sizes_merged = []
fit_time_merged = []
merged_text = en_data_loader.load_data() + '\n' + de_data_loader.load_data()
for operations in num_operations:
    encoder = BytePairEncoder()
    start_time = time.time()
    encoder.fit(merged_text,operations=operations)
    end_time = time.time()
    elapsed_time = (end_time - start_time)/60
    # Document the encoded text
    with open("../results/BPE_results.txt", "a") as file:
        file.write(f"Sample of German data set encoded with {operations}-operations BPE (merged) \n")
        file.write(' '.join(encoder.encode(en_data_loader.load_data()[:500])))
        file.write('\n\n')
        file.write(f"Sample of English data set encoded with {operations}-operations BPE (merged) \n")
        file.write(' '.join(encoder.encode(de_data_loader.load_data()[:500])))
        file.write("\n\n\n")
    vocab_sizes_merged += [len(encoder.get_vocab())]
    fit_time_merged += [elapsed_time]
    encoder.save_model(f'../logs/BPE_EN-DE_{operations}')
```

```
Processing Text: 100%|██████████| 1000/1000 [00:30<00:00, 32.73it/s, New Token=erwachsene, New Rule=erwach sene -> erwachsene]
Processing Text: 100%|██████████| 5000/5000 [01:41<00:00, 49.34it/s, New Token=bögen, New Rule=bö gen -> bögen]
Processing Text:  91%|█████████ | 13716/15000 [03:30<00:19, 65.17it/s, New Token=lotsenboots, New Rule=lotsen boots -> lotsenboots]
```

Quantiative Analysis

In [ ]:

```python
data = {'German Vocab Size' : vocab_sizes_de, 'English Vocab Size' : vocab_sizes_en, 'Merged Vocab Size' : vocab_sizes_merged, 'Operation
df = pd.DataFrame(data)
df = df.set_index('Operations')
df = df.style.highlight_max(axis = 1, color='yellow')
print("Comparison of total number of operations to vocabulary size:")
df
```

```
Comparison of total number of operations to vocabulary size:
```

Out[ ]:

	German Vocab Size	English Vocab Size	Merged Vocab Size
Operations			
1000	1071	1054	1074
5000	5071	5054	5074
15000	11922	5892	13790

In [ ]:

```python
data = {'Time on German (min)' : fit_time_de, 'Time on English (min)' : fit_time_en, 'Time on Merged (min)' : fit_time_merged, 'Operation
df = pd.DataFrame(data)
df = df.set_index('Operations')
df = df.style.highlight_max(axis = 1, color='yellow')
print("Comparison of total number of operations to learning time:")
df
```

```
Comparison of total number of operations to learning time:
```

Out[ ]:

	Time on German (min)	Time on English (min)	Time on Merged (min)
Operations			
1000	0.339906	0.144886	0.512873
5000	1.146667	0.632568	1.692632
15000	2.185942	0.543540	3.511735

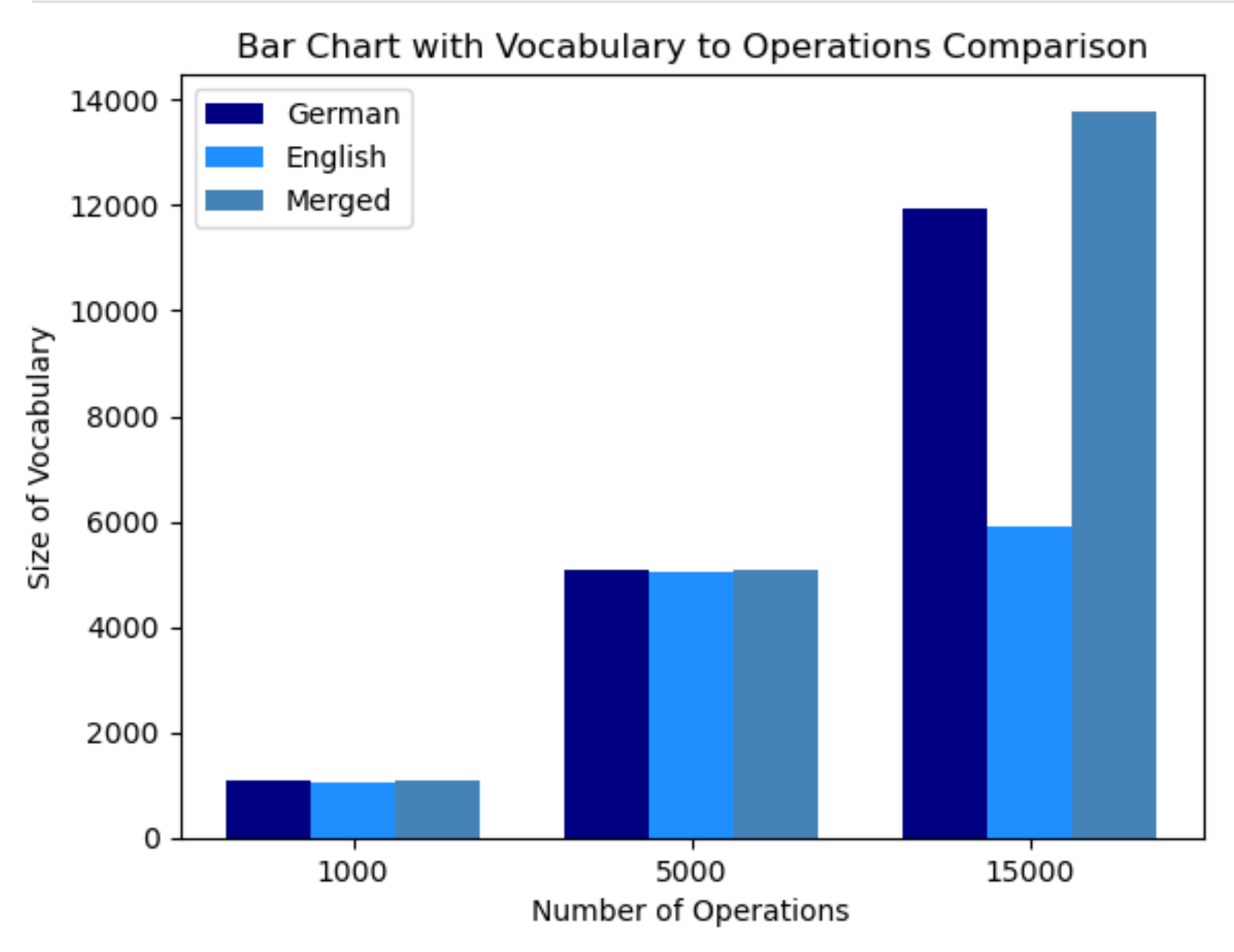
Visualization

In [ ]:

```python
# plot the data
x = np.array([0, 1, 2])
bar_width = 0.25
fig, ax = plt.subplots()
ax.bar(x - bar_width, vocab_sizes_de, width=bar_width, label='German', color=(0.0, 0.0, 0.5))
ax.bar(x, vocab_sizes_en, width=bar_width, label='English', color=(0.12, 0.56, 1.0))
ax.bar(x + bar_width, vocab_sizes_merged, width=bar_width, label='Merged', color=(0.27, 0.51, 0.71))

# set labels and title
ax.set_ylabel('Size of Vocabulary')
ax.set_xlabel('Number of Operations')
ax.set_title('Bar Chart with Vocabulary to Operations Comparison')
ax.set_xticks(x)
ax.set_xticklabels(num_operations)
ax.legend()

plt.show()
```



In [ ]:

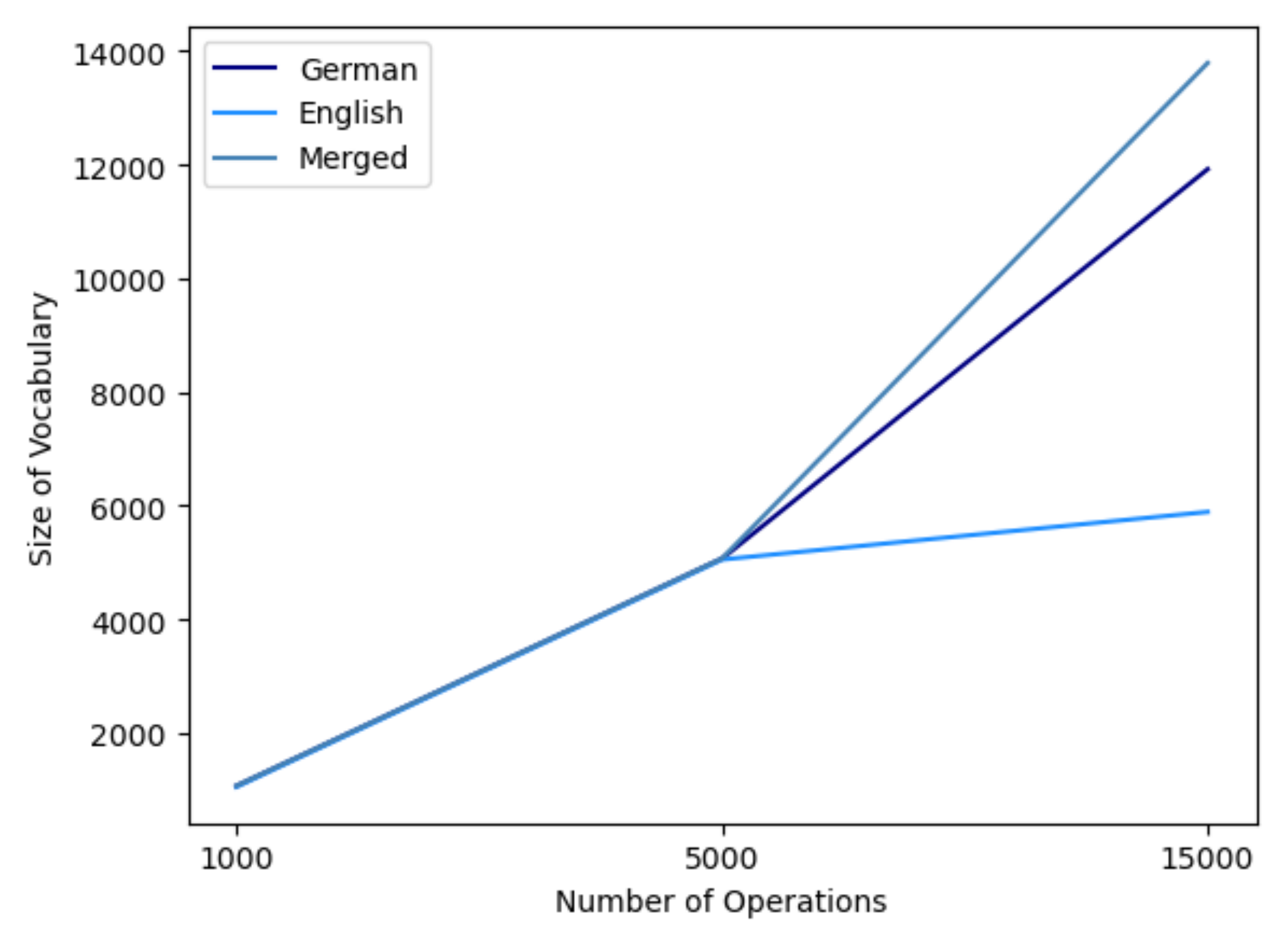
```python
# Create a figure and axis object
fig, ax = plt.subplots()

x = np.array([1, 2, 3])
# Plot the curves
ax.plot(x, vocab_sizes_de, label='German',color=(0.0, 0.0, 0.5))
ax.plot(x, vocab_sizes_en, label='English',color=(0.12, 0.56, 1.0))
ax.plot(x, vocab_sizes_merged, label='Merged',color=(0.27, 0.51, 0.71))
# Add a legend
ax.legend()

ax.set_ylabel('Size of Vocabulary')
ax.set_xlabel('Number of Operations')

ax.set_xticks(x)
ax.set_xticklabels(num_operations)

# Show the plot
plt.show()
```



Qualitative Analysis

Whilst the BPE algorithm stopped to learn more words, the vocabulary learned on the German dataset grew even more. One possible reason could be the amount of long and combined words, which can be build in the German language. One example would be "Antriebsradsystem".

There are also some differences between the vocabulary learned on German data compared to Vocabulary learned on the merged data. For example: "antr@@ ie@@ b@@ s@@ rads@@ y@@ ste@@ m" and "an@@ trie@@ b@@ s@@ rads@@ yste@@ m "

In [ ]:

```python
with open("../results/BPE_results.txt", "r") as file:
    print(file.read())
```

```
Start BPE testing...

Sample of English data set encoded with 1000-operations BPE
<s> two young , white males are outside near many bushes . </s> <s> several men in hard hats are operating a giant pul@@ ley sy@@ ste@@ m
. </s> <s> a little girl climbing into a wooden play@@ house . </s> <s> a man in a blue shirt is standing on a ladder cleaning a window .
</s> <s> two men are at the sto@@ ve preparing food . </s> <s> a man in green holds a guitar while the other man obser@@ ves his shirt . <
/s> <s> a man is smiling at a stuffed lion </s> <s> a t@@ rendy girl talking on her cellphone while g@@ liding s@@ low@@ ly down the stree
t . </s> <s> a woman with a large purse is walking by </s>


Sample of English data set encoded with 5000-operations BPE
<s> two young , white males are outside near many bushes . </s> <s> several men in hard hats are operating a giant pulley system . </s> <s
> a little girl climbing into a wooden playhouse . </s> <s> a man in a blue shirt is standing on a ladder cleaning a window . </s> <s> two
men are at the stove preparing food . </s> <s> a man in green holds a guitar while the other man observes his shirt . </s> <s> a man is sm
iling at a stuffed lion </s> <s> a trendy girl talking on her cellphone while gliding slowly down the street . </s> <s> a woman with a lar
ge purse is walking by </s>


Sample of English data set encoded with 15000-operations BPE
<s> two young , white males are outside near many bushes . </s> <s> several men in hard hats are operating a giant pulley system . </s> <s
> a little girl climbing into a wooden playhouse . </s> <s> a man in a blue shirt is standing on a ladder cleaning a window . </s> <s> two
men are at the stove preparing food . </s> <s> a man in green holds a guitar while the other man observes his shirt . </s> <s> a man is sm
iling at a stuffed lion </s> <s> a trendy girl talking on her cellphone while gliding slowly down the street . </s> <s> a woman with a lar
ge purse is walking by </s>


Sample of German data set encoded with 1000-operations BPE
<s> zwei junge weiße männer sind im freien in der nähe vieler bü@@ sche . </s> <s> mehrere männer mit schutzhelmen be@@ dienen ein antr@@
ie@@ b@@ s@@ rads@@ y@@ ste@@ m . </s> <s> ein kleines mädchen klettert in ein spiel@@ haus aus holz . </s> <s> ein mann in einem blauen h
emd steht auf einer leiter und putzt ein fenster . </s> <s> zwei männer stehen am her@@ d und bereiten essen zu . </s> <s> ein mann in grü
n hält eine gitarre , während der andere mann sein hemd ansieht . </s> <s> ein mann lächelt einen aus@@ gestopften l@@ ö@@ wen an . </s> <
s> ein sch@@ ickes mädchen spricht mit dem handy während sie </s>


Sample of German data set encoded with 5000-operations BPE
<s> zwei junge weiße männer sind im freien in der nähe vieler büsche . </s> <s> mehrere männer mit schutzhelmen bedienen ein antrie@@ b@@
s@@ radsystem . </s> <s> ein kleines mädchen klettert in ein spielhaus aus holz . </s> <s> ein mann in einem blauen hemd steht auf einer l
eiter und putzt ein fenster . </s> <s> zwei männer stehen am herd und bereiten essen zu . </s> <s> ein mann in grün hält eine gitarre , wä
hrend der andere mann sein hemd ansieht . </s> <s> ein mann lächelt einen ausgestopften löwen an . </s> <s> ein schickes mädchen spricht m
it dem handy während sie </s>


Sample of German data set encoded with 15000-operations BPE
<s> zwei junge weiße männer sind im freien in der nähe vieler büsche . </s> <s> mehrere männer mit schutzhelmen bedienen ein antriebsradsy
stem . </s> <s> ein kleines mädchen klettert in ein spielhaus aus holz . </s> <s> ein mann in einem blauen hemd steht auf einer leiter und
putzt ein fenster . </s> <s> zwei männer stehen am herd und bereiten essen zu . </s> <s> ein mann in grün hält eine gitarre , während der
andere mann sein hemd ansieht . </s> <s> ein mann lächelt einen ausgestopften löwen an . </s> <s> ein schickes mädchen spricht mit dem han
dy während sie </s>


Sample of German data set encoded with 1000-operations BPE (merged)
<s> two young , white males are outside near many bushes . </s> <s> several men in hard hats are operating a gi@@ an@@ t pul@@ ley s@@ yst
e@@ m . </s> <s> a little girl climbing into a wooden play@@ house . </s> <s> a man in a blue shirt is standing on a lad@@ der cleaning a
window . </s> <s> two men are at the sto@@ ve preparing food . </s> <s> a man in green holds a guitar while the other man ob@@ ser@@ ves h
is shirt . </s> <s> a man is smiling at a stu@@ ffed lion </s> <s> a tren@@ d@@ y girl talking on her cellphone while g@@ liding s@@ low@@
ly down the street . </s> <s> a woman with a large pur@@ se is walking by </s>

Sample of English data set encoded with 1000-operations BPE (merged)
<s> zwei junge weiße männer sind im freien in der nähe vieler bü@@ sche . </s> <s> mehrere männer mit schutz@@ helmen bedienen ein an@@ tr
ie@@ b@@ s@@ rads@@ yste@@ m . </s> <s> ein kleines mädchen klettert in ein spiel@@ haus aus holz . </s> <s> ein mann in einem blauen hemd
steht auf einer le@@ iter und put@@ zt ein fenster . </s> <s> zwei männer stehen am her@@ d und bereiten essen zu . </s> <s> ein mann in g
rün hält eine gitarre , während der andere mann sein hemd ansieht . </s> <s> ein mann lächelt einen aus@@ gestop@@ ften l@@ ö@@ wen an . <
/s> <s> ein sch@@ ickes mädchen spricht mit dem handy während sie </s>


Sample of German data set encoded with 5000-operations BPE (merged)
<s> two young , white males are outside near many bushes . </s> <s> several men in hard hats are operating a giant pulley system . </s> <s
> a little girl climbing into a wooden play@@ house . </s> <s> a man in a blue shirt is standing on a ladder cleaning a window . </s> <s>
two men are at the stove preparing food . </s> <s> a man in green holds a guitar while the other man observes his shirt . </s> <s> a man i
s smiling at a stuffed lion </s> <s> a trendy girl talking on her cellphone while gliding slowly down the street . </s> <s> a woman with a
large purse is walking by </s>

Sample of English data set encoded with 5000-operations BPE (merged)
<s> zwei junge weiße männer sind im freien in der nähe vieler büsche . </s> <s> mehrere männer mit schutzhelmen bedienen ein antrie@@ bs@@
radsystem . </s> <s> ein kleines mädchen klettert in ein spiel@@ haus aus holz . </s> <s> ein mann in einem blauen hemd steht auf einer le
iter und putzt ein fenster . </s> <s> zwei männer stehen am herd und bereiten essen zu . </s> <s> ein mann in grün hält eine gitarre , wäh
rend der andere mann sein hemd ansieht . </s> <s> ein mann lächelt einen ausgestopften löwen an . </s> <s> ein schickes mädchen spricht mi
t dem handy während sie </s>


Sample of German data set encoded with 15000-operations BPE (merged)
<s> two young , white males are outside near many bushes . </s> <s> several men in hard hats are operating a giant pulley system . </s> <s
> a little girl climbing into a wooden playhouse . </s> <s> a man in a blue shirt is standing on a ladder cleaning a window . </s> <s> two
men are at the stove preparing food . </s> <s> a man in green holds a guitar while the other man observes his shirt . </s> <s> a man is sm
iling at a stuffed lion </s> <s> a trendy girl talking on her cellphone while gliding slowly down the street . </s> <s> a woman with a lar
ge purse is walking by </s>

Sample of English data set encoded with 15000-operations BPE (merged)
<s> zwei junge weiße männer sind im freien in der nähe vieler büsche . </s> <s> mehrere männer mit schutzhelmen bedienen ein antriebsradsy
stem . </s> <s> ein kleines mädchen klettert in ein spielhaus aus holz . </s> <s> ein mann in einem blauen hemd steht auf einer leiter und
putzt ein fenster . </s> <s> zwei männer stehen am herd und bereiten essen zu . </s> <s> ein mann in grün hält eine gitarre , während der
andere mann sein hemd ansieht . </s> <s> ein mann lächelt einen ausgestopften löwen an . </s> <s> ein schickes mädchen spricht mit dem han
dy während sie </s>
```