```
import os
import csv
import pandas as pd
import torch
import torch.nn as nn
import torch.optim as optim
from torch.utils.data import Dataset, DataLoader
from torch.nn.utils.rnn import pad_sequence
from torch.nn import Transformer
import numpy as np
import spacy

# English-Sinhala translation pairs
translation_pairs = [
    ("Hello", "හෙලෝ"),
    ("Good morning", "සුභ උදෑසනක්"),
    ("How are you?", "ඔබ කෙසේ වෙයි?"),
    ("I am fine", "මම හොඳයි"),
    ("Goodbye", "සමුගැනීමට"),
    ("Thank you", "ස්තූතියි"),
    ("Please", "කරුණාකර"),
    ("Yes", "ඔව්"),
    ("No", "නැත"),
    ("Sorry", "කාලෙකට"),
    ("Excuse me", "කලින් ඉවුරුදු හැටි"),
    ("Thank you", "ස්තූතියි"),
    ("Please", "කරුණාකර"),
    ("Yes", "ඔව්"),
    ("No", "නැත"),
    ("Sorry", "කාලෙකට"),
    ("Excuse me", "කලින් ඉවුරුදු හැටි"),
    ("How are you?", "ඔබ කොහෙද?"),
    ("I'm fine, thank you.", "ඔහු මාසයේදී සමුගැනීමට පිළිගැනීමක් වෙත හිමිවේ."),
   ( "What is your name?", "ඔබගේ නම කුමක්ද?"),
   ( "My name is...", "මගේ නම..."),
    ("How old are you?", "ඔබ කල් වියදාද?"),
    ("I am ___ years old.", "මම ___ වන්දී පියාදි වෙයි."),
    ("Where are you from?", "ඔබ කොහෙදදී අපිටද?"),
    ("I am from...", "මම ... පිළිබඳව පොදු වියදෙනවා."),
    ("What do you do for a living?", "ඔබ ජීවත්වූ සංවාදය කුමක්ද?"),
    ("I am a...", "මම ... බිහිවූ වියදී වෙයි."),
    ("How can I help you?", "මම ඔබට කෙළින්නේ කෙසේද?"),
    ("Can you help me, please?", "කරුණාකර මාට සහභාගී කරන්නෙ කෙසේද?"),
    ("I need your assistance.", "මම ඔබගේ උපකාරය අවශ්‍ය වේ."),
    ("What time is it?", "වේලාව කුමක්ද?"),
    ("It is ___ o'clock.", "එය ___ ක් ට වේ."),
    ("Where is the bathroom?", "ශ්‍රී පුල් කොළ කොහොමද?"),
    ("The bathroom is over there.", "ශ්‍රී පුල් කොළ එහි දිගටම ඇත."),
    ("I am hungry.", "මම අඳුරු වෙයි."),
    ("I am thirsty.", "මම සමාලෝචනය වෙයි."),
    ("I am tired.", "මම දුරස්සට වෙයි."),
    ("I am sleepy.", "මම පයියෝයි වෙයි."),
    ("I am cold.", "මම විශ්වාසයක් වෙයි."),
    ("I am hot.", "මම හොඳින් වෙයි."),
    ("I am sick.", "මම රෝගී වෙයි."),
    ("I am lost.", "මම මැරෙනවා."),
    ("I am in trouble.", "මම අගය දක්වා ඇත."),
    ("I love you.", "මම ඔබ කැමති නොහැකිය."),
    ("I miss you.", "මම ඔබ අලුත්ම කියනවා."),
    ("I want to see you.", "මම ඔබ දැන් දැනුවත් බලා ගන්නවා."),
    ("I want to be with you.", "මම ඔබ සමග විස්තර ප්‍රකාශකරනවා."),
    ("Can I have...?", "මම ... අතුලු හැමෝම් ලබා දීමට කාටහන් කරයිද?"),
    ("I would like...", "මම ... වෙත ප්‍රවාහනය කරනවා."),
    ("Where is the nearest...?", "ආසන්නයේදී ... කොටුව කොහෙද?"),
    ("Could you show me the way to...?", "ඔබ මට ... වෙත මා පවතින මාර්ගය පෙන්වනවාද?"),
    ("How much does this cost?", "මෙමට අඩු කිරීම කොහොමද?"),
    ("I would like to buy..." ,"මම ... මිලදී ගන්න බැහැරවිය හැකිය."),
    ("Do you accept credit cards?", "ඔබට ක්‍රෙඩිට් කාඩ්පත් පිළිබඳව හැකියාවක් තිබේද?"),
    ("Can I try it on?", "මම එය උසස් කරනවාද?"),
    ("Where can I find...?", "මට ... ගැන ඉල්ලීමට කොහෙද?"),
    ("I am looking for...", "මම ... ගැන සොයා ගැනීම කරයි."),
    ("Can I speak to...?", "මට ... හවසට ඇතිවිට කියයිද?"),
    ("May I come in?", "මට විනාශ විය හැක්කේද?"),
    ("Wait a moment.", "කාලය වත්කර බලා ගන්න."),
    ("Let me think.", "මට කරුණාකර යෝජනා කරන්න."),
    ("I understand.", "මට තිබෙනවා."),
    ("I don't understand.", "මට තිබෙනවාද?"),
    ("Can you repeat that, please?", "කරුණාකර එය ආරම්භ කරන්නෙ කුමක්ද?"),
    ("Can you speak more slowly?", "කරුණාකර මොඩියුලු කරන්නෙ කුමක්ද?"),
    ("I am sorry, I don't speak Sinhala.", "කම්කොටුවේ, මම සිංහල කතා කරන්නෙ නැත."),
    ("Could you write it down, please?", "කරුණාකර එය ලියන්නෙ කුමක්ද?"),
```

```
("I need to practice my Sinhala.", "මම මගේ සිංහල විස්තර හැදුවේ අවශ්‍ය වේ."),
("What is this?", "මේ කුමක්ද?"),
("This is a...", "මෙය ... ය."),
("Where are we?", "අපි කොහෙද?"),
("We are here.", "අපි මෙතැනින්ද."),
("What is your phone number?", "ඔබගේ දුරකථන අංකය කුමක්ද?"),
("My phone number is...", "මගේ දුරකථන අංකය ... වෙයි."),
("Where do you live?", "ඔබ කොහෙදදී ජීවත්ද?"),
("I live in...", "මම ... ජීවිතයට සිටියි."),
("Do you have any siblings?", "ඔබට කිසිදු සහෝදරයක් තිබෙද?"),
("Yes, I have...", "ඔව්, මට ... තිබෙද."),
("No, I don't have any siblings.", "නැත, මට කිසිවක් නැත."),
("What is your occupation?", "ඔබගේ වැඩිපුරමිකරුවන්ට ඔබව අදාලව කියනවාද?"),
("I work as a...", "මම ... වැඩිපුරමිකරු ලෙස වැඩිපුරම් වෙයි."),
("Are you married?", "ඔබ විවාහයට ඇත්තේද?"),
("Yes, I am married.", "ඔව්, මට විවාහයට ඇත්තේ."),
("No, I am not married.", "නැත, මට විවාහයට නැත."),
("Do you have any children?", "ඔබට කිසිවක් ළමයින් තිබෙද?"),
("Yes, I have... children.", "ඔව්, මට ... අයිතියන් තිබෙද."),
("No, I don't have any children.", "නැත, මට කිසිවක් අයිතියක් නැත."),
("What is your favorite food?", "ඔබට ප්‍රියතම ආහාර කුමක්ද?"),
("My favorite food is...", "මගේ ප්‍රියතම ආහාරය ... ය."),
("What is your favorite color?", "ඔබට ප්‍රියතම වර්ණය කුමක්ද?"),
("My favorite color is...", "මගේ ප්‍රියතම වර්ණය ... ය."),
("What is your hobby?", "ඔබට විවාහකම කුමක්ද?"),
("My hobby is...", "මගේ විවාහකම ... ය."),
("What is your dream?", "ඔබට සිත්තමක් කුමක්ද?"),
("My dream is...", "මගේ සිත්තම ... ය."),
("What is your goal?", "ඔබට අත්අඩංගුවක් කුමක්ද?"),
("My goal is...", "මගේ අත්අඩංගුව ... ය."),
("Apple", "ඇපල්"),
("Banana", "බැනානා"),
("Orange", "අරන්ජ්"),
("Grapes", "කුරුඳු"),
("Pineapple", "පින්කාපල්"),
("Papaya", "පෙපාල්"),
("Watermelon" "ගොඩක්කා"),
("Coconut", "පොල්"),
("Peach", "පීච්"),
("Strawberry", "ස්ට්‍රොබෙරි"),
("Cherry", "චෙරි"),
("Lemon", "ලිමෝන්"),
("Pomegranate", "අන්නා"),
("Avocado", "ඇවොකාඩෝ"),
("Cucumber", "කෙලා"),
("Carrot", "කැරට්"),
("Tomato", "තමන්"),
("Potato", "අලුත්කාසි"),
("Onion", "කොම්බස්"),
("Garlic", "අලු"),
("Ginger", "කිංගර්"),
("Spinach", "පාලම්මිල්"),
("Broccoli", "බ්‍රොකෙලි"),
("Cabbage", "ගෝස"),
("Lettuce", "ලෙට්ස්"),
("Pumpkin", "වට්ටක්කා"),
("Peas", "කරටක්කා"),
("Beans", "බන්දු"),
("Corn", "කොන්"),
("Rice", "බෝල"),
("Bread", "පාන්"),
("Butter", "බටර්"),
("Cheese", "චීස්"),
("Milk", "කිරි"),
("Yogurt", "තාලු"),
("Egg", "බිත්තර"),
("Chicken", "කුකුල්"),
("Beef", "ගෙවිල්"),
("Pork", "පොර්ක්"),
("Fish", "මීනා"),
("Shrimp", "ඉන්දිකාව"),
("Crab", "කකුල්"),
("Lobster", "ලොබිස්ටර්"),
("Squid", "කැලි"),
("Octopus", "ඉන්දිකාව"),
("Rabbit", "දුරුම්බර"),
("Deer", "මුහුදු"),
("Goat", "කටුවක්"),
("Sheep", "ලියා"),
("Cow", "ගවුව"),
("Horse", "බොල"),
("Elephant", "අලුන්ස්"),
("Tiger", "වළිය"),
```

```
("Lion", "සිංහ"),
("Leopard", "කටුවක්"),
("Cheetah", "හකුරු"),
("Giraffe", "ගිරාෆ්"),
("Zebra", "සීබ්‍රා"),
("Monkey", "මුංකුව"),
("Gorilla", "ගොරිලා"),
("Orangutan", "ඔරන්ගුටාන්"),
("Chimpanzee", "චිම්පැන්සී"),
("Kangaroo", "කැන්ගරුව"),
("Koala", "කෝලා"),
("Panda", "පැන්ඩා"),
("Penguin", "පැන්ඩු"),
("Seal", "දැවුම්බර"),
("Dolphin", "ඩොල්ෆින්"),
("Whale", "වැල්"),
("Shark", "සරක්"),
("Turtle", "කටුවක්"),
("Frog", "මුඩුවක්"),
("Snake", "නයියා"),
("Lizard", "මැටික්කාව"),
("Crocodile", "කොරලෝ"),
("Alligator", "පපල්ක්ස්"),
("Bird", "කුකුළා"),
("Eagle", "ඇගලී"),
("Hawk", "කුලාන්ට"),
("Falcon", "සුරකුල්"),
("Owl", "කුරුළා"),
("Pigeon", "කුලුකුලා"),
("Parrot", "නිල්හා"),
("Peacock", "නිල්හා"),
("Swan", "කුකුළා"),
("Duck", "මහදුනා"),
("Goose", "ගූස්"),
("Hen", "දුරුම්බරයින්"),
("Rooster", "පියාකුව"),
("Sparrow", "කුකුළා"),
("Bat", "වෙව්වක්"),
("Butterfly", "බටලීන්"),
("Bee", "මැදුරු"),
("Ant", "මහරු"),
("Spider", "අලියා"),
("Mosquito", "මේකේන්"),
("Fly", "මාලිමම"),
("Cockroach", "නැවාල"),
("Dragonfly", "කොට්ටක්කා"),
("Ladybug", "කිකිපාවා"),
("Centipede", "මහරුන්ගාරයා"),
("Snail", "පූකා"),
("Worm", "පොලිය"),
("Caterpillar", "කිරිප්පිය"),
("Cricket", "කන්ටුව"),
("Grasshopper", "කොණ්ණා"),
("Beetle", "බීටල්"),
("Scorpion", "කලපුරු"),
("Spider", "අලියා"),
("Earthworm", "භූමියා"),
("Dragon" "මල්ලවාසිකා"),
("Unicorn", "එල්වෝ"),
("Phoenix", "පිනිස්"),
("Mermaid", "මමයිර්මේඩ්"),
("Elf", "එල්ෆ්"),
("Fairy", "ෆෙයිරි"),
("Goblin" "ගොබ්ලින්"),
("Giant", "විහාලයින්"),
("Troll", "ට්‍රොල්"),
("Wizard", "විසාරවන්"),
("Witch", "විච්"),
("Warlock", "වෝල්ක්"),
("Vampire", "වැම්පයර්"),
("Zombie", "සොම්බි"),
("Ghost", "දෑස"),
("Skeleton", "කණ්ඩායම්"),
("Mummy", "මම්"),
("Werewolf", "වීරවෝල්ෆ්"),
("Yeti", "යෙටි"),
("Alien", "නායක"),
("Robot", "රොබෝට්"),
("Dinosaur", "දන්දික්සෝරා"),
("UFO", "UFO"),
("Rocket", "කන්දකාමිකා"),
("Spaceship", "විදුලිය"),
("Astronaut", "නියමිතයා"),
```

```
        ("Meteor", "මීටියෝරය"),
        ("Comet", "කෝමෙට්"),
        ("Black hole", "කළු ගර්මය"),
        ("Galaxy", "ගැලළු"),
        ("Star", "තරු"),
        ("Sun", "ඉර"),
        ("Moon", "ගිනි"),
        ("Planet", "ලෝකය"),
        ("Earth", "පාලනය"),
        ("Mars", "මර්ස්"),
        ("Venus", "වීනස්"),
        ("Jupiter", "ජුපිටර්"),
        ("Saturn", "ශක්නා"),
        ("Uranus", "යුරේනස්"),
        ("Neptune", "නේප්ටුන්"),
        ("Pluto", "ප්ලූ"),
       ("Standard deviation", "සම්මත ද්විජ්තිය"),
        ("Alpha", "අල්ෆා"),
        ("Beta", "බීටා"),
        ("Gamma", "ගැමා"),
        ("Delta", "ඩෙල්ටා"),
        ("Theta", "තෙටා"),
        ("Rho", "රෝ"),
        ("Vega","වෙගා"),
        ("Hello", "හෙලෝ"),
        ("Goodbye", "ගොඩක් යාම"),
        ("Yes", "ඔව්"),
        ("No", "නැත"),
        ("Please", "කරන්න"),
        ("Thank", "ඔබට ස්තුතියි"),
        ("Sorry", "කාලෙකට"),
        ("Excuse", "වඩා අසමත්"),
        ("Me", "මට"),
        ("How", "කොහොම"),
        ("Are", "අය"),
        ("You", "ඔය"),
        ("I'm", "මම"),
        ("Fine", "හොඳයි"),
        ("Thank", "ඔබට"),
        ("What", "මොකද"),
        ("Your", "ඔයාගේ"),
        ("Name", "නම"),
        ("Nice", "හොඳයි"),
        ("Meet", "ලැබුනා"),
        ("Where", "කොහොමද"),
        ("From", "සිටියා"),
        ("Old", "පැරණියි"),
        ("Years", "වසර"),
        ("Do", "කරන්න"),
        ("Weather", "කාලාන්තය"),
        ("Sunny", "සුදු දුරු"),
        ("Raining", "පර්වර්වයි"),
        ("Cloudy", "අව්වයි"),
        ("Windy", "ගිහිල්ලා"),
        ("Hungry", "ආහාරයි"),
        ("Thirsty", "දිසායි"),
        ("Tired", "කාලෙකට බොහොමයි"),
        ("Bored", "කියලා අහන්න හිතුවා"),
        ("New", "නව"),
        ("Nothing", "කිසිදු නැත"),
        ("Let's", "අපි"),
        ("Go", "යන්න"),
        ("Don't", "නැත"),
        ("Know", "දන්නවා"),
        ("Great", "මහා"),
        ("Terrible","ක්ෂණික"),
        ("Love", "ආදරෙයි"),
        ("Miss", "අලුත් වෙනවා"),
        ("Need", "අවශ්‍යයි"),
        ("Help", "උදව් වන්න"),
        ("Lost", "අහසයි"),
        ("Congratulations", "සුබ පැතුම්"),
        ("Happy", "සම්පූර්ණ"),
        ("Birthday", "උපන් වන දිනය"),
        ("Merry", "පොඩ්ඩන්ද"),
        ("Christmas", "නත්තල් උල්ලංඝාව"),
        ("New Year", "නව වසර"),
        ("Luck", "වාසය"),
        ("Luck", "ක්රියාන්විතය"),
        ("Forgive", "උත්තර කරනවා"),
        ("Happy", "සම්පූර්ණ"),
        ("Love", "ආදරෙයි"),
```

```
        ("Miss", "අලුත් වෙනවා"),
        ("Sorry", "කාලෙකට"),
        ("Need", "අවශ්‍යයි"),
        ("Lost", "අහාසයි"),
        ("Congratulations", "සුබ පැතුම්"),
        ("Happy", "සම්පූර්ණ"),
        ("Birthday", "උපන් වන දිනය"),
        ("Merry", "පොඩ්ඩන්ද"),
        ("Christmas", "නත්තල් උල්ලංඝාව"),
        ("New Year", "නව වසර"),
        ("Luck", "වාසය"),
        ("Luck", "ක්‍රියාන්විතය"),
        ("Forgive", "උත්තර කරනවා"),
        ("I'm", "මම"),
        ("Lost", "අහාසයි"),
        ("I'm", "මම"),
        ("Lost", "අහාසයි"),
        ("Sorry", "කාලෙකට"),
        ("I", "මට"),
        ("Forgive", "උත්තර කරනවා"),
        ("I", "මට"),
        ("Fish", "මෙස"),
        ("Sea", "සමුගන්න"),
        ("River", "ගඟ"),
        ("Mountain", "කඳුරැවෙල"),
        ("Tree", "ගස"),
        ("Flower", "මල්"),
        ("Sun", "අල"),
        ("Moon", "ගම"),
        ("Star", "තරු"),
        ("Sky", "තුන්කාලය"),
        ("Cloud", "නිවාස"),
        ("Rain", "මල්"),
        ("Wind", "වාර්තාව"),
        ("Thunder" ,"අහල්"),
        ("Lightning" ,"විලාසය"),
        ("Fire", "ගිනි"),
        ("Earth", "බුද්ධ"),
        ("Sand", "කඳුලු"),
        ("Stone", "ගල්"),
        ("House", "ගෙදර"),
        ("School", "පාසල"),
        ("Hospital", "රෝහල"),
        ("Market", "වෙළඳපොළ"),
        ("Park", "උද්‍යානය"),
        ("Street", "වීදුරු"),
        ("Car", "මෝටර්"),
        ("Bicycle", "වෙල්කඩය"),
        ("Bus", "බස්"),
        ("Train", "දුම්රිය")
    ]

    # File name
    file_name = "english_sinhala_pairs.csv"

    # Writing translation pairs to CSV file
    with open(file_name, mode='w', newline='', encoding='utf-8') as file:
        writer = csv.writer(file)
        writer.writerows(translation_pairs)

    print(f"CSV file '{file_name}' has been created with English-Sinhala translation pairs.")

    # Custom Dataset class
    class TranslationDataset(Dataset):
        def __init__(self, file_path):
            self.df = pd.read_csv(file_path)

        def __len__(self):
            return len(self.df)

        def __getitem__(self, idx):
            source, target = self.df.iloc[idx]
            return source, target

    # Define custom collate function
    def custom_collate(batch):
        sources, targets = zip(*batch)

        # Create English vocabulary
        english_vocab = create_english_vocabulary(translation_pairs)

        # Convert English sentences to integer lists using vocabulary lookup
        source_seqs = [convert_to_intlist(source, english_vocab) for source in sources]
```

```
        target_seqs = [tokenize_si(target) for target in targets]  # Sinhala remains tokenized strings

        padded_sources = pad_sequence(source_seqs, batch_first=True)
        padded_targets = pad_sequence(target_seqs, batch_first=True)
        return padded_sources, padded_targets

    # Function to convert English sentence to integer list using vocabulary lookup
    def convert_to_intlist(text, vocab):
        # Create a vocabulary mapping (word -> integer index)
        word2idx = {word: idx for idx, word in enumerate(vocab)}
        # Get the index of the '<UNK>' token if it exists, otherwise assign it to the last index
        unk_idx = word2idx.get('<UNK>', len(word2idx) - 1)
        # Convert each word in the sentence to its index in the vocabulary
        return [word2idx.get(word, unk_idx) for word in spacy_en.tokenizer(text)]

    # Tokenize function for Sinhala (remains unchanged)
    def tokenize_si(text):
        return text.split()

    # Initialize spacy for English
    spacy_en = spacy.load('en_core_web_sm')

    # Function to create English vocabulary
    def create_english_vocabulary(text_pairs):
        # Extract all unique English words from the translation pairs
        words = set()
        for source, _ in text_pairs:
            words.update(spacy_en.tokenizer(source))
        # Add the padding token (e.g., <pad> with index 0)
        words.add("<pad>")  # You can choose a different padding token if needed
        # Convert the set of words to a list and sort for efficient indexing later
        english_vocab = sorted(list(words))
        return english_vocab

    # Model parameters
    SRC_VOCAB_SIZE = len(create_english_vocabulary(translation_pairs))
    TRG_VOCAB_SIZE = len(set(word for pair in translation_pairs for word in tokenize_si(pair[1])))
    D_MODEL = 256
    N_HEAD = 4
    NUM_ENCODER_LAYERS = 3
    NUM_DECODER_LAYERS = 3  # Add the number of decoder layers

    # Model Architecture
    class TransformerModel(nn.Module):
        def __init__(self, src_vocab_size, trg_vocab_size, d_model, n_head, num_encoder_layers, num_decoder_layers):
            super(TransformerModel, self).__init__()
            self.model_type = 'Transformer'
            self.src_embedding = nn.Embedding(src_vocab_size, d_model)
            self.trg_embedding = nn.Embedding(trg_vocab_size, d_model)
            self.transformer = Transformer(d_model=d_model, nhead=n_head, num_encoder_layers=num_encoder_layers,
                                           num_decoder_layers=num_decoder_layers)
            self.fc_out = nn.Linear(d_model, trg_vocab_size)

        def forward(self, src, trg):
            src_embedded = self.src_embedding(src)
            trg_embedded = self.trg_embedding(trg)
            output = self.transformer(src_embedded, trg_embedded)
            output = self.fc_out(output)
            return output

    # Initialize DataLoader
    dataset = TranslationDataset(file_name)
    dataloader = DataLoader(dataset, batch_size=64, shuffle=True, collate_fn=custom_collate)

    # Initialize model, criterion, and optimizer
    model = TransformerModel(SRC_VOCAB_SIZE, TRG_VOCAB_SIZE, D_MODEL, N_HEAD, NUM_ENCODER_LAYERS, NUM_DECODER_LAYERS)
    criterion = nn.CrossEntropyLoss(ignore_index=0)  # Assuming 0 is the index for padding token
    optimizer = optim.Adam(model.parameters(), lr=0.0005)

    # Training loop
    device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
    model.to(device)
    model.train()
    for epoch in range(10):  # Change the number of epochs as needed
        epoch_loss = 0
        for src, trg in dataloader:
            src, trg = src.to(device), trg.to(device)
            optimizer.zero_grad()
            output = model(src, trg[:, :-1])  # Exclude <eos> from inputs
            output_dim = output.shape[-1]
            output = output.contiguous().view(-1, output_dim)
            trg = trg[:, 1:].contiguous().view(-1)
            loss = criterion(output, trg)
```

```
            loss = criterion(output, trg)
            loss.backward()
            optimizer.step()
            epoch_loss += loss.item()
        print(f'Epoch: {epoch+1} | Loss: {epoch_loss / len(dataloader)}')
```