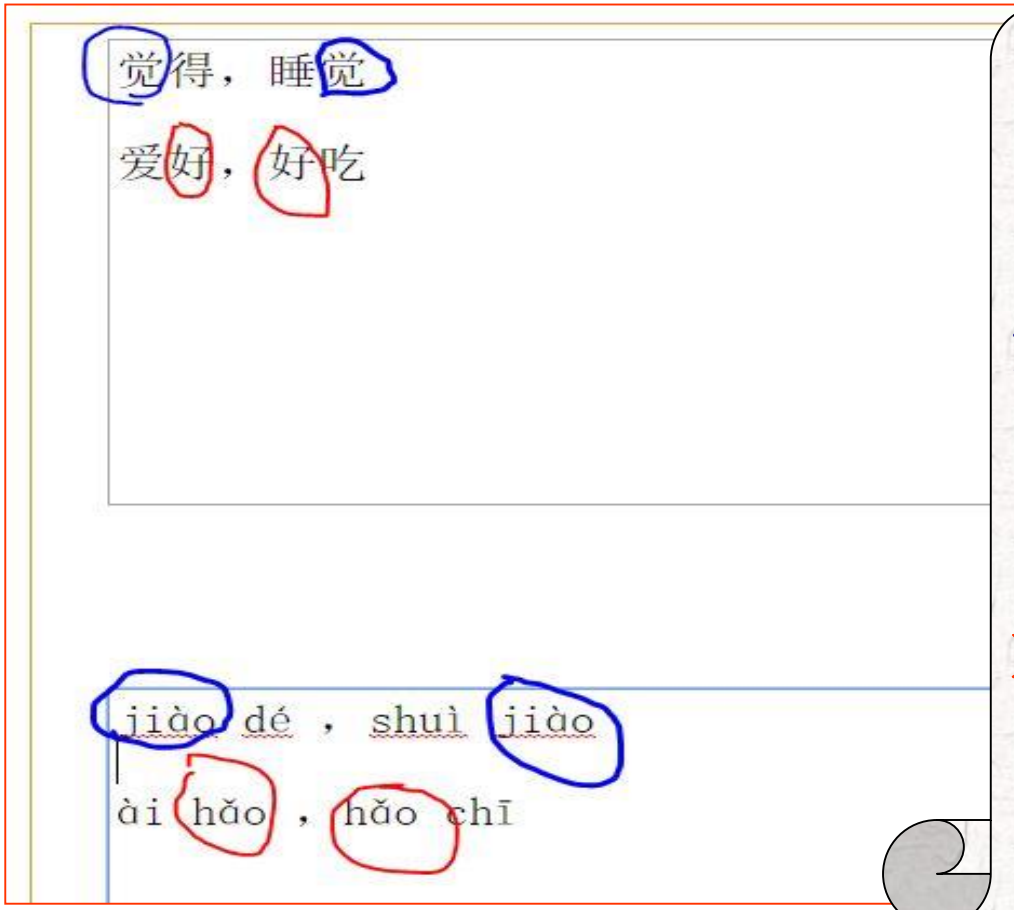
A Tool for Converting Chinese Characters to Pinyin

Hongchen Wu

I. Why ?

1. to indicate tones in Pinyin

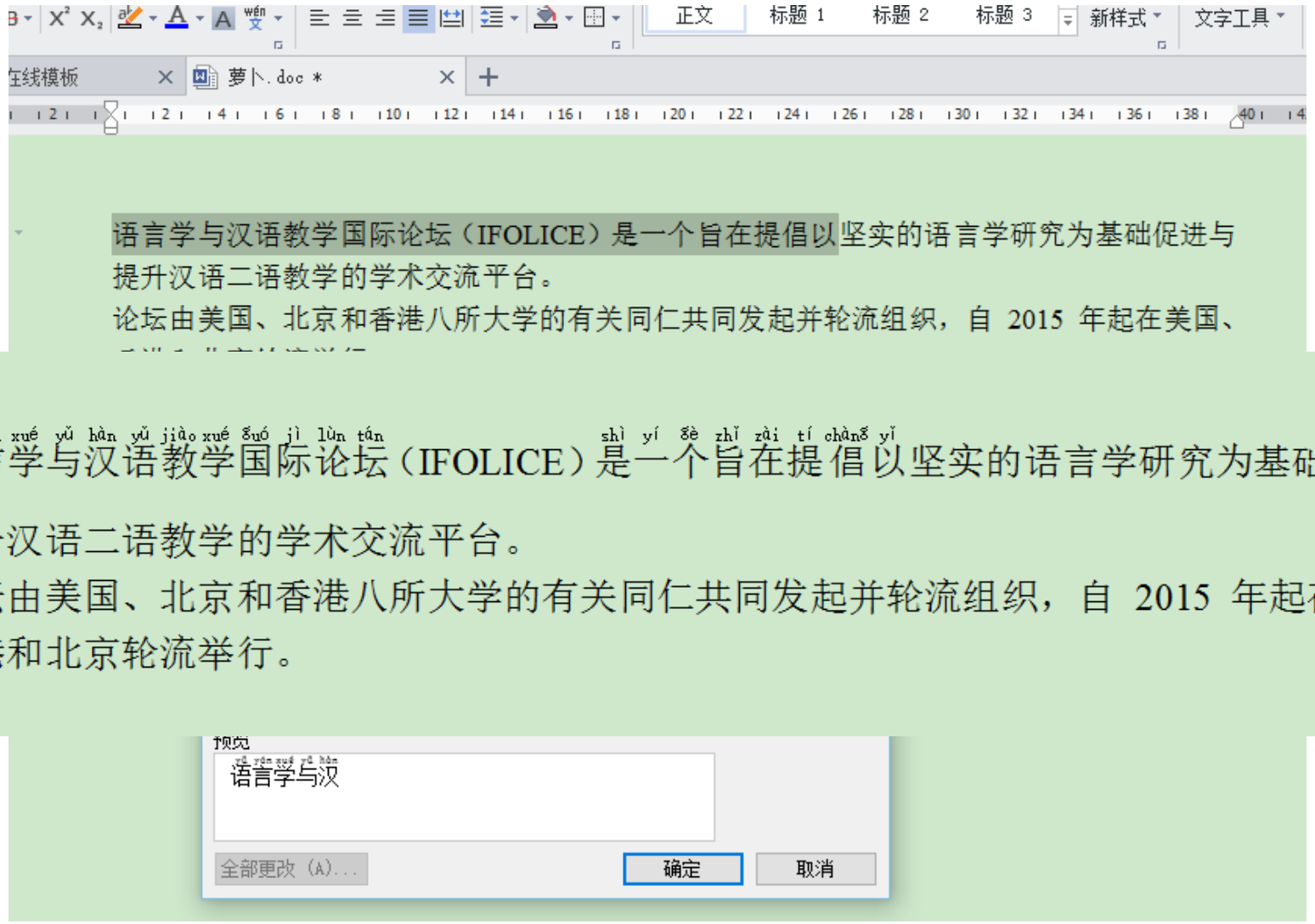
- 2. to recoginize "duo yin zi"



觉得,jué de, ' to think'
睡觉, shuì jiào, 'sleep'

爱好, ài hào, 'hobby'
好吃,hǎo chī 'delicious'

- 3. to easily edit/ prepare teaching materials offline



语言学与汉语教学国际论坛（IFOLICE）是一个旨在提倡以坚实的语言学研究为基础促进与提升汉语二语教学的学术交流平台。

论坛由美国、北京和香港八所大学的有关同仁共同发起并轮流组织，自 2015 年起在美国、香港和北京轮流举行。

yǔ yán xué yǔ hàn yǔ jiāo xué guó jì lùn tán （ I F O L I C E ） shì yī gè zhǐ zài tí chàng yǐ jiān shí de yǔ yán xué yán jiū wéi jī chǔ cù jìn yǔ tí shēng hàn yǔ èr yǔ jiāo xué de xué shù jiāo liú píng tái 。

lùn tán yóu měi guó 、 běi jīng hé xiāng gǎng bā suǒ dà xué de yǒu guān tóng rén gòng tóng fā qǐ bìng lún liú zǔ zhī ， zì 2 0 1 5 nián qǐ zài měi guó 、 xiāng gǎng hé běi jīng lún liú jǔ háng 。

II. How?

- 1. Database

  two official documents

-- “duo yin zi” (called “异读词审音表”)

-- 3500 most frequently used Chinese characters

- 2. Package

  cjklib, nltk

- 3. Input & Ouput:

  Input:

-- Chinese character(s)/sentence(s)

-- <string>

  Output:

-- corresponding standard pronunciation of the input

-- <string>

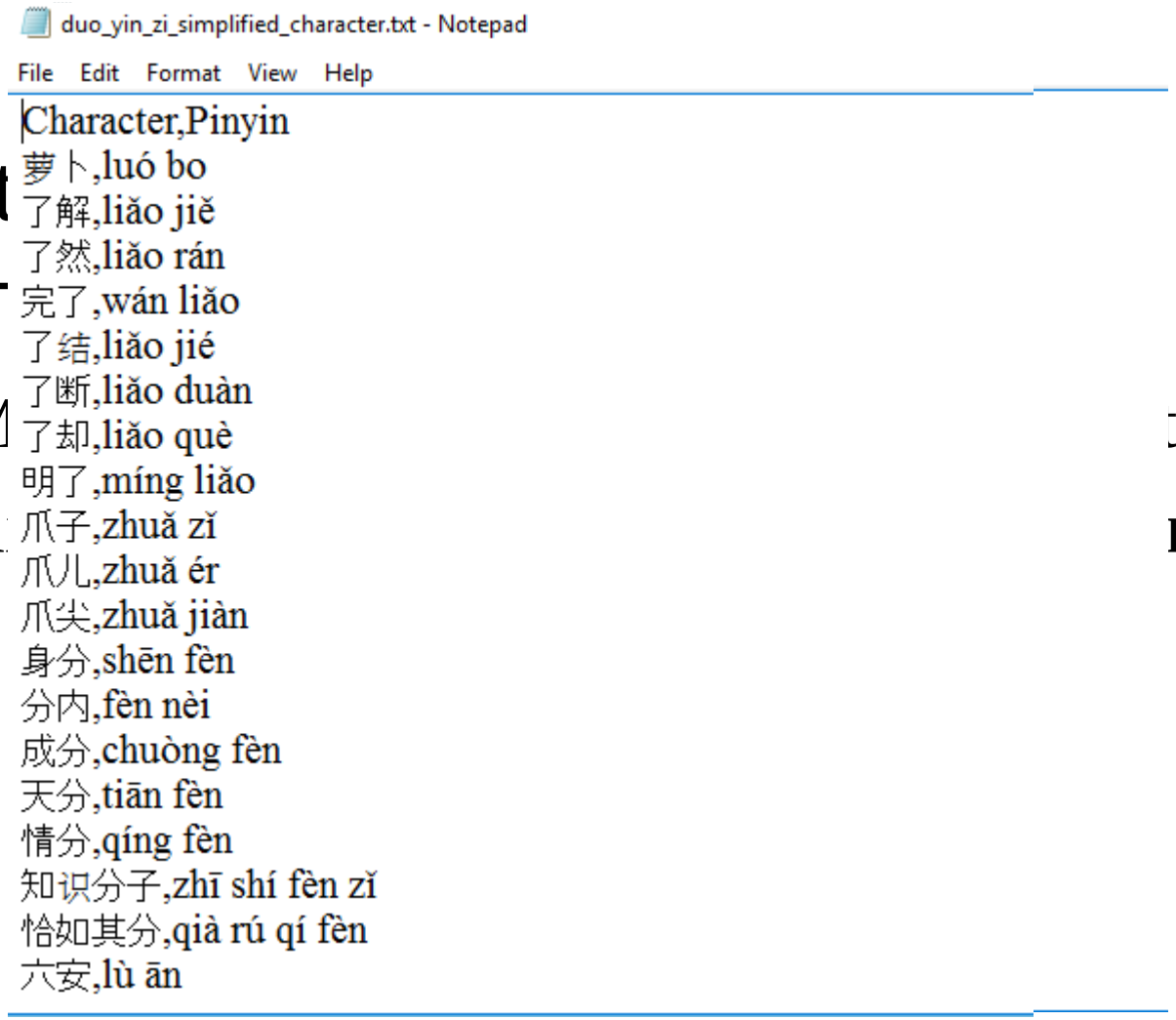
- 4. St

(1) T

3500_M ters.txt

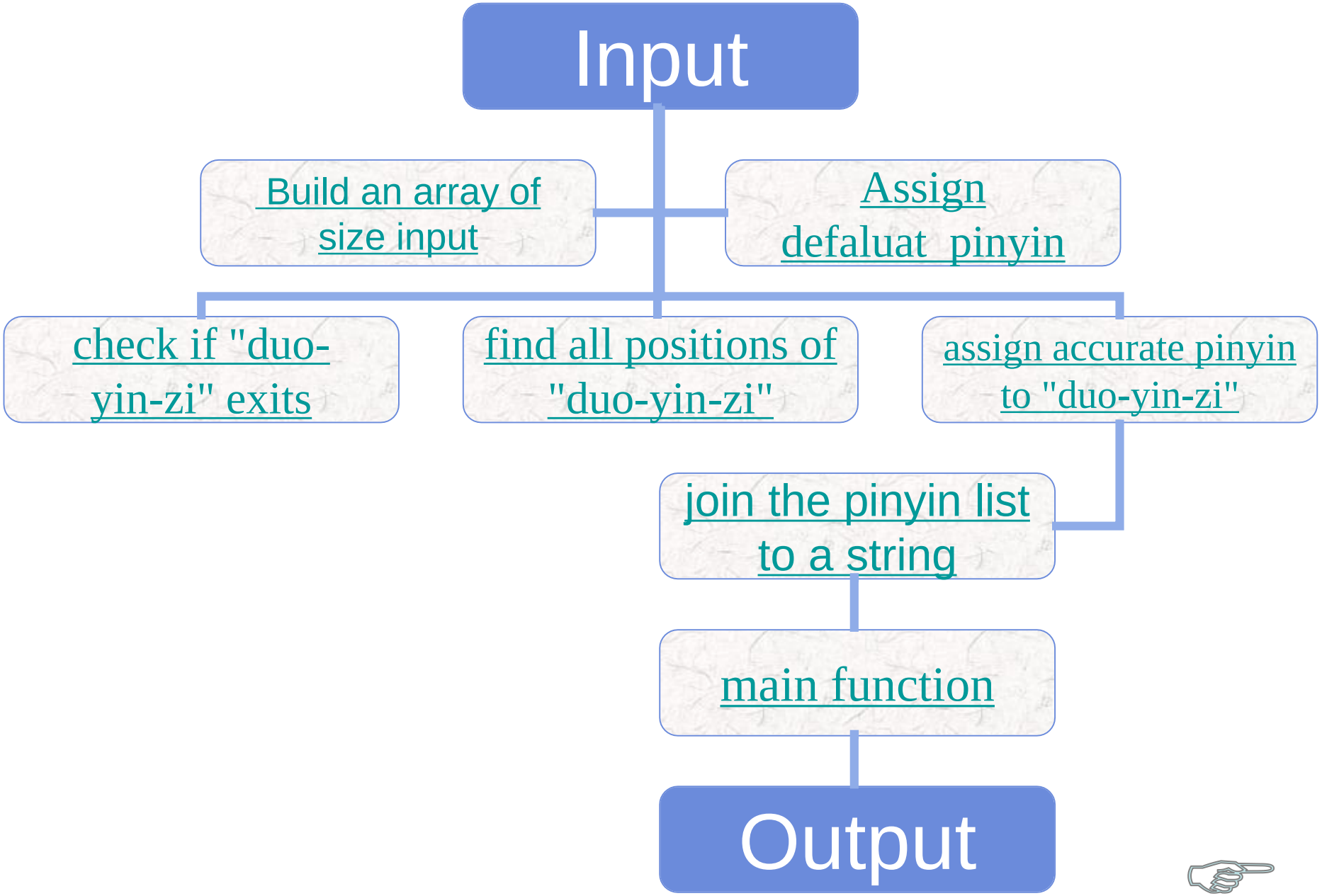
100_Ch nciation

s.txt



- (2) create dictinonaries

```
## create a dict with less common pronunciation ##
file100 = open("100_Chinese_Characters_with_Multiple_Pronunciations.txt",encoding="UTF-8"))
dict100 = nltk.defaultdict(str)
for line in file100:
        y = line.split(",")
        character2 = y[0]
        pinyin2 = y[1]
        y1 = len(pinyin2)-1
        pinyin2 = pinyin2[0:y1]
        dict100[character2] = pinyin2
```


Input
Build an array of size input
Assign defaluat pinyin
check if "duo-yin-zi" exits
find all positions of "duo-yin-zi"
assign accurate pinyin to "duo-yin-zi"
join the pinyin list to a string
main function
Output

① Build an array of size input

```
input_len = len(Chinese)
sentencePinYin = [None] * input_len
```

- ② For all null entries in the string, fill it with default pinyin

```
positionCounter = 0
for word in Chinese:
        sentencePinYin[positionCounter] = dict3500[word]
        positionCounter = positionCounter + 1
```

- ③ Find ‘duo-yin-zi' in the input and where they exist in the string

```
for key in dict100.keys():
            #if it exists, get the duoyinzi pinyin and overwrite the default pinyin
            if key in Chinese:
                    #find all indexes where the phrase exists using find() function
                    location = -1 #start at -1
                    locations = [] #keep track of all locations where phrase is found
                    #loop through Chinese input, finding all phrases
                    while True:
                            location = Chinese.find(key, location + 1) #location where
                            # use find() instead of index() to avoid "error" message: i
                            if location == -1:
                                    break #if not found, break loop
                            else:
                                    locations.append(location) #store location
```

- ④ assign the accurate pinyin of the 'duo-yin-zi' to where they exist in the string

```
pinYinSplit = dict100[key].split(" ") #break phrase into words
for location in locations: #for each location where phrase was found
        for pinYin in range(len(pinYinSplit)): #assign sentencePinYin[location] to first word,
                sentencePinYin[location] = pinYinSplit[pinYin]
                location = location + 1
```

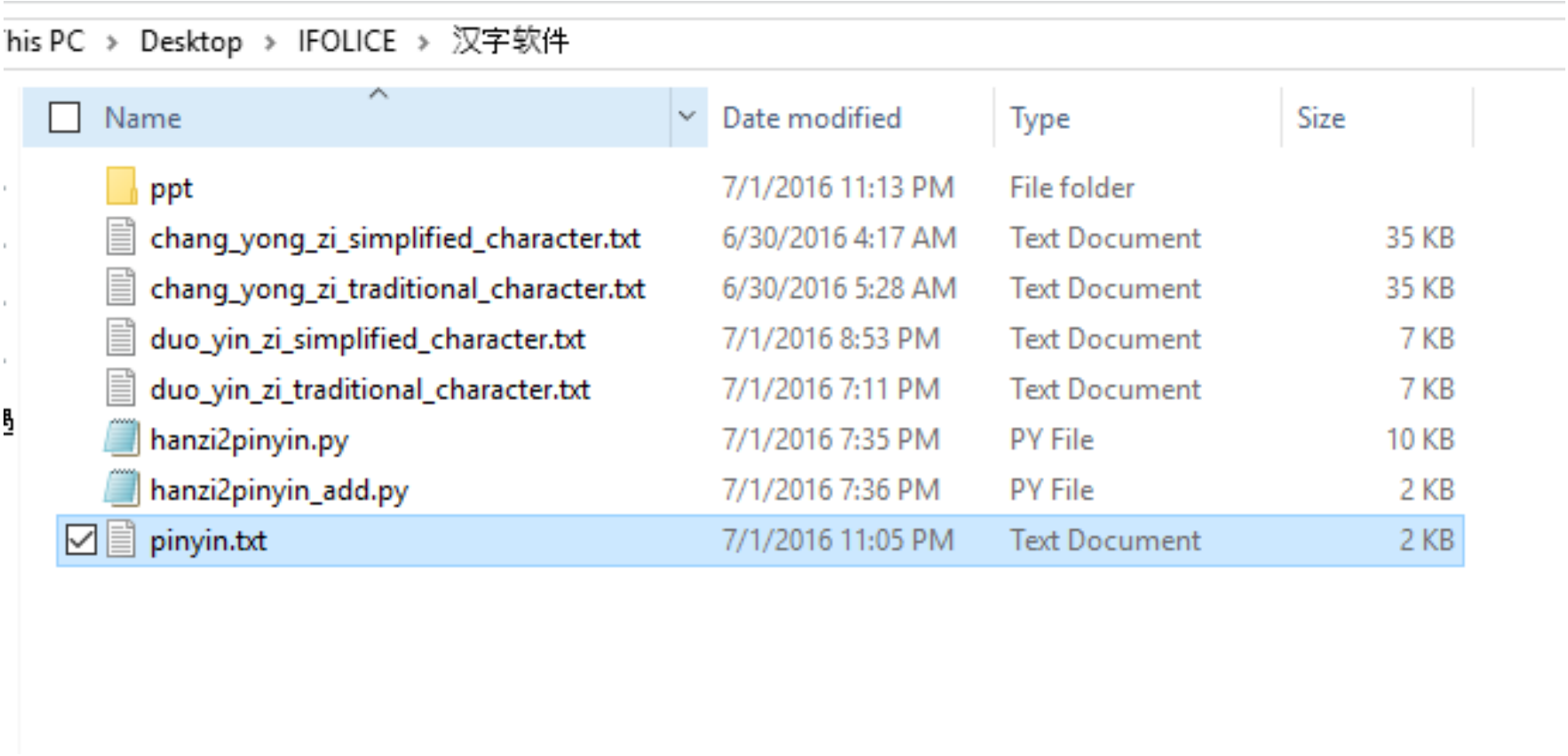
- ⑤ print pinyin as a continuious string

```
return (" ".join(sentencePinYin))
```

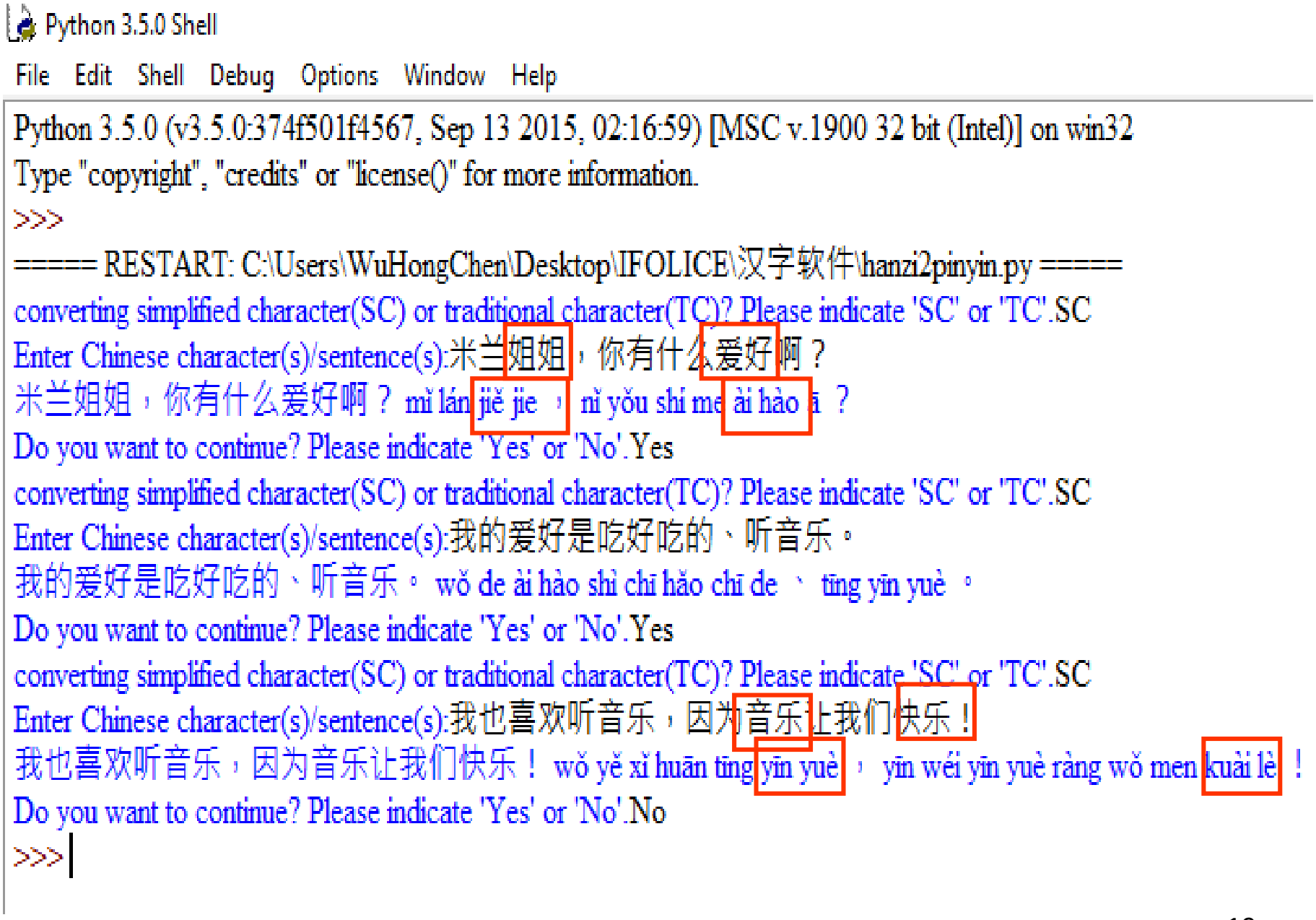
```
def main():
    while True:
        zi_ti = input("converting simplified character(SC) or traditional character(TC)? Please indicate 'SC' or 'TC'.")
        Chinese = input("Enter Chinese character(s)/sentence(s):")
        if zi_ti == "SC":
            #print(Chinese)
            mydata = hanzi2pinyin_simplified(Chinese)
            #print(mydata)
            myfile = open('pinyin.txt', 'ab')
            myfile.write(Chinese.encode('utf-8') + mydata.encode("utf-8") )
            myfile.close()
            myfile = open('pinyin.txt', 'a')
            myfile.write("\n")
            print(Chinese, hanzi2pinyin_simplified(Chinese))
        else:
            mydata = hanzi2pinyin_traditional(Chinese)
            myfile = open('pinyin.txt', 'ab')
            myfile.write(Chinese.encode('utf-8') + mydata.encode("utf-8") )
            myfile.close()
            myfile = open('pinyin.txt', 'a')
            myfile.write("\n")
            print(Chinese, hanzi2pinyin_traditional(Chinese))
        D =input("Do you want to continue? Please indicate 'Yes' or 'No'.")
        if D=="Yes":
            continue
        else:
            break
```

ile

III. Demo


his PC › Desktop › IFOLICE › 汉字软件
Name
Date modified
Type
Size
ppt
7/1/2016 11:13 PM
File folder
chang_yong_zi_simplified_character.txt
6/30/2016 4:17 AM
Text Document
35 KB
chang_yong_zi_traditional_character.txt
6/30/2016 5:28 AM
Text Document
35 KB
duo_yin_zi_simplified_character.txt
7/1/2016 8:53 PM
Text Document
7 KB
duo_yin_zi_traditional_character.txt
7/1/2016 7:11 PM
Text Document
7 KB
hanzi2pinyin.py
7/1/2016 7:35 PM
PY File
10 KB
hanzi2pinyin_add.py
7/1/2016 7:36 PM
PY File
2 KB
pinyin.txt
7/1/2016 11:05 PM
Text Document
2 KB

```
===== RESTART: C:\Users\WuHongChen\Desktop\IFOLICE\汉字软件\hanzi2pinyin.py =====
converting simplified character(SC) or traditional character(TC)? Please indicate 'SC' or 'TC'.SC
Enter Chinese character(s)/sentence(s):您好！您觉得我的拼音软件做得好吗？
您好！您觉得我的拼音软件做得好吗？ nín hǎo ！ nín jué de wǒ de pīn yīn ruǎn jiàn zuò dé hǎo ma
？
Do you want to continue? Please indicate 'Yes' or 'No'.Yes
converting simplified character(SC) or traditional character(TC)? Please indicate 'SC' or 'TC'.SC
Enter Chinese character(s)/sentence(s):我觉得你这个软件很好，很方便。
我觉得你这个软件很好，很方便。 wǒ jué de nǐ zhè gè ruǎn jiàn hěn hǎo ， hěn fāng biàn 。
```


Python 3.5.0 Shell
File Edit Shell Debug Options Window Help
Python 3.5.0 (v3.5.0:374f501f4567, Sep 13 2015, 02:16:59) [MSC v.1900 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
===== RESTART: C:\Users\WuHongChen\Desktop\IFOLICE\汉字软件\hanzi2pinyin.py =====
converting simplified character(SC) or traditional character(TC)? Please indicate 'SC' or 'TC'.SC
Enter Chinese character(s)/sentence(s):米兰姐姐，你有什么爱好啊？
米兰姐姐，你有什么爱好啊？ mǐ lán jiě jie ， nǐ yǒu shí me ài hào ā ？
Do you want to continue? Please indicate 'Yes' or 'No'.Yes
converting simplified character(SC) or traditional character(TC)? Please indicate 'SC' or 'TC'.SC
Enter Chinese character(s)/sentence(s):我的爱好是吃好吃的、听音乐。
我的爱好是吃好吃的、听音乐。 wǒ de ài hào shì chī hǎo chī de 、 tīng yīn yuè 。
Do you want to continue? Please indicate 'Yes' or 'No'.Yes
converting simplified character(SC) or traditional character(TC)? Please indicate 'SC' or 'TC'.SC
Enter Chinese character(s)/sentence(s):我也喜欢听音乐，因为音乐让我们快乐！
我也喜欢听音乐，因为音乐让我们快乐！ wǒ yě xǐ huān tīng yīn yuè ， yīn wéi yīn yuè ràng wǒ men kuài lè ！
Do you want to continue? Please indicate 'Yes' or 'No'.No
>>>

pinyin.txt - Notepad

File Edit Format View Help

您好！您觉得我的拼音软件做得好吗？nín hǎo ！ nín jué de wǒ de pīn yīn ruǎn jiàn zuò dé hǎo ma ？
我觉得你这个软件很好，很方便。wǒ jué de nǐ zhè gè ruǎn jiàn hěn hǎo ， hěn fāng biàn 。
米兰姐姐，你有什么爱好啊？mǐ lán jiě jie ， nǐ yǒu shí me ài hào ā ？
米兰姐姐，你有什么爱好啊？mǐ lán jiě jie ， nǐ yǒu shí me ài hào ā ？
我的爱好是吃好吃的、听音乐。wǒ de ài hào shì chī hǎo chī de 、 tīng yīn yuè 。
我也喜欢听音乐，因为音乐让我们快乐！wǒ yě xǐ huān tīng yīn yuè ， yīn wéi yīn yuè ràng wǒ men kuài lè ！

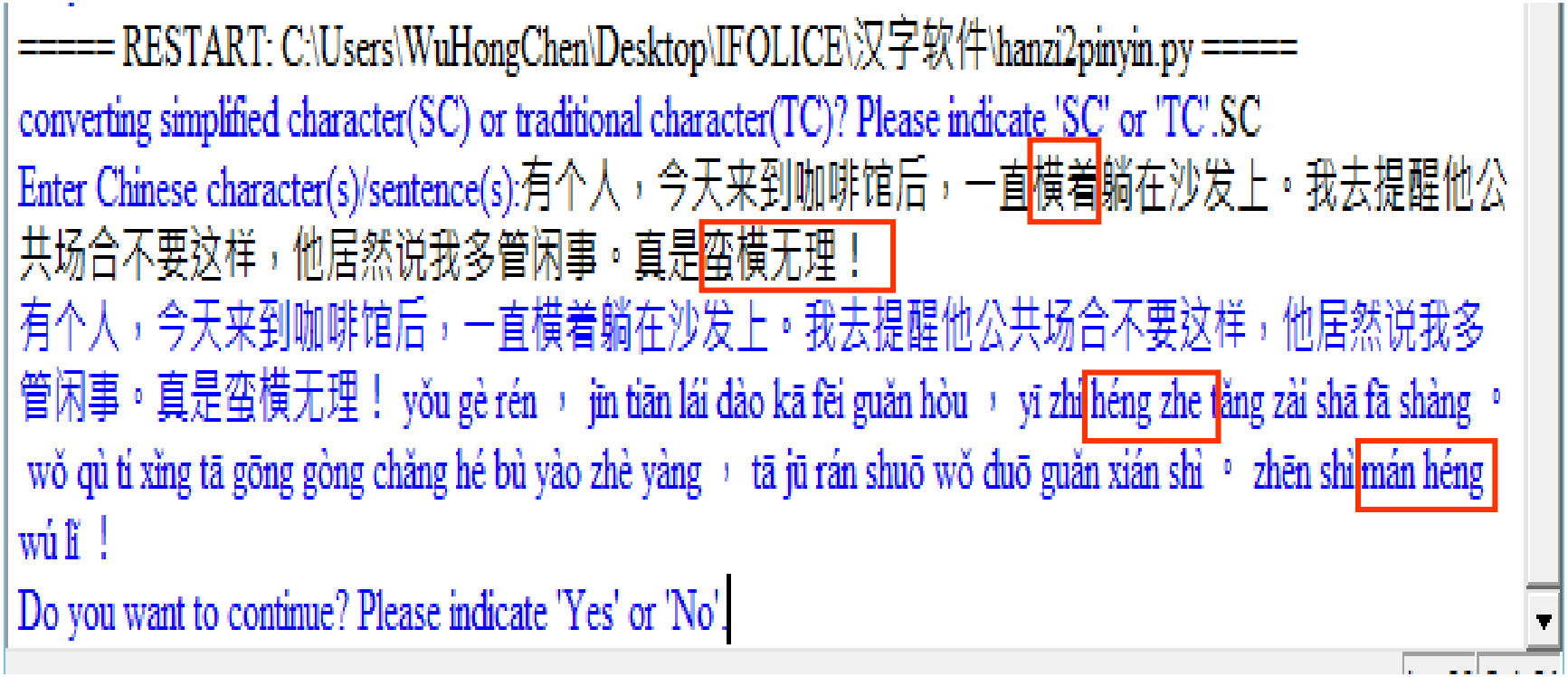
pinyin.txt - Notepad

File Edit Format View Help

您好！您觉得我的拼音软件做得好吗？nín hǎo ！ nín jué de wǒ de pīn yīn ruǎn jiàn zuò dé hǎo ma ？
我觉得你这个软件很好，很方便。wǒ jué de nǐ zhè gè ruǎn jiàn hěn hǎo ， hěn fāng biàn 。
米兰姐姐，你有什么爱好啊？mǐ lán jiě jie ， nǐ yǒu shí me ài hào ā ？
米兰姐姐，你有什么爱好啊？mǐ lán jiě jie ， nǐ yǒu shí me ài hào ā ？
我的爱好是吃好吃的、听音乐。wǒ de ài hào shì chī hǎo chī de 、 tīng yīn yuè 。
我也喜欢听音乐，因为音乐让我们快乐！wǒ yě xǐ huān tīng yīn yuè ， yīn wéi yīn yuè ràng wǒ men kuài lè ！
本文通过可接受判断测试和组句测试两种测试方法，对英语母语者和韩语母语者学习者习得“不、没”的体标记选择和语法意义的区别进行了定量研究，并通过测试结果的分析，对习得过程中的语际影响、句法-语义接口、词语习得等问题进行了探讨。běn wén tōng guò kě jiē shòu pàn duàn cè shì hé zǔ jù cè shì liǎng zhǒng cè shì fāng fǎ ， duì yīng yǔ mǔ yǔ zhě hé hán yǔ mǔ yǔ zhě xué xí zhě xí dé “ bù 、 méi ” de tǐ biāo jì xuǎn zé hé yǔ fǎ yì yì de qū bié jìn xíng le dìng liáng yán jiū ， bìng tōng guò cè shì jié guǒ de fēn xī ， duì xí dé guò chéng zhōng de yǔ jì yǐng xiǎng 、 jù fǎ - yǔ yì jiē kǒu 、 cí yǔ xí dé děng wèn tí jìn xíng le tàn tǎo 。

IV. Editable dictionary

Do you want to continue? Please indicate 'Yes' or 'No'.Yes
converting simplified character(SC) or traditional character(TC)? Please indicate 'SC' or 'TC'.SC
Enter Chinese character(s)/sentence(s):您好！我是武宏琛！认识您很高兴！
您好！我是武宏琛！认识您很高兴！ nín hǎo ！ wǒ shì wǔ hóng 琛 ！ rèn shí nín hěn gāo xìng ！
Do you want to continue? Please indicate 'Yes' or 'No'.


===== RESTART: C:\Users\WuHongChen\Desktop\IFOLICE\汉字软件\hanzi2pinyin.py =====
converting simplified character(SC) or traditional character(TC)? Please indicate 'SC' or 'TC'.SC
Enter Chinese character(s)/sentence(s):有个人，今天来到咖啡馆后，一直横着躺在沙发上。我去提醒他公共场合不要这样，他居然说我多管闲事。真是蛮横无理！
有个人，今天来到咖啡馆后，一直横着躺在沙发上。我去提醒他公共场合不要这样，他居然说我多管闲事。真是蛮横无理！ yǒu gè rén ， jīn tiān lái dào kā fēi guǎn hòu ， yì zhí héng zhe tǎng zài shā fā shàng 。 wǒ qù tí xǐng tā gōng gòng chǎng hé bù yào zhè yàng ， tā jū rán shuō wǒ duō guǎn xián shì 。 zhēn shì mán héng wú lǐ ！
Do you want to continue? Please indicate 'Yes' or 'No'.

```
def edit(edit_item):
    edit_item=input("Enter the item with the format like '一磅,yī bàng'")
    return edit_item

def main():
    while True:
        editInput = input("What kind of duo_yin_zi phrase you wan to add? simplified character(SC) or
        edit_item = edit(editInput)
        print(edit_item)
        if editInput =="SC":
            myfile=open("duo_yin_zi_simplified_character.txt","ab")
            myfile.write(edit_item.encode('utf-8'))
            myfile.close()
            myfile = open('duo_yin_zi_simplified_character.txt', 'a')
            myfile.write("\n")
            myfile.close()
        else:
            myfile=open("duo_yin_zi_traditional_character.txt","ab")
            myfile.write(edit_item.encode('utf-8'))
            myfile.close()
            myfile = open('duo_yin_zi_traditional_character.txt', 'a')
            myfile.write("\n")
            myfile.close()
        D =input("Do you want to continue? Please indicate 'Yes' or 'No'.")
        if D=="Yes":
            continue
        else:
            break
```

```
=== RESTART: C:\Users\WuHongChen\Desktop\IFOLICE\汉字软件\hanzi2pinyin_add.py ===
What kind of duo_yin_zi phrase you wan to add? simplified character(SC) or traditional character(TC)? Please indicate 'SC' or 'TC'SC
Enter the item with the format like '一磅,yi bàng'武宏琛,wǔ hóng chēn
武宏琛,wǔ hóng chēn
Do you want to continue? Please indicate 'Yes' or 'No'.No
>>>
```

Ln: 37 Col: 4

```
>>>
===== RESTART: C:\Users\WuHongChen\Desktop\IFOLICE\汉字软件\hanzi2pinyin.py =====
converting simplified character(SC) or traditional character(TC)? Please indicate 'SC' or 'TC'.SC
Enter Chinese character(s)/sentence(s):您好！我是武宏琛！认识您很高兴！
您好！我是武宏琛！认识您很高兴！ nín hǎo ！ wǒ shì wǔ hóng chēn ！ rèn shí nín hěn gāo xìng ！
Do you want to continue? Please indicate 'Yes' or 'No'.No
>>>
```

```
=== RESTART: C:\Users\WuHongChen\Desktop\IFOLICE\汉字软件\hanzi2pinyin_add.py ===
What kind of duo_yin_zi phrase you wan to add? simplified character(SC) or traditional character(TC)? Ple
ase indicate 'SC' or 'TC'SC
Enter the item with the format like '一磅,yi bàng'蛮横,mán hèng
蛮横,mán hèng
Do you want to continue? Please indicate 'Yes' or 'No'.No
>>>
===== RESTART: C:\Users\WuHongChen\Desktop\IFOLICE\汉字软件\hanzi2pinyin.py =====
converting simplified character(SC) or traditional character(TC)? Please indicate 'SC' or 'TC'.SC
Enter Chinese character(s)/sentence(s):有个人，今天来到咖啡馆后，一直横着躺在沙发上。我去提
醒他公共场合不要这样，他居然说我多管闲事。真是蛮横无理！
有个人，今天来到咖啡馆后，一直横着躺在沙发上。我去提醒他公共场合不要这样，他居然
说我多管闲事。真是蛮横无理！ yǒu gè rén ， jīn tiān lái dào kā fēi guǎn hòu ， yī zhí héng zhe tǎng zài
 shā fā shàng 。 wǒ qù tí xǐng tā gōng gòng chǎng hé bù yào zhè yàng ， tā jū rán shuō wǒ duō guǎn xián shì
 。 zhēn shi mán hèng wú lǐ ！
Do you want to continue? Please indicate 'Yes' or 'No'
```

Thank you!

hongchen.wu@stonybrook.edu