

VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY  
UNIVERSITY OF TECHNOLOGY  
FACULTY OF COMPUTER SCIENCE AND ENGINEERING



## SOFTWARE ENGINEERING (CO200B)

---

Assignment (Semester: 222)

# Xử lý từ điển đã OCR

---

**Advisors:** Quấn Thành Thơ  
**Students:** Nguyễn Phúc Minh Quân - 2110479.  
Phan Trần Minh Đạt - 2111025

Ho Chi Minh City, 28/03/2023



## Mục lục

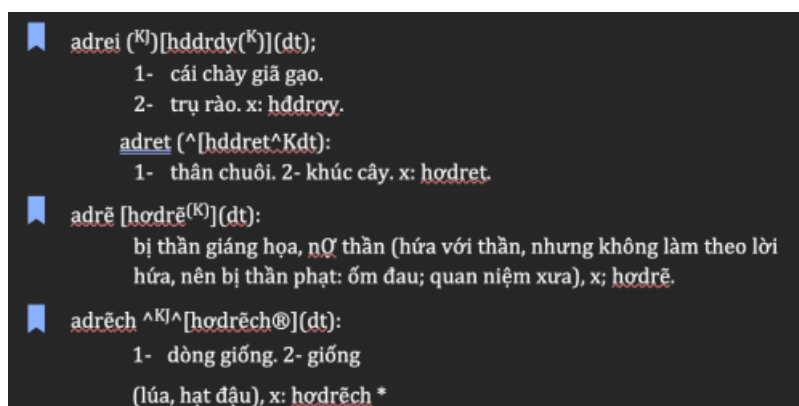
<b>1</b>	<b>Giới thiệu đề tài</b>	<b>2</b>
<b>2</b>	<b>Hướng giải quyết</b>	<b>2</b>
<b>3</b>	<b>Nội dung chi tiết</b>	<b>4</b>
3.1	Đọc nội dung từ điển 1 . . . . .	4
3.1.1	Source code . . . . .	4
3.1.2	Mô tả . . . . .	4
3.1.3	Kết quả thu được . . . . .	5
3.2	Đọc nội dung từ điển 2 . . . . .	5
3.2.1	Source code . . . . .	5
3.2.2	Mô tả . . . . .	7
3.2.3	Kết quả thu được . . . . .	7
<b>4</b>	<b>Nhận xét</b>	<b>7</b>
4.1	Những gì nhóm đã làm được . . . . .	7
4.2	Khó khăn . . . . .	7

## 1 Giới thiệu đề tài

- Cho các từ điển song ngữ Việt-Bahnar đã được số hóa (file Word). Hãy hiện thực phần mềm/công cụ để tách riêng từng cặp từ song ngữ thành một hàng và lưu dưới dạng bảng trong Excel.
- Ví dụ:  
Ở từ điển 1, "d" là danh từ

Ao âm - d	<i>Ao tonõ Ao</i>
Ao chặt -d	<i>hrăt Ao kôm</i>
Ao com lê - d	<i>plẻ Ao djăl</i>
Ao cọc - d	<i>Ao ti djăl Ao</i>
Ao cọc tay - d	<i>kojung Ao ti</i>
Ao dài - d	<i>tai Ao trôk</i>
Ao dài tay - d	<i>Ao đâm Ao</i>
Ao dơ - d	<i>pokao</i>
Ao đâm - d	<i>Ao tonỏ; ao bang bả</i>

Ở từ điển 2, "dt" là danh từ



- Yêu cầu:
  - Phần mềm có khả năng tách từ tiếng Việt và tiếng Bahnar theo từng hàng và trích xuất được từ loại tương ứng.
  - Có khả năng nhận vào nhiều file words và xuất ra chung 1 file excel.
  - Có giao diện để thuận tiện sử dụng.

## 2 Hướng giải quyết

Trong quá trình khảo sát các tập tin Word đã OCR, nhóm nhận thấy một phần dữ liệu của từ điển 1 được lưu trữ dưới dạng bảng, là dạng dữ liệu tương đối dễ thao tác và chuyển đổi. Ví dụ:

Bạch - d	Hơi
Bạch hầu - d	Pơangeh jĩ ako
Bạch tạng - d	Mong
Bài - d	B'ai
Bài giảng - d	B'ai bơ tho
Bài hát - d	B'ai hơi
Bài học - d	Tơdrong pơhrăm; b'ai
Bài sai - d	B'ai giải
bài tập - d	B'ai pơhrăm
Bài thơ - d	Nờr pơđơk

Đối với từ điển 2, nhóm nhận thấy rằng từng đoạn văn bản dịch từ tiếng Bahnar sang tiếng Việt với dạng chung là từ vựng Bahnar, cách phát âm của từ Bahnar (nếu có), loại từ và cuối cùng là định nghĩa tiếng Việt.

achom<sup>1(L)</sup>[hơchhôm®](đt): đựng  
nhau, x: hơchhôm.

achom<sup>2(I)</sup>[pơgăm®](dt):  
thuốc độc lấy từ nhựa dây  
mrei (tắm clmt độc vào mũi  
tên để bắn thú dữ, kẻ thù), x:  
pơgăm<sup>2</sup>

achốt<sup>(K)</sup>[hơchhố®](đt): 1- tù, tựa.  
2.- ngưng lại, đình chỉ (một  
công việc) 3- giới hạn. 4- nói  
dứt khoát, x: hơchốt.

achũ<sup>^</sup>[hơchũ^](đt):  
va nhẹ đầu vào. x: hơchũ.

- Để tự động hóa việc chuyển đổi các dữ liệu dạng bảng như trên sang tập tin Excel, nhóm quyết định sử dụng ngôn ngữ lập trình Python cùng với thư viện python-docx, một thư viện cho phép người dùng truy cập, tạo mới hoặc thay đổi một tập tin Word. Hướng dẫn cài đặt và sử dụng thư viện có thể tìm thấy [tại đây](#).
- Bên cạnh việc truy xuất các dữ liệu dạng bảng từ file Word bằng thư viện trên, nhóm cũng cần trích xuất từ loại sang một cột riêng biệt.
  - Đối với từ điển 1, nhóm tiến hành xử lý chuỗi trên cột đầu tiên, nhằm tách riêng phần từ vựng và từ loại thông qua ký tự ngăn cách “-” đối với từ đi.
  - Đối với từ điển 2, nhóm tiến hành đọc văn bản theo từng đoạn, từ loại sẽ đứng trước dấu “:” và ở trong dấu ngoặc đơn.
- Quá trình chuyển đổi và xử lý dữ liệu trên file Word diễn ra khá suôn sẻ, nên nhóm đã thu hoạch được một số kết quả khả quan như bên dưới.

## 3 Nội dung chi tiết

### 3.1 Đọc nội dung từ điển 1

#### 3.1.1 Source code

```
import docx
import pandas as pd
my_path = "Word/Tu_vung_doi_chieu_GiaLai/Viet Ba Na OCR/"
paths = ["tu_vung_doi_chieu_p1_done.docx", "tu_vung_doi_chieu_p2_done.docx", "tu_vung_doi_chieu_p3_done.docx"]
i = 0
# Vietnamese_data = []
# Bahnar_data = []
# word_type_data = []
for path in paths:
    file = docx.Document(my_path + path)
    Vietnamese = []
    Bahnar = []
    word_type = []
    for table in file.tables:
        for row in table.rows:
            if len(row.cells) != 2:
                continue
            try:
                VN, type = row.cells[0].text.split('-')
                Vietnamese.append(VN.strip())
                word_type.append(type.strip())
                Bahnar.append(row.cells[1].text.strip())
            except:
                pass
    path = "Excel/df" + str(i) + ".xlsx"
    data = {'language0': Vietnamese,
            'language1': Bahnar,
            'word_type': word_type}
    df = pd.DataFrame(data)
    df.to_excel(path, index = False)
    i += 1
```

#### 3.1.2 Mô tả

- Trước hết, hai thư viện *docx* và *pandas* được thêm vào chương trình. Thư viện *docx* dùng để tương tác với file Word và *pandas* dùng để chuyển đổi bảng từ vựng sau khi xử lý sang một data frame để truyền vào file Excel.
- Các file Word được truyền vào chương trình thông qua đường dẫn đến tập tin.
- Chương trình lần lượt duyệt qua các file. Với mỗi bảng trong file, chương trình sẽ trích xuất dữ liệu trong cột đầu tiên và tách chúng thành hai phần: từ vựng và từ loại.
- Đưa các nội dung đã trích xuất vào các mảng tương ứng và tạo một data frame từ chúng. Truyền data frame đó vào file Excel.

### 3.1.3 Kết quả thu được

	A	B	C	D
1	language0	language1	word_type	
2	Ái an	Tơ hực dih bắ	d	
3	Ái tình	Hực bắ	d	
4	A lô	Alô	đg	
5	Ăm	Ami`n	đg	
6	Ăm em	Ami`n oh; pôk	đg	
7	Ăm lấy	Ami`n ayo`k	dg	
8	An	Hoai; khi`	t	
9	An ninh	We`i kơchấp	t	
10	An tâm	Khi` ôh	đg	
11	An táng	B`u` bơngai lôch	đg	
12	An ủi	Pơlung	đg	
13	Ấn	An	d	
14	Ấn mang	Palôch	d	
15	Anh	Anho`ng	d	
16	Anh ấy	Sư;anho`ng anoh	d	
17	Anh bạn	Anho`ng bôl buắ	d	
18	Anh cả	Anho`ng kơdrắ	d	
19	Anh chàng	Anho`ng dăm	d	
20	Anh chị	Anho`ng mai pôm b`ắ me`	d	
21	Anh dũng	Nuih	d	
22	Anh đẹp trai	Anho`ng alăng ro`	d	
23	Anh đó	Anho`ng anoh	d	
24	Anh em	Anho`ng oh pôm b`ắ me`	d	
25	Ắc quy	Binh qui	d	
26	Ăm ắp	Blai; mơmân	1	

## 3.2 Đọc nội dung từ điển 2

### 3.2.1 Source code

```
import docx
import pandas as pd
import re
my_path = "E:/Study/CNPM/MR/Word/Tu dien Ba Na - Việt Kon Tum.docx"
file = docx.Document(my_path)
language0 = []
language1 = []
para = file.paragraphs
```

```
for i in range(96, 17126):
    try:
        index = para[i].text.find('x:')
        if index == -1:
            pass
        else:
            para[i].text = para[i].text[:index]
    except IndexError:
        pass
    try:
        x, y = file.paragraphs[i].text.split(':', 1)
        if (x == "" or y == ""):
            continue
        language0.append(x.strip())
        language1.append(y.strip())
    except:
        continue
word_type = []
for i, word in enumerate(language0):
    try:
        if word[-1] == ')':
            if word[-5] == '(':
                word_type.append(word[-4:-1])
            else:
                word_type.append(word[-3:-1])
        else:
            word_type.append("")
    except:
        word_type.append("")
    try:
        language0[i] = re.sub("[\(\[\].*?[\]\]]", "", language0[i])
        language0[i] = re.sub('\^', "", language0[i])
        language0[i] = re.sub('@', "", language0[i])
        language0[i] = re.sub('\(', "", language0[i])
        language0[i] = re.sub('\[', "", language0[i])
        language0[i] = re.sub('\)', "", language0[i])
        language0[i] = re.sub('\]', "", language0[i])
        language0[i] = ''.join([x for x in language0[i] if not x.isdigit()])
        language0[i] = language0[i].strip()
    except:
        pass
path = "Excel/df4.xlsx"
print(len(language0))
print(len(language1))
print(len(word_type))
data = {'language0': language0,
        'language1': language1,
        'word_type': word_type}
df = pd.DataFrame(data)
```

```
df.to_excel(path, index = False)
```

### 3.2.2 Mô tả

- Tương tự từ điển 1, các file Word được truyền vào chương trình thông qua đường dẫn đến tập tin.
- Chương trình duyệt qua văn bản và chia văn bản thành từng đoạn. Đối với mỗi đoạn, chương trình sẽ ra từng thành phần bao gồm từng vừng Bahnar, nghĩa tiếng Việt và từ loại dựa vào các dấu “[:]”.
- Ngoài ra, do lỗi trong quá trình OCR nên văn bản có những từ không định dạng được nên sẽ loại ra khỏi chương trình.
- Đưa các nội dung đã trích xuất vào các mảng tương ứng và tạo một data frame từ chúng. Truyền data frame đó vào file Excel.

### 3.2.3 Kết quả thu được

adar adeh	chạm rỗi, từ
từ, thông thả. Nễ tổ 'don, bôn bôn adar adeh duh truh	Đừng lo lắng, chúng ta đi từ từ cũng sẽ tới. Jang adar adeh, nễ hơroh horei: Làm thông thả, đừng hấp tấp.
ade	1- một loại lỗ ồ. 2- rong dưới sông, hồ ao.
adiêng	bí tích. Topoh Adiêng: Bảy phép Bí Tích.
adoi	cũng, đều.
adon	buồng chuối, bông lúa.
adra	giàn bếp.
adra kiêk	xương đòn (xương nối từ ức tới vai),
adrah	độn, ghé.
adrah	đựng cụ phát ra âm thanh chạy bằng nước để đuổi chim, hay thú vật đặt ở rẫy.
adrak dt]	chì.
a drap	cắm cùm.
hang động vật. X	par *
adráp	lại, lần nữa.
adrâu	cái xăm kéo cá.
adreh	độn, ghé.
bị thần giáng họa, nỡ thần hứa với thần, nhưng không làm theo lời hứa, r	ôm đau; quan niệm xưa), x; hodrê.
adrêk	thon dần
adrêl	ngay lập tức, ngay khi.
adrê	1- bỏ. Adrê khop: Bỏ đạo. 'bok Roh xang adrê kon akân boih: ông Roh đã bỏ vợ con rồi.
chết. Mẽ nhân xang adrê nhân minh xomâm boih	Mẹ chúng tôi chết đã một năm rồi.
adrin	cố gắng. Ih athai adrin bờ jung: Anh phải cố
gắng làm việc. Adrin pokel nải ầu	cố gắng làm xong hôm nay.
adring	cùng một lúc, một lượt,
adro	con ve ve.
adro KJ	1- góa
adroi	trước. Hấp bắk adroi kơ inh: Nó đi trước tôi. Adroi xồ hấp jì bongai kiê: Trước kia nó là người ăn cắp. Bỏ hời adroi: Đi kẻ trước

## 4 Nhận xét

### 4.1 Những gì nhóm đã làm được

Nhóm đã tận dụng được thư viện và ngôn ngữ Python để thực hiện được việc đọc các từ điển và xuất ra thành excel dạng bảng như yêu cầu đề bài.

### 4.2 Khó khăn

- Chưa làm được giao diện theo yêu cầu đề bài.
- Do OCR nên các từ điển có nhiều lỗi (lỗi về font chữ, lỗi định dạng, ...). Nhóm đang tìm cách tự động hóa sửa lỗi nhưng chưa thành công, vì vậy nhóm quyết định sửa thủ công. Các lỗi định dạng trong từ điển 1 và tiến trình sửa:



1. đã xong:  
23, 24, 25, 27, 28, 31-34, 3942, 45, 46, 49, 50, 54, 55, 60, 61, 74, 75, 79, 80, 83, 84, 85, 93, 94, 106, 107, 110, 111, 117, 118, 121, 122, 127-130, 135, 136, 139, 140, 143-146, 152, 153, 158, 159, 162-167, 171, 172, 177-182, 188, 194, 195, 198-201, 210-213, 248, 254, 255, 258, 259, 262, 263, 268, 269, 271-274, 282, 283, 304, 318-321, 324-325.
  2. đã xong:  
1, 2, 5, 26-30, 41, 42, 48, 57, 58, 75-78, 97-101, 113, 114, 121-125, 127, 131, 132, 136, 137, 140-143, 148, 149, 157, 158, 162, 163, 169-174, 183-189, 209, 210, 213, 214, 217-220, 223, 228, 229, 235, 236, 240, 241, 245, 246, 252, 255, 262, 263, 275, 276, 298, 307, 308, 312, 325, 328, 341, 342, 346, 347, 352, 353, 360, 361, 373, 374, 378, 379, 383, 384,
  3. đã xong:  
4, 5, 10, 11, 21, 22, 42, 43, 61-65, 73, 74, 84, 85, 92, 93, 96, 97, 105, 106, 122-126, 131-133, 137-140, 145-148, 158, 159, 163, 164, 173-178, 181, 182, 190, 191, 197, 198, 203, 209, 210, 217, 218, 226, 227, 242, 243, 246-248, 255-257, 264, 265, 269-272, 278, 279, 288-298, 309, 310, 316, 317, 320-329, 339-342, 350, 351, 353-356, 364, 379, 380, 384, 385, 389-392, 407, 408, 413-415, 419-422, 430, 436-445
  4. chưa hoàn thành:  
3, 5, 8, 11-13, 15-22, 26, 28, 29, 31, 33, 37, 38, 40, 42, 45, 51, 52, 56, 59, 62-68, 71, 72, 86, 88, 90, 92, 94, 96, 103, 106-110, 114, 116, 119, 123, 124, 127, 128.
- Ngoài ra, trong các file excel được xuất ra có những dữ liệu nhiều vẫn chưa xử lý được.